Chronic Obstructive Pulmonary Diseases:
Journal of the COPD Foundation®

Original Research

Deep Learning Integration of Chest Computed Tomography Imaging and Gene Expression Identifies Novel Aspects of COPD

Junxiang Chen, PhD[1] Zhonghui Xu, MS[2] Li Sun, MS[1] Ke Yu, MS[1] Craig P. Hersh, MD, MPH[2,3] Adel Boueiz, MD, MS[2,3] John E. Hokanson, MPH, PhD[4] Frank C. Sciurba, MD[5] Edwin K. Silverman MD, PhD[2,3] Peter J. Castaldi, MD, MS[2,6*] Kayhan Batmanghelich, PhD[1*]

Abstract

Rationale: Chronic obstructive pulmonary disease (COPD) is characterized by pathologic changes in the airways, lung parenchyma, and persistent inflammation, but the links between lung structural changes and blood transcriptome patterns have not been fully described.

Objectives: The objective of this study was to identify novel relationships between lung structural changes measured by chest computed tomography (CT) and blood transcriptome patterns measured by blood RNA sequencing (RNA-seq).

Methods: CT scan images and blood RNA-seq gene expression from 1223 participants in the COPD Genetic Epidemiology (COPDGene®) study were jointly analyzed using deep learning to identify shared aspects of inflammation and lung structural changes that we labeled image-expression axes (IEAs). We related IEAs to COPD-related measurements and prospective health outcomes through regression and Cox proportional hazards models and tested them for biological pathway enrichment.

Results: We identified 2 distinct IEAs: IEA_{emph} which captures an emphysema-predominant process with a strong positive correlation to CT emphysema and a negative correlation to forced expiratory volume in 1 second and body mass index (BMI); and IEA_{airway} which captures an airway-predominant process with a positive correlation to BMI and airway wall thickness and a negative correlation to emphysema. Pathway enrichment analysis identified 29 and 13 pathways significantly associated with IEA_{emph} and IEA_{airway}, respectively (adjusted $p<0.001$).

Conclusions: Integration of CT scans and blood RNA-seq data identified 2 IEAs that capture distinct inflammatory processes associated with emphysema and airway-predominant COPD.

1. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States
2. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States
3. Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States
4. Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States
5. Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States
6. Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts, United States

Equal contribution

Abbreviations:

AP=airway predominant; **BMI**=body mass index; **COPD**=chronic obstructive pulmonary disease; **CT**=computed tomography; **CSRL**=context-aware self-supervised representation; **DICOM**=Digital Imaging and Communications in Medicine standard; **FA**=factor analysis axe; **FA_{airway}**=airway disease FA; **FA_{emph}**=emphysema FA; **FDR**=false discovery rate; **FEV_1**=forced expiratory volume in 1 second; **FVC**=forced vital capacity; **%gas trapping**=%low attenuation area using −856 Hounsfield unit threshold on expiratory CT scan; **GO**=gene ontology; **GOLD**=Global initiative for chronic Obstructive Lung Disease; **HU**=Hounsfield unit; **IEA**=image-expression axe; **IEA_{emph}**=emphysema IEA; **IEA_{airway}**=airway disease IEA; **MLP**=multilayer perceptron; **mMRC**=modified Medical Research Council; **PCA**=principal component analysis; **PCA-Is**=PCA-image only axes; **perc15**=15th percentile Hounsfield unit in inspiratory CT scan; **Pi10**=the average wall thickness for a hypothetical airway of 10-mm lumen perimeter on CT; **%pred**=percentage predicted; **PRISm**=preserved ratio-impaired spirometry; **RNAseq**=RNA sequencing; **RRS**=relatively resistant smoker; **SE**=severe emphysema; **SNR**=single nucleotide polymorphism; **SGRQ**=St George's Respiratory Questionnaire; **%WA segmental**=the percentage of airway wall area for 3rd generation bronchi

Funding Support:

This work was supported by the National Heart, Lung, and Blood Institute (NHLBI) R01HL141813, K08 HL141601, R01 HL124233, R01 HL126596, R01 HL147326, U01 HL089897, and U01 HL089856. The COPD Gene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion.

Citation:

Chen J, Xu Z, Sun L, et al. Deep learning integration of chest computed tomography and gene expression identifies novel aspects of COPD *Chronic Obstr Pulm Dis.* 2023;10(4):355-368. doi: https://doi.org/10.15326/jcopdf.2023.0399

Publication Dates:

Date of Acceptance: June 30, 2023
Published Online Date: July 6, 2023

Address correspondence to:

Junxiang Chen, PhD
Department of Biomedical Informatics
University of Pittsburgh
5607 Baum Blvd.
Pittsburgh, PA 15206.
Phone: (412) 624-5100
Email: juc91@pitt.edu

Keywords:

machine learning; emphysema; genomics

This article has an online supplement

Introduction

Chronic obstructive pulmonary disease (COPD) is one of the most prevalent chronic diseases,[1] responsible for approximately 3 million deaths annually.[2] COPD is characterized by persistent respiratory symptoms and poorly reversible airflow limitation.[3] It is associated with an abnormal inflammatory response of the lungs to cigarette smoke or other noxious particles,[4] which results in lung structural changes, including the loss or narrowing of airways (airway disease) and parenchymal destruction (emphysema).[5] In addition to its characteristic lung structural changes, the changes in blood transcriptome patterns have been linked to COPD exacerbations[6,7] and lung function decline.[8]

Although lung structural changes and the changes in blood transcriptome patterns are characteristic aspects of COPD, their relationship remains unclear. Therefore, we are motivated to apply a deep learning method to analyze computed tomography (CT) imaging and blood RNA-sequencing (RNA-seq) data to identify novel relationships between them.

Based on the paradigm of COPD as a collection of treatable traits,[9] we hypothesize that COPD heterogeneity can be described by continuous measures corresponding to distinct disease processes that are present to varying degrees in affected individuals. We refer to these continuous measures as "disease axes,"[10] and we further hypothesize that integrative analysis of CT images and blood RNA-seq data can identify disease axes that reveal patterns of association between lung structural abnormalities and blood transcriptomic profile change. We tested these hypotheses by training a deep learning model on data from 1223 participants in the COPD Genetic Epidemiology (COPDGene®) study[11] with CT scans and blood gene expression data. Our analysis identified 2 disease axes that capture patterns of CT features consistent with emphysema and airway-predominant disease processes that are also associated with emphysema core-peel distribution and specific inflammatory pathways.

Methods

A comprehensive description of methods is included in the online supplement, and all analysis code is available in a GitHub repository (https://github.com/batmanlab/IEA).

Participant Enrollment and Data Collection

COPDGene enrolled 10,198 participants with a minimum 10-pack-year lifetime smoking history at 21 centers across the United States (NCT00608764).[11] Individuals with a history of lung diseases other than asthma, such as pulmonary fibrosis, extensive bronchiectasis, and cystic fibrosis, are excluded from this study. Five-year follow-up data are available for 6717 participants, and 10-year follow-up visits are currently being completed. Participants underwent spirometry, questionnaire assessments, standardized inspiratory and expiratory chest CT imaging, and genome-wide single nucleotide polymorphism (SNP) genotyping. At the second visit (year 5), complete blood counts were conducted, PAXgene blood RNA tubes were collected, and RNA-seq was performed. Each center obtained institutional review board approval, and all participants provided written informed consent.

Learning Image-Expression Axes

Only participants whose CT scans were obtained on Siemens scanners with the b31f kernel and with RNA-seq data available at the second visit were analyzed. CT features were extracted from Digital Imaging and Communications in Medicine (DICOM) standard image files using the following procedure. Every inspiratory chest CT scan was divided into 581 patches, each with a volume of 32^3 mm^3. We extracted 128 features from each patch using context-aware self-supervised representation learning (CSRL),[12] resulting in a 581×128-dimensional matrix for each scan.

Image-expression axes (IEAs) were constructed where CSRL features were the input to a multilayer perceptron (MLP) that produced a low-dimensional representation for each patch. Further supervised dimension reduction was performed to obtain participant-level IEAs using a product

of an expert's model.[13] At this stage, we applied statistically independent constraints with the Hilbert-Schmidt independence criterion[14] to ensure that each IEA captured an independent disease process. A final linear layer used IEAs as the input to predict the expression levels of the genes simultaneously. The parameters of the model were jointly optimized via Adam, an optimization algorithm used to train a machine-learning model.[15] In the training process, we evaluated the impact of feature selection on genes by testing each gene for the association with the top-128 principal components of the CSRL features in the training dataset. We also evaluated various thresholds for gene inclusion determined by the *p*-value of the F-test for each gene.

We randomly split the data into training and testing sets with sizes of 923 and 300, respectively. Model training was performed in the training set using 5-fold cross-validation, giving us 5 models. The final IEAs were given by taking the average value of the IEAs from the 5 models.

Association of Image-Expression Axes With Clinical Measurements

We computed Pearson correlation coefficients between IEAs and clinical measurements to understand their association. A full description of the measurements is included in the online supplement. Multivariable analyses were conducted by training ordinary least squares models for continuous measurements and logistic regression models for categorical measurements. We conducted a survival analysis (starting from the second visit) with the Cox proportional hazards model.[16] We applied the IEA model to 1527 participants from another subset of the COPDGene dataset to provide independent replication of our IEA associations. These participants had their CT scans available but without RNA-seq data and, therefore, were not used for model training. A full description of these models, including model covariates, is provided in the online supplement.

Comparison of Image-Expression Axes to Other Disease Axes

We compared IEAs to the following disease axes:

1. COPD factor analysis axes (FAs): Previously published phenotype disease axes identified through factor analysis.[17]
2. Principal component analysis image-only axes (PCA-Is): Disease axes constructed by applying PCA to the CSRL features.

The comparative analyses included the calculation of Pearson correlation coefficients between IEAs and other disease axes, association analyses to clinical measurements, and comparison of nested models for clinical outcomes utilizing IEAs, FAs, and PCA-Is in determining whether IEAs improved the performance of models already containing FAs and/or PCA-Is.

Differential Expression and Usage Analyses

To identify the genes and pathways associated with IEAs, we conducted a differential gene expression analysis using the *voom* function[18] of the *limma* package.[19] Voom prepares RNA-Seq data for linear modeling. It works in conjunction with limma, a library specifically designed to assess differential expression through linear models. Multiple comparisons were corrected with the Benjamini-Hochberg method to control the false discovery rate (FDR)[20] at 10%. Gene ontology (GO) pathway enrichment analysis was performed for pathways in the "biological process" category using the Top GO (v2.33.1) method.[21] The threshold for statistical significance was an adjusted *P*-value<0.001.

Results

A total of 1223 participants in the COPDGene study with complete CT scan and blood RNA-seq data were analyzed, and the flow diagram for the selection of participants for analysis is shown in Supplemental Figure E1 in the online supplement. The analyzed participants were 50% female, 82% non-Hispanic White, and 18% African American, and the average age of the participants was 67 years. The Global initiative for chronic Obstructive Lung Disease (GOLD)[3] spirometric stage distribution of participants was 42.5% in GOLD 0, 44.0% in GOLD 1–4, and 13.5% with preserved ratio-impaired spirometry (PRISm). For model training and validation, participants were split into training and test sets, with no statistically significant differences in demographic or key clinical characteristics between these groups (Table 1).

Image-Expression Axe Model Training and Reproducibility Analysis

A schematic overview of the model training process is shown in Figure 1, and the data flow is summarized in Supplemental Figure E2 in the online supplement. To maximize the stability of the IEAs and reduce the effects of sampling variability, we used nested cross-validation in the model training process to select the number of genes included in the model and the number of IEAs identified. For gene selection, we tested genes in the training data for the association to the top 128 principal components of the CSRL image features using an F-test, and a series of *p*-value thresholds for gene inclusion were explored, ranging from $p=1x10^{-6}$ to $p=1$. The resulting IEAs were found to be stable across the entire range of *p*-value thresholds, and the threshold corresponding to $p=0.01$ was selected (Pearson's r for IEAs across cross-validation folds ≥0.96, Supplemental Table E1 in the online supplement). With this threshold, 4685 genes were included in the final model. The number

Table 1. Participant Characteristics in Training and Test Data

	Training (n=923)	Test (n=300)	*p*-value
Demographics			
Age	66.6±8.4	66.6±9.2	0.961
Gender, %females	49.7%	50.7%	0.778
Race, %African American	18.6%	19.3%	0.788
Clinical Measurements			
%Current Smoker	30.3%	28.5%	0.661
Body Mass Index	29.1±6.4	29.1±5.9	0.646
Pack Years	44.8±24.4	44.3±25.0	0.860
FEV_1 %pred	77.7±25.1	76.0±24.8	0.226
FEV_1 / FVC	0.67±0.15	0.66±0.15	0.227
SGRQ Total Score	21.3±19.4	22.4±21.1	0.818
mMRC Dyspnea Score	1.1±1.3	1.2±1.4	0.464
6-Minute-Walk Distance	1292±443	1260±433	0.308
Frequent Exacerbator-History	6.3%	7.7%	0.403
Quantitative CT			
perc15	-922±28	-923±29	0.637
%Emphysema at -950HU	6.9±10.3	7.2±10.8	0.497
%Gas Trapping	21.4±20.2	21.1±19.1	0.612
Pi10	2.2±0.6	2.3±0.6	0.159
%WA Segmental	49.9±8.2	50.2±8.2	0.491
$Q_{perc15_{peel-core}}$	-2.5±1.6	-2.5±1.6	0.459
GOLD Stages			
%GOLD 0	43.2%	39.7%	0.300
%GOLD 1	8.5%	10.3%	0.349
%GOLD 2	19.4%	20.5%	0.663
%GOLD 3	11.5%	9.9%	0.471
%GOLD 4	4.6%	5.5%	0.554
%PRISm	12.9%	14.0%	0.611
Longitudinal			
ΔFEV_1 %predicted	-0.3±2.1	-0.3±2.3	0.789
ΔFEV_1 / FVC	-0.0±0.0	-0.0±0.0	0.235
Frequent Exacerbator-Future	4.8%	8.6%	0.084
5-year Mortality	14.7%	15.2%	0.769

Continuous variables are expressed as means and standard deviations.
Categorical variables are expressed as percentages.
P-values are obtained using the Kruskal-Wallis test for continuous variables and Chi-square test for proportions, comparing the training and test data.
ΔFEV_1 %predicted and ΔFEV_1/FVC are computed by subtracting the visit 3 values from the visit 2 values of FEV_1 % of predicted or FEV_1/FVC and dividing them by the number of years between the 2 visits.
$Q_{perc15_{peel-core}}=100*log(perc15_{peel}/ perc15_{core})$, where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary.

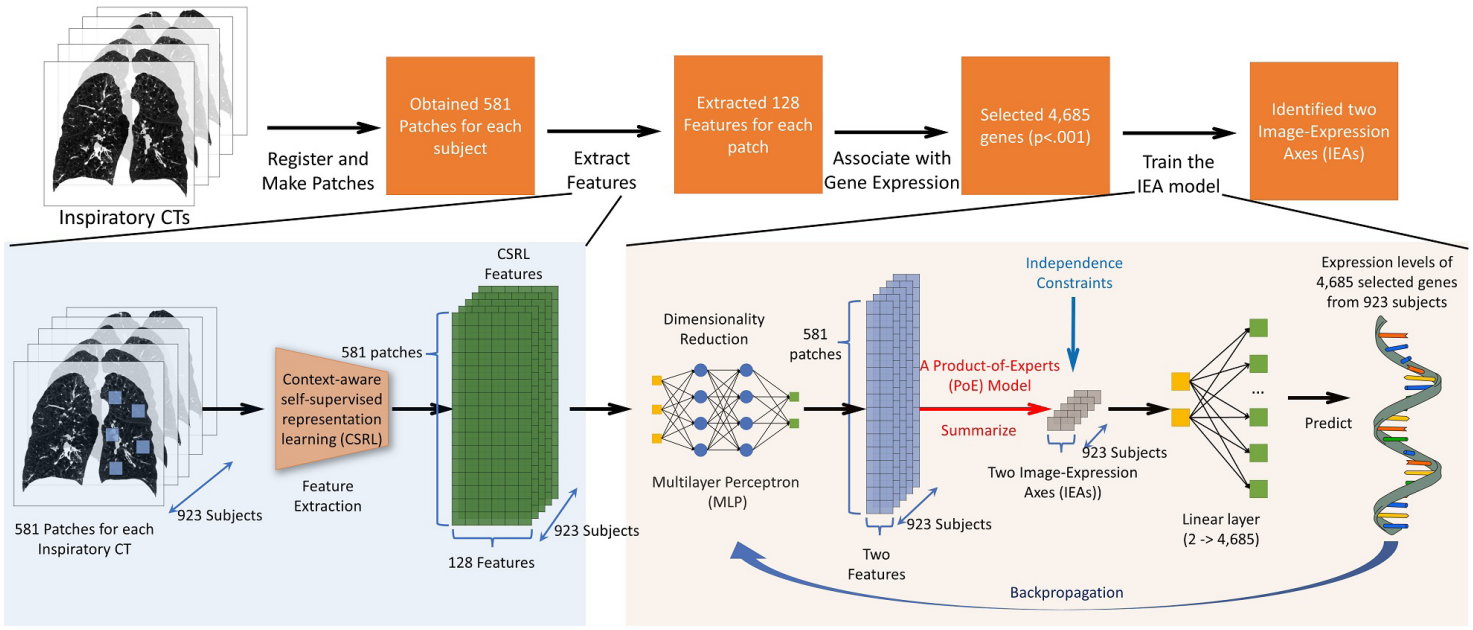
FEV_1=forced expiratory volume in 1 second; FVC=forced vital capacity; SGRQ=St George's Respiratory Questionnaire; mMRC=modified Medical Research Council dyspnea scale; CT=computed tomography; perc15=15th percentile Hounsfield unit in inspiratory CT scan; HU=Hounsfield unit; %gas trapping=%low attenuation area using −856 Hounsfield unit threshold on expiratory CT scan; Pi10=the average wall thickness for a hypothetical airway of 10-mm lumen perimeter on CT; %WA segmental=the percentage of airway wall area for 3rd generation bronchi; GOLD=Global initiative for chronic Obstructive Lung Disease; PRISm=preserved ratio-impaired spirometry

of IEAs identified by the model was determined by the amount of gene expression variance explained, which was the highest with 2 IEAs (Figure 2).

Image-Expression Axe Association to Clinical and Radiographic Features and Prospective Outcomes

To provide a clinical interpretation of the IEAs, we calculated their correlation to a range of COPD-related clinical and imaging measurements Figure 3. We refer to the first IEA as the emphysema axis (IEA_{emph}), because it demonstrates a pattern of clinical associations consistent with quantitative emphysema. Specifically, higher levels of IEA_{emph} were associated with lower lung function, emphysema, lower body mass index (BMI), and a lower likelihood of being a current smoker. The second IEA is consistent with airway disease and is referred to as IEA_{airway}. Higher levels of IEA_{airway} were associated with higher BMI, thicker airways, and less emphysema. The analysis comparing IEAs with blood cell counts reveals that IEA_{emph} is positively associated with the neutrophil count, proportion of neutrophils, and monocyte count, but negatively correlated with lymphocyte count and proportion. On the other hand, IEA_{airway} is

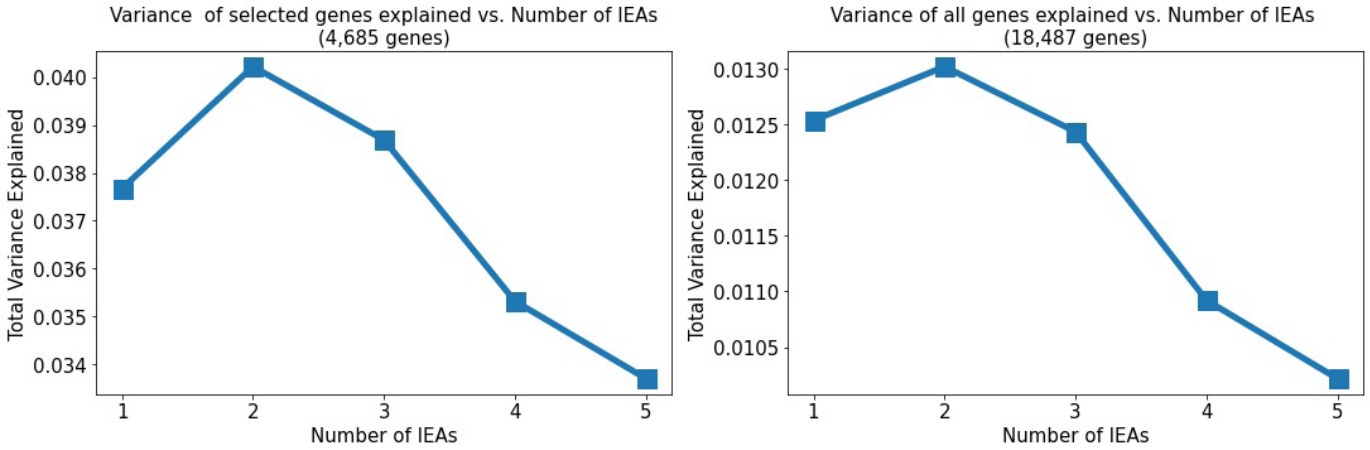
Figure 1. Overview of the Machine Learning Workflow



CT images were processed as 32^3 mm^3 patches from which 128 features were constructed using the CSRL.[12] These features were the input for a MLP that processed a 581×128×923 tensor to output a 581×2×923 participant-level data representation. The participant-level latent representation (IEAs) is given by summarizing the patch-level features into a matrix of 2×923. We introduce a linear layer (2 -> 4685) that estimates gene expression for each participant, taking the IEA as the input. We apply independence constraints to ensure IEAs are independent of each other. The overall objective function is given by minimizing the mean-squared error of the gene expression levels in prediction.

CT=computed tomography; IEAs=image-expression axes; CSRL=context-aware self-supervised representation learning; MLP=multilayer perceptron

Figure 2. Total Variance of the Gene Expression Explained Versus the Number of Image-Expression Axes



The figure on the left shows the plot of the 4685 selected genes. The figure on the right shows the plot for all the genes. The figures show that when the number of IEAs is 2, the total variance explained is maximized. We choose the number of IEAs to be two.

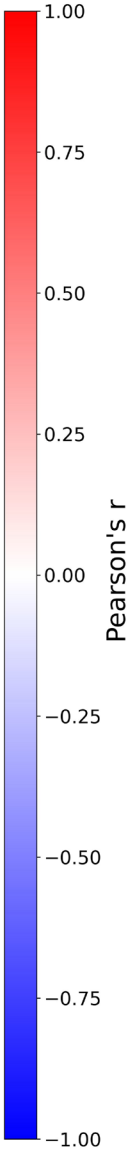
IEAs=image-expression axes

positively associated with both white blood cell count and neutrophil count. The IEAs were uncorrelated with each other, suggesting that they may capture different underlying disease processes.

Figure 3 reveals that both IEAs showed a negative association with emphysema peel/core distribution ($Q_{perc15_{peel\text{-}core}}$), indicating that higher IEA values are linked to more emphysema in the central regions of the lung. To further explore this association, we conducted sensitivity analyses by dividing the lung into concentric bands based on the distance to the lung boundary and defining the peel region with different bands (as detailed in the online supplement). The results suggest that IEA_{emph} exhibits a consistently positive association with $Q_{perc15_{peel\text{-}core}}$, regardless of the band used to define it. However, it is currently unclear whether the association between IEA_{airway} and $Q_{perc15_{peel\text{-}core}}$ reflects a real biological phenomenon at the extreme periphery of the lung, or if it is an artifact of segmentation.

Figure 3. Pearson's Correlation Between Image-Expression Axes and COPD-Related Characteristics and Health Outcomes

	IEA_{emph}	IEA_{airway}
Clinical Measurements		
Smoking Status	-0.31^{a}	-0.10
Body Mass Index (BMI)	-0.14^{b}	0.68^{a}
Pack Years	0.28^{a}	0.12^{b}
FEV_1 %Predicted	-0.45^{a}	0.00
FEV_1/FVC	-0.65^{a}	0.19^{a}
SGRQ Total Score	0.40^{a}	0.08
mMRC Dyspnea Score	0.40^{a}	0.12^{b}
6-Minute-Walk Distance	-0.22^{a}	-0.30^{a}
Frequent Exacerbator (History)	0.12^{b}	-0.03
Quantitative CT		
perc15	-0.76^{a}	0.54^{a}
%Emphysema at -950HU	0.72^{a}	-0.34^{a}
%Gas Trapping	0.74^{a}	-0.32^{a}
Pi10	0.14^{b}	0.17^{c}
%WA Segmental	0.09	0.34^{a}
$Q_{perc15_{peel-core}}$	-0.36^{a}	-0.32^{a}
Complete Blood Count		
White Blood Cell Count	0.11	0.13^{b}
Neutrophil Count	0.24^{a}	0.12^{b}
Neutrophil %	0.33^{a}	0.10
Lymphocyte Count	-0.26^{a}	0.05
Lymphocyte %	-0.39^{a}	-0.10
Monocyte Count	0.16^{c}	0.11
Monocyte %	0.10	-0.06
Eosinophil Count	0.09	0.05
Eosinophil %	0.04	0.01
Basophil Count	0.04	-0.03
Basophil %	-0.02	-0.11
Longitudinal		
ΔFEV_1%predicted	-0.03	0.14
ΔFEV_1/FVC	-0.03	0.07
Frequent Exacerbator (Future)	0.28^{a}	-0.10
5-year Mortality	0.31^{a}	0.04
Image-Expression Axes (IEAs)		
IEA_{emph}	1.00^{a}	-0.03
IEA_{airway}	-0.03	1.00^{a}



[a]$p<.01$
[b]$p<.001$
[c]$p<.05$

$Q_{perc15_{peel-core}}$=100 log($perc15_{peel}$/$perc15_{core}$), where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary.
ΔFEV_1 %predicted and ΔFEV_1/FVC are computed by subtracting the visit 3 value from the visit 2 value of FEV_1 % of predicted or FEV_1/FVC and dividing it by the number of years between the 2 visits.

FEV_1=forced expiratory volume in 1 second; FVC=forced vital capacity; SGRQ=St George's Respiratory Questionnaire; mMRC=modified Medical Research Council dyspnea scale; CT=computed tomography; perc15=15th percentile Hounsfield unit in inspiratory CT scan; %gas trapping=%low attenuation area using −856 Hounsfield unit threshold on expiratory CT scan; Pi10=the average wall thickness for a hypothetical airway of 10-mm lumen perimeter on CT; %WA segmental=the percentage of airway wall area for 3rd generation bronchi; IEA=image-expression axes; IEA_{emph}=emphysema IEA; IEA_{airway}=airway disease IEA

To determine whether the IEAs provided clinical information in addition to standard demographic variables, we tested the significance of adding IEAs to regression models for various COPD-related measures (Table 2, Table 3, and Supplemental Tables E2 and E3 in the online supplement). After adjusting for standard demographic variables, both IEA_{emph} and IEA_{airway} were significantly associated with forced expiratory volume in 1 second (FEV_1) percentage predicted (%pred), FEV_1 to forced vital capacity (FVC) ratio, St George's Respiratory Questionnaire (SGRQ) total score, modified Medical Research Council (mMRC) dyspnea score, and 6-minute-walk distance, as well as neutrophil count. Additionally, IEA_{emph} was also associated with being a frequent exacerbator-history, a frequent exacerbator-future, an all-cause mortality rate, monocyte proportion, and eosinophil count. On the other hand, IEA_{airway} is associated with white blood cell count, neutrophil proportion, lymphocyte count, and lymphocyte proportion.

Table 2. Multivariable Associations of Image-Expression Axes to Continuous COPD-Related Characteristics and Health Outcomes

	$\beta_{IEA_{emph}}$ (95%CI)	$\beta_{IEA_{emph}}$ (95%CI)
Clinical Measurements		
Body Mass Index	-0.12[a] (-0.22, -0.03)	0.71[b] (0.63, 0.79)
FEV_1 %predicted	-0.54[b] (-0.66, -0.42)	-0.04 (-0.14, 0.07)
FEV_1/FVC	-0.66[b] (-0.76, -0.56)	0.18[b] (0.09, 0.26)
SGRQ Total Score	0.56[b] (0.44, 0.68)	0.14[a] (0.04, 0.24)
mMRC Dyspnea Score	0.51[b] (0.40, 0.63)	0.16[a] (0.06, 0.26)
6-Minute-Walk Distance	-0.24[b] (-0.37, -0.12)	-0.28[b] (-0.39, -0.18)
Quantitative CT		
perc15	-0.71[b] (-0.76, -0.66)	0.54[b] (0.50, 0.59)
%Emphysema at -950HU	0.81[b] (0.73, 0.89)	-0.30[b] (-0.37, -0.23)
%Gas Trapping	0.73[b] (0.65, 0.81)	-0.36[b] (-0.42, -0.29)
Pi10	0.22[a] (0.09, 0.36)	0.20[b] (0.08, 0.32)
%WA Segmental	0.11 (-0.02, 0.24)	0.36[b] (0.25, 0.47)
$Q_{perc15_{peel-core}}$	-0.36[b] (-0.47, -0.24)	-0.30[b] (-0.41, -0.20)
Blood Cell Counts		
White Blood Cell Count	0.10 (-0.03, 0.24)	0.13[c] (0.01, 0.24)
Neutrophil Count	0.18[a] (0.04, 0.31)	0.09 (-0.02, 0.21)
Neutrophil %	-0.12 (-0.25, 0.01)	0.09 (-0.03, 0.20)
Lymphocyte Count	0.07 (-0.06, 0.20)	0.10 (-0.01, 0.21)
Lymphocyte %	0.12 (-0.02, 0.26)	0.07 (-0.05, 0.19)
Monocyte Count	0.02 (-0.11, 0.16)	-0.03 (-0.15, 0.09)
Monocyte %	0.17[a] (0.05, 0.30)	0.05 (-0.06, 0.15)
Eosinophil Count	-0.21[b] (-0.33, -0.09)	-0.04 (-0.15, 0.06)
Eosinophil %	0.03 (-0.10, 0.16)	-0.05 (-0.16, 0.07)
Basophil Count	0.06 (-0.08, 0.20)	0.02 (-0.10, 0.14)
Basophil %	0.02 (-0.12, 0.16)	-0.09 (-0.21, 0.03)
Longitudinal		
ΔFEV_1 %predicted	-0.20 (-0.44, 0.03)	0.03 (-0.17, 0.23)
ΔFEV_1/FVC	-0.11 (-0.35, 0.13)	-0.00 (-0.20, 0.20)

[a] $p<.01$
[b] $p<.001$
[c] $p<.05$

The table reports the β coefficients and corresponding 95% confidence intervals for IEA_{emph} and IEA_{airway} in linear models using the indicated COPD-related measurement or health outcomes as the response variable. All models were adjusted for age, gender, race, pack years, smoking status.
$Q_{perc15_{peel-core}}$=100 log($perc15_{peel}$/$perc15_{core}$), where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary.
ΔFEV_1 %predicted and ΔFEV_1/FVC are computed by subtracting the visit 3 value from the visit 2 value of FEV_1 % of predicted or FEV_1/FVC and dividing it by the number of years between the 2 visits.

COPD=chronic obstructive pulmonary disease; CI=confidence interval; FEV_1=forced expiratory volume in 1 second; FVC=forced vital capacity; SGRQ=St George's Respiratory Questionnaire; mMRC=modified Medical Research Council dyspnea scale; CT=computed tomography; perc15=15th percentile Hounsfield unit in inspiratory CT scan; %gas trapping=%low attenuation area using −856 Hounsfield unit threshold on expiratory CT scan; Pi10=the average wall thickness for a hypothetical airway of 10-mm lumen perimeter on CT; %WA segmental=the percentage of airway wall area for 3rd generation bronchi

To provide independent replication of our IEA associations, the IEA model was applied to 1527 participants from another subset of the COPDGene dataset that had not been used for model training. All the significant associations to clinical and longitudinal measures remained significant with very similar effect estimates indicating a high level of reproducibility for IEAs (Supplemental Tables E4, E5, and E6 in the online supplement).

COPD Subgroups Defined by Image-Expression Axes and Comparison to Existing COPD Subtypes

To further understand the clinical characteristics of COPD subgroups defined by IEAs, we divided the IEA space into 4 quadrants (Figure 4) and computed the average characteristics of each subgroup (Supplemental Table E8 in the online supplement). As expected, participants with low IEA_{emph}/low IEA_{airway} values had the least obstruction (mean FEV_1 89.2% predicted), the highest percentage of GOLD spirometric grade 0 participants, low emphysema, and the thinnest airways. Participants with high IEA_{emph}/low IEA_{airway} values had characteristics consistent with emphysema-predominant COPD, namely high emphysema and low BMI, with about 70% of GOLD grade 4 participants present in this group. Participants with low IEA_{emph}/high IEA_{airway} values had an airway-predominant profile with thick airway walls, elevated BMI, and the greatest proportion of PRISm participants.

Table 3. Multivariable Associations of Image-Expression Axes to Frequent Exacerbations and Mortality

Logistic Regression Model	$\beta_{IEA_{emph}}$ (95%CI)	Odds Ratio (95%CI)	$\beta_{IEA_{airway}}$ (95%CI)	Odds Ratio (95%CI)
Frequent Exacerbator-History	0.85[a] (0.58, 1.12)	2.34[a] (1.79, 3.07)	0.12 (-0.10, 0.35)	1.13 (0.90, 1.41)
Frequent Exacerbator-Future	0.92[a] (0.47, 1.37)	2.51[a] (1.60, 3.94)	-0.09 (-0.48, 0.30)	0.91 (0.62, 1.34)
Cox Proportional Hazard Model	$\beta_{IEA_{emph}}$ (95%CI)	Hazard Ratio (95%CI)	$\beta_{IEA_{airway}}$ (95%CI)	Hazard Ratio (95%CI)
Mortality	0.66[a] (0.45, 0.87)	1.93[a] (1.57, 2.38)	0.17 (-0.00, 0.34)	1.19 (1.00, 1.41)

[a]p<.001

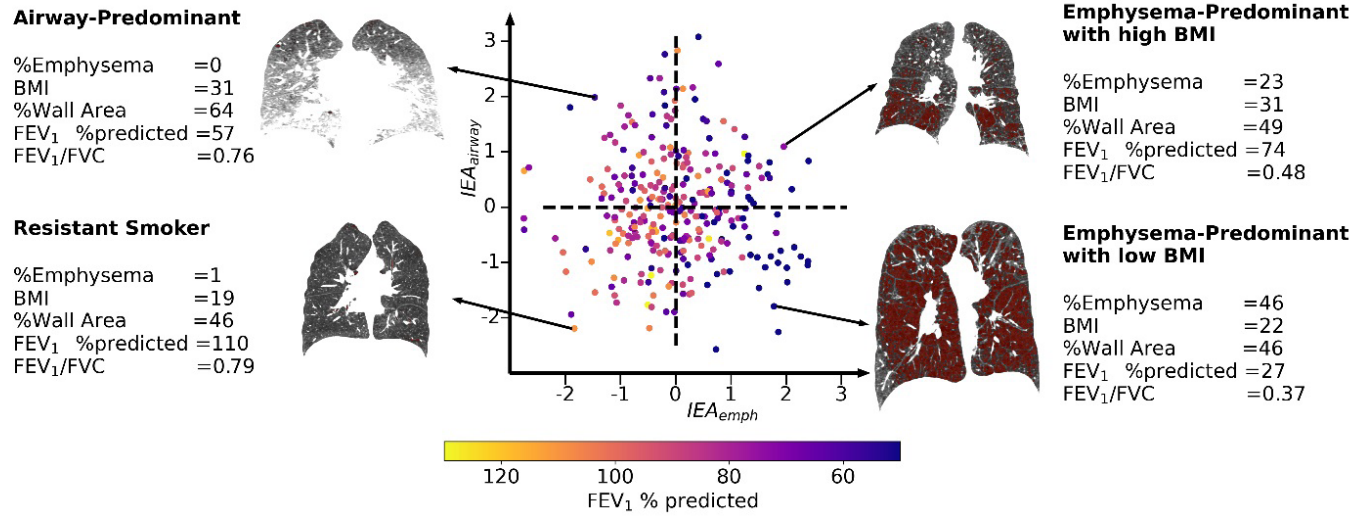
The table reports the β coefficients of the IEA_{emph} and IEA_{airway} and corresponding 95% confidence intervals from logistic regression models for frequent exacerbator status and a Cox proportional hazards model for mortality.

All models adjusted for age, gender, race, pack years, and smoking status as the covariates.

Frequent exacerbator-history indicated if the participants had at least 2 self-reported exacerbations during the 12 months before the second visit. Frequent exacerbator-future indicated if the participants had at least 2 self-reported exacerbations during the past 12 months before the third visit.

CI=confidence interval; IEA_{emph}=emphysema image-expression axes; IEA_{airway}=airway disease image-expression axes

Figure 4. Visualization of Participants Projected Along Each Identified Image-Expression Axis Dimension



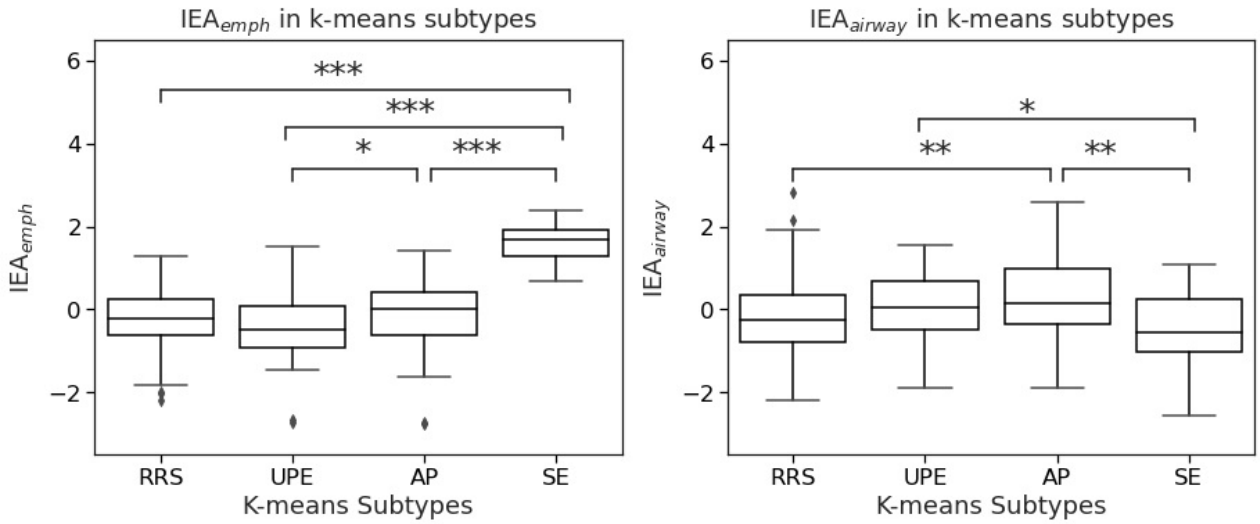
IEA_{emph} is the emphysema axis, where higher values indicate more severe emphysema. IEA_{airway} is the airway disease axis, where a higher value represents higher BMI and thicker airways. The space defined by these IEAs was used to stratify the cohort into 4 subgroups based on dividing the IEA space into 4 quadrants. Lung CT scans and clinical characteristics are shown for 1 participant in each quadrant, where the red mask represents the emphysema regions (<-950 HU). The characteristics of the 4 quadrants are summarized in Supplemental Table E8 in the online supplement.

BMI=body mass index; FEV_1=forced expiratory volume in 1 second; FVC=forced vital capacity; IEA=image-expression axis; IEA_{emph}=emphysema IEA; IEA_{airway}=airway disease IEA; CT=computed tomography; HU=Hounsfield units

Participants with high IEA_{emph}/high IEA_{airway} had the highest SGRQ total score, highest mMRC dyspnea scores, and shortest 6-minute walk distance. In terms of COPD progression, the groups differed significantly in mortality risk (p<0.001) and frequent exacerbation status (2 or more exacerbations in 1 year) but did not change in FEV_1. The group with the highest mortality was high IEA_{emph}/high IEA_{airway} followed by high IEA_{emph}/low IEA_{airway}. The latter group also had the highest percentage of participants with frequent exacerbations, both for retrospective (p<0.001, Chi-square test of all groups) and prospective exacerbations (p=0.048).

To place IEA_{emph} and IEA_{airway} in the context with previously reported subtypes and disease axes in COPDGene, we compared these axes directly with the previously reported k-means subtypes[22] and FAs.[17] In Figure 5, we observe that the highest values of IEA_{emph} are found in the severe emphysema k-means subtype, and the highest values of IEA_{airway} are found in the airway-predominant k-means subtype, confirming our clinical interpretation of these disease axes. Since the FAs also showed patterns consistent with emphysema (FA_{emph}) and airway-predominant disease (FA_{airway}), we compared the IEAs to the FAs and observed that IEA_{emph} and FA_{emph} showed a reasonably strong correlation (Pearson's r=0.58), but the IEA_{airway} and FA_{airway} axes showed only modest correlation (Pearson's r=0.28, see Supplemental Table E9 in the online supplement). Examination of the pattern of clinical associations for IEA_{airway} and FA_{airway}

Figure 5. Distribution of the Emphysema Image-Expression Axis and Airway Disease Image-Expression Axis Values Grouped by Previously Published COPD K-Means Clustering Subtypes[22]



*p<0.5
**p<.01
***p<.001

P-values are obtained using the Kruskal-Wallis test.
No symbol indicates non-significant results.

IEA=image-expression axes; IEA$_{emph}$=emphysema IEA; IEA$_{airway}$=airway disease IEA; RRS= relatively resistant smokers; UPE=upper zone dominant emphysema; AP=airway predominant; SE=severe emphysema

revealed that while airway axes were positively correlated to airway wall thickness, IEA$_{airway}$ is negatively correlated to emphysema, whereas, FA$_{airway}$ is positively correlated (Supplemental Table E10 in the online supplement). To determine whether the IEAs provided additional information about COPD phenotypes (FEV$_1$ % pred, FEV$_1$/FVC, SGRQ, mMRC, retrospective frequent exacerbations, and 6-minute walk distance) and COPD progression (mortality and prospective frequent exacerbations) above and beyond FAs, we constructed baseline models for each COPD phenotype and progression measurements with FAs included and then compared them to models including both FAs and IEAs. In most cases, the models with IEAs included outperformed the baseline models (p<0.001 for all COPD phenotypes and mortality, Supplemental Tables E11 and E12 in the online supplement).

Comparison to Principal Components Based on Images Alone

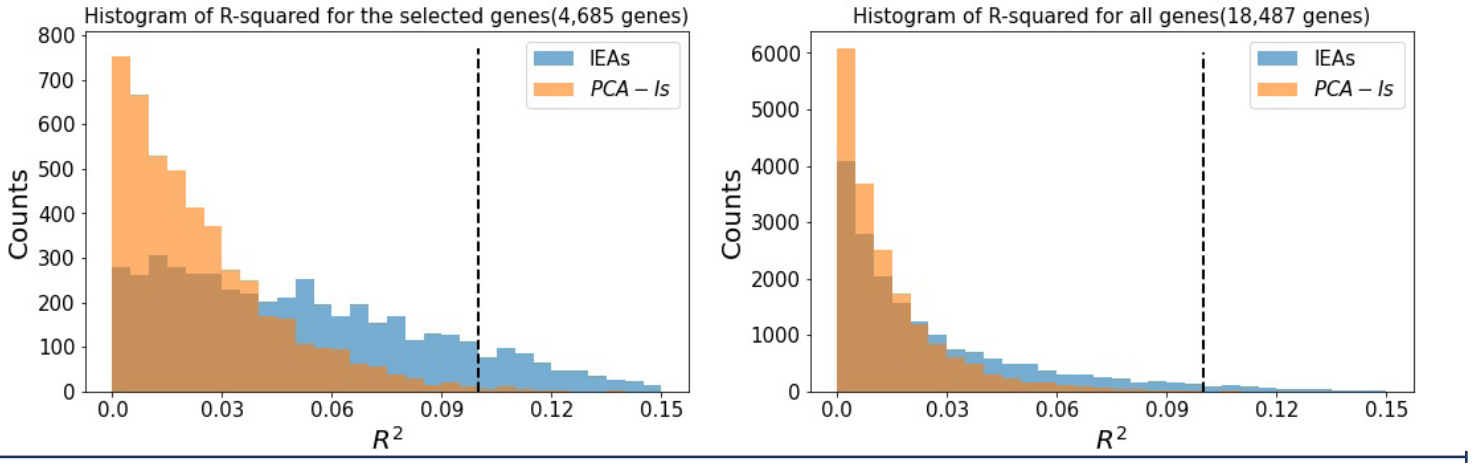
After observing that IEAs contain additional clinically relevant information relative to standard features extracted from CT images, we sought to determine whether the additional information came only from applying dimension reduction to the CT images (CSRL features), or whether there was added value from our algorithm that combined the CT features with gene expression. To make this comparison, we constructed disease axes from images only by using PCAs to extract the top 2 principal components (PCs) of the CSRL features, denoted as PCA-Is. We then compared the predictive performance of linear models that utilize both IEAs and PCA-Is with the nested version that involves the PCA-Is only, and we observed that the models including IEAs were superior to models with PCA-Is only for all of the 6 studied COPD phenotypes as well as prospective exacerbations and mortality (p<0.001, Supplemental Tables E13 and E14 in the online supplement). These results suggest that by incorporating gene expression data during training, IEAs extract more clinically important information than similar methods that utilize imaging features only.

Image-Expression Axes Are Associated With Inflammatory Pathways

To understand the biological aspect of the IEAs, we first confirmed that IEAs explained a greater proportion of gene expression variance than PCA-Is, as demonstrated in Figure 6, which shows that IEAs explain a greater proportion of variation on a per-gene basis than PCA-Is (557 genes with R^2>10% for IEAs versus 68 genes with R^2>10% for PCA-Is).

To identify specific biological processes associated with each IEA, we performed differential expression and pathway enrichment analysis. We identified 6494 and 3815 genes

Figure 6. Histograms for the Variances of Genes Explained (R^2) by the Image-Expression Axes and Principal Component Analysis- Image Only Axes



The figure on the left shows the histogram for the 4685 selected genes. Within these genes, there are 557 genes with R^2>10% for IEAs and 68 genes with R^2>10% for PCA-Is. The figure on the right shows the histogram for all the genes (18,487 genes in total). There are 622 genes with R^2>10% for IEAs and 69 genes with R^2>10% for PCA-Is.

IEAs=image-expression axes; PCA-Is=principal component analysis-image only axes; RRS=relatively resistant smokers; UPE=upper zone dominant emphysema; AP=airway predominant; SE=severe emphysema

associated at an FDR of 10% with IEA_{emph} and IEA_{airway}, respectively (Supplemental Tables E15 and E16 in the online supplement). GO pathway enrichment identified 29 and 13 enriched pathways (p-value<0.001) for IEA_{emph} and IEA_{airway}, respectively (Supplemental Tables E17 and E18 in the online supplement). The most significantly associated pathway for IEA_{emph} was neutrophil degranulation, whereas IEA_{airway} had the strongest enrichment for RNA processing. The most significant pathway results are shown in Table 4.

Discussion

In this paper, we used deep learning to identify novel connections between lung imaging features and blood gene expression. The deep learning model provided novel disease axes, i.e., IEAs, that captured elements of shared variability between CT scans and blood RNA-seq. We demonstrated: (1) that these IEAs are associated with important COPD-related physiologic and functional measures, (2) that these associations contained information that is independent from pre-existing, standard clinical and imaging variables, (3) that the IEA_{emph} axis is significantly associated with prospective mortality in multivariable models, and (4) that IEAs capture distinct patterns of connection between lung structural changes and blood transcriptome patterns.

Many of our main results are consistent with our current understanding of COPD. IEAs capture the 2 cardinal pathologies of COPD, emphysema, and airway disease; but clearer links between these aspects of lung structure and blood transcriptome patterns emerge from the joint analysis of CT images and blood RNA-seq. First, neutrophilic inflammation was strongly associated with emphysema but not the airway axis. This agrees with the prominent role of neutrophils in alpha-1 antitrypsin deficiency-associated emphysema,[23] and it provides further support for the role of neutrophils in the emphysema of *typical* COPD. The IEA_{emph} axis is negatively correlated with FEV_1 and it is positively correlated with neutrophil and monocyte counts with corresponding negative correlation to lymphocyte counts. This result is consistent with previous observations that FEV_1 itself is positively correlated with lymphocyte counts, and negatively correlated with neutrophil and monocyte counts.[24] There is evidence as well for a role for specific adaptive immune processes that show significant enrichment for both IEA_{emph} and IEA_{airway}, though the inflammatory signals that we observed in blood differ from the B-cell predominated signatures that have been observed in some lung transcriptomic studies of emphysema.[25] This discrepancy is expected, because the B-cell signature in the lungs may be driven by the aggregation of B-cells in submucosal lymphoid aggregates, which would not be expected to be observable in peripheral blood samples. The biological pathway enrichments we observed are consistent with previous reports of the association of emphysema to biomarkers related to systemic inflammation, oxidative stress, and elevated plasma fibrinogen levels.[26] IEA_{airway} is strongly correlated to BMI, which coincides with a previous hypothesis that obesity-related adipose tissue hypoxia and systemic hypoxia due to reduced pulmonary function contribute to the systemic inflammation of COPD.[27] Future studies, including single-cell transcriptomic data, could better identify the association of emphysema and airway disease with specific types of innate and adaptive inflammatory cells.

While our IEAs seem most descriptive of emphysema and airway disease, they are not completely correlated to standard CT measurements of emphysema and airway disease, and they differ notably from machine-learning disease axes based on imaging

Table 4. Top-10 Significant Gene Ontology Enrichment Terms for Image-Expression Axes

GO terms	Description	Annotated	Significant	Adjusted *P*-value
Top-10 Significant Gene Ontology Enrichment Terms for IEA_{emph}				
GO:0043312	Neutrophil Degranulation	461	239	2.60e-18
GO:0019083	Viral Transcription	174	99	4.20e-17
GO:0000184	Nuclear-Transcribed mRNA Catabolic Process, Nonsense-Mediated decay	119	79	2.00e-14
GO:0006614	SRP-Dependent Cotranslational Protein Targeting to Membrane	99	65	2.20e-13
GO:0006413	Translational Initiation	185	93	2.30e-13
GO:0002181	Cytoplasmic Translation	98	51	2.20e-07
GO:0006364	rRNA Processing	224	124	6.40e-07
GO:0006954	Inflammatory Response	529	250	4.40e-06
GO:0006955	Immune Response	1758	750	2.00e-05
GO:0045730	Respiratory Burst	33	20	3.30e-05
Top-10 Significant Gene Ontology Enrichment Terms for IEA_{airway}				
GO:0006396	RNA Processing	981	204	1.80e-08
GO:0098869	Cellular Oxidant Detoxification	77	33	1.00e-06
GO:0042744	Hydrogen Peroxide Catabolic Process	24	14	1.70e-05
GO:0006879	Cellular Iron ion Homeostasis	56	25	7.30e-05
GO:0001895	Retina Homeostasis	43	16	9.10e-05
GO:0030449	Regulation of Complement Cctivation	79	30	9.20e-05
GO:0019886	Antigen Processing and Presentation of Exogenous Peptide Antigen via MHC Class II	90	32	1.00e-04
GO:0006413	Translational Initiation	185	53	1.00e-04
GO:0006614	SRP-Dependent Cotranslational Protein Targeting to Membrane	99	34	2.40e-04
GO:0050900	Leukocyte Migration	377	89	5.20e-04

GO pathway enrichment analysis was performed using the GO *biological process* gene sets with *p*-values calculated with the Fisher exact test statistic using the weight01 algorithm in top GO[21`] (v2.33.1) that accounts for dependency in GO topology.

GO=gene ontology; IEA_{emph}=emphysema image-expression axes; IEAairway=airway disease image-expression axes

alone (PCA-Is) or based on imaging and spirometry (FAs).[17] While none of these representations of emphysema and airway disease is demonstrably superior to the others, the IEA axes have a clear interpretation due to the integrative nature of the deep learning algorithm, whose goal was to find shared variability between CT images and transcriptomic patterns in the blood. The clinical relevance of these axes was demonstrated through regression models showing that IEAs were significantly associated with a wide range of COPD-related measures, including mortality. In the future, these algorithms can be extended to incorporate additional sources of molecular or imaging data.

While the IEAs primarily captured patterns of emphysema and airway-predominant COPD, they were also significantly correlated with core versus the periphery (peel) emphysema distribution. Previous work has demonstrated numerous clinically relevant associations to aspects of emphysema distribution, most notably for core/peel and apical/basal emphysema distribution,[28-32] and a machine learning analysis of images alone also identified peel-core emphysema distribution as an important dimension of COPD-related variability.[33] Our analysis suggests that blood transcriptome patterns are most strongly associated with core/peel rather than apical/basal emphysema distribution and that the amount of emphysema in the core region is associated with the more severe disease along both the IEA_{emph} and IEA_{airway} axes. Since quantification of the lung peel can be influenced by technical factors related to lung segmentation, we conducted sensitivity analyses that confirmed a consistent association for the IEA_{emph} axis, whereas the IEA_{airway} association was clearly present only when the analysis included the outermost lung regions. Accordingly, we have high confidence in the IEA_{emph} association to core/peel distribution, but it is not clear whether the IEA_{airway} axis association reflects a true biological relationship or technical artifacts.

By collecting CT scans, blood transcriptomics, and detailed phenotype data on thousands of current and former smokers enriched for COPD, the COPDGene study provides a novel opportunity for the application of machine learning to better understand the connections between the lung structure in COPD and molecular mechanisms of systemic inflammation. Like all machine learning models, the construction of our model required many explicit and implicit design choices. In our model, we extracted patch-level representations via self-supervised learning. Such methods are capable of extracting generalized and semantically meaningful features.[34] The linear independence assumption in our model was intended to identify IEAs that captured distinct underlying disease processes and increase the reproducibility of our model. Unlike previous studies that explore the relationship between COPD imaging and omics by associating previously discovered image patterns

to omics data,[22,35,36] our method potentially identifies new image patterns that have not been previously explored.

The main strengths of this study are:

1. The joint analysis of full DICOM data from CT images and gene expression data via deep learning is novel and provides new biological and clinical insight into COPD.
2. The sample size is large, allowing for more power to identify novel discoveries.
3. We used a number of techniques to improve the reproducibility of our disease axes, including cross-validation, sensitivity analysis, and the use of constraints in our modeling procedure.

The main limitations are: (1) our study is limited to blood RNA biomarkers, which capture systemic inflammation but no other important aspects of the COPD inflammatory response, such as lung gene expression and protein biomarkers, and (2) our analysis was limited to the COPDGene study. In the future, such analyses could be conducted in other ongoing studies collecting CT scan and omics data in populations enriched for COPD.

In summary, deep learning applied to CT images, and transcriptomic biomarkers in COPD identified 2 main inflammatory processes related to CT image features that can be broadly defined as emphysema and airway disease. The emphysema-related process was most enriched for pathways related to neutrophilic inflammation. The airway axis differed from previously reported disease axes learned from phenotypic data alone and it was negatively correlated with emphysema. Finally, there was also a strong relationship between the core-peel distribution of emphysema and blood transcriptome patterns. In the future, these integrative machine learning methods can be refined for more fine-grained interpretability and extended to include other sources of biological information.

Acknowledgements

Author contributions: JC, PJC, and KB designed the study. JC performed the modeling and statistical analysis and wrote the initial manuscript. ZX conducted differential expression and usage analyses. LS and KY conducted image pre-processing and feature extraction. CPH, AB, J H, FCS, EKS, and PJC assisted with the analysis of the COPDGene data. All authors contributed to the production of the final manuscript with revision for important intellectual content.

Declaration of Interest

PJC has received grant support from Bayer and consulting fees from Novartis and GSK. CPH reports grant support from Bayer, Boehringer-Ingelheim, and Vertex, and consulting fees from AstraZeneca and Takeda. EKS has received grant support from Bayer and GSK. All other authors have nothing to declare.

References

1. Adeloye D, Chua S, Lee C, et al. Global and regional estimates of COPD prevalence: systematic review and meta-analysis. *J Glob Health*. 2015;5(2):020415. https://doi.org/10.7189/jogh.05.020415

2. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study. *Lancet*. 2013;385(9963):117-171. https://doi.org/10.1016/S0140-6736(14)61682-2

3. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for prevention, diagnosis, and management of COPD: 2021 report. GOLD website. Published 2021. Accessed February 1, 2023. https://goldcopd.org/gold-reports/gold-report-2021-v1-0-11nov20_wmv/

4. Rabe K, Hurd S, Anzueto A, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2007;176(6):532-555. https://doi.org/10.1164/rccm.200703-456SO

5. Patel B, Coxson H, Pillai S, et al. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2008;178(5):500-505. https://doi.org/10.1164/rccm.200801-059OC

6. Bertrams W, Griss K, Han M, et al. Transcriptional analysis identifies potential biomarkers and molecular regulators in pneumonia and COPD exacerbation. *Sci Rep*. 2020;10(1):241. https://doi.org/10.1038/s41598-019-57108-0

7. Singh D, Fox SM, Tal-Singer R, Bates S, Riley JH, Celli B. Altered gene expression in blood and sputum in COPD frequent exacerbators in the ECLIPSE cohort. *PLos One*. 2014;9(9):e107381. https://doi.org/10.1371/journal.pone.0107381

8. Moll M, Boueiz A, Ghosh AJ, et al. Development of a blood-based transcriptional risk score for chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2022;205(2):161-170. https://doi.org/10.1164/rccm.202107-1584oc

9. Agusti A, Bel E, Thomas M, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J*. 2016;47(2):410-419. https://doi.org/10.1183/13993003.01359-2015

10. Kinney GL, Santorico SA, Young KA, et al. Identification of chronic obstructive pulmonary disease axes that predict all-cause mortality: the COPDGene study. *Am J Epidemiol*. 2018;187(10):2109-2116. https://doi.org/10.1093/aje/kwy087

11. Regan E, Hokanson J, Murphy J, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2011;7(1):32-43. https://doi.org/10.3109/15412550903499522

12. Sun L, Yu K, Batmanghelich K. Context matters: graph-based self-supervised representation learning for medical images. *Proc AAAI Conf Artif Intell*. 2021;35(6):4874-4882. https://doi.org/10.1609/aaai.v35i6.16620

13. Hinton G. Training products of experts by minimizing contrastive divergence. *Neural Comput*. 2002;14(8):1771-1800. https://doi.org/10.1162/089976602760128018

14. Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Jain S, Simon HU, Tomita E, eds. *Algorithmic Learning Theory. ALT 2005. Lecture Notes in Computer Science*. Vol 3734. Springer, Berlin, Heidelberg; 2005:63.77. https://doi.org/10.1007/11564089_7

15. Kingma D, Ba J. Adam: a method for stochastic optimization. *ArXiv*. 2017;412.6980.v9. https://doi.org/10.48550/arXiv.1412.6980

16. Cox DR. Regression models and life-tables. *J R Stat Series B Stat Methodol*. 1972;34(2):187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

17. Young KA, Regan EA, Han MK, et al. Subtypes of COPD have unique distributions and differential risk of mortality. *Chronic Obstr Pulm Dis*. 2019;6(5):400-413. https://doi.org/10.15326/jcopdf.6.5.2019.0150

18. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat*. 2016;10(2):946-963. https://doi.org/10.1214/16-AOAS920

19. Ritchie M, Phipson B, Wu D, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. https://doi.org/10.1093/nar/gkv007

20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Series B Stat Methodol*. 1995;57(1):289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

21. Alexa A, Rahnenführer J. Gene set enrichment analysis with top GO. *Bioconductor Improv*. 2009;27:1-26. https://bioconductor.org/packages/devel/bioc/vignettes/topGO/inst/doc/topGO.pdf

22. Castaldi PJ, Cho MH, San José Estépar R, et al. Genome-wide association identifies regulatory loci associated with distinct local histogram emphysema patterns. *Am J Respir Crit Care Med*. 2014;190(4):399-409. https://doi.org/10.1164/rccm.201403-0569OC

23. McCarthy C, Reeves EP, McElvaney NG. The role of neutrophils in alpha-1 antitrypsin deficiency. *Ann Am Thorac Soc*. 2016;13 Suppl 4:S297-304. https://doi.org/10.1513/AnnalsATS.201509-634KV

24. Halper-Stromberg E, Yun JH, Parker MM, et al. Systemic markers of adaptive and innate immunity are associated with chronic obstructive pulmonary disease severity and spirometric disease progression. *Am J Respir Cell Mol Biol*. 2018;58(4):500-509. https://doi.org/10.1165/rcmb.2017-0373OC

25. Faner R, Cruz T, Casserras T, et al. Network analysis of lung transcriptomics reveals a distinct B-cell signature in emphysema. *Am J Respir Crit Care Med*. 2016;193(11):1242-1253. https://doi.org/10.1164/rccm.201507-1311OC

26. Papaioannou A, Mazioti A, Kiropoulos T, et al. Systemic and airway inflammation and the presence of emphysema in patients with COPD. *Respir Med*. 2010;104(2):275-282. https://doi.org/10.1016/j.rmed.2009.09.016

27. Tkacova R. Systemic inflammation in chronic obstructive pulmonary disease: may adipose tissue play a role? Review of the literature and future perspectives. *Mediators Inflamm*. 2010;2010:585989. https://doi.org/10.1155/2010/585989

28. Mair G, Miller JJ, McAllister D, et al. Computed tomographic emphysema distribution: relationship to clinical features in a cohort of smokers. *Eur Respir J*. 2009;33(3):536-542. https://doi.org/10.1183/09031936.00111808

29. Gurney JW, Jones KK, Robbins RA, et al. Regional distribution of emphysema: correlation of high-resolution CT with pulmonary function tests in unselected smokers. *Radiology*. 1992;183(2):457-463. https://doi.org/10.1148/radiology.183.2.1561350

30. Nakano Y, Sakai H, Muro S, et al. Comparison of low attenuation areas on computed tomographic scans between inner and outer segments of the lung in patients with chronic obstructive pulmonary disease: incidence and contribution to lung function. *Thorax*. 1999;54(5):384-389. https://doi.org/10.1136/thx.54.5.384

31. Haraguchi M, Shimura S, Hida W, Shirato K. Pulmonary function and regional distribution of emphysema as determined by high-resolution computed tomography. *Respiration*. 1998;65(2):125-129. https://doi.org/10.1159/000029243

32. Boueiz A, Chang Y, Cho MH, et al. Lobar emphysema distribution is associated with 5-year radiological disease progression. *Chest*. 2018;153(1):65-76. https://doi.org/10.1016/j.chest.2017.09.022

33. 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(12):3652-3662. https://doi.org/10.1109/TMI.2021.3094660

34. Ohri K, Kumar M. Review on self-supervised image recognition using deep neural networks. *Knowl Based Syst*. 2021;224:107090. https://doi.org/10.1016/j.knosys.2021.107090

35. Cho M, Castaldi P, Hersh C, et al. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *Am J Respir Crit Care Med*. 2015;192(5):559-569. https://doi.org/10.1164/rccm.201501-0148OC

36. Jeong I, Lim J-H, Oh DK, Kim WJ, Oh Y-M. Gene expression profile of human lung in a relatively early stage of COPD with emphysema. *Int J Chron Obstruct Pulmon Dis*. 2018;13:2643-2655. https://doi.org/10.2147/COPD.S166812