Original Research

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

Junxiang Chen, PhD1 Zhonghui Xu, MS2 Li Sun, MS1 Ke Yu, MS1 Craig P. Hersh, MD, MPH2,3 Adel Boueiz, MD, MS2,3 John E. Hokanson, MPH, PhD4 Frank C. Sciurba, MD5 Edwin K. Silverman MD, PhD,2,3 Peter J. Castaldi, MD, MS2,6* Kayhan Batmanghelich, PhD1*

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

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

Running Head: Deep Learning on CT and Gene Expression in COPD

Keywords: machine learning; emphysema; genomics

Abbreviations: COPD, Chronic Obstructive Pulmonary Disease; CT, Computed Tomography; IEA, Image-Expression Axes; FEV1, Forced expiratory volume in 1 second; FVC, Forced vital
capacity; BMI, Body Mass Index; CSRL, Context-aware Self-supervised Representation Learning; MLP, Multilayer Perceptron; PoE, Product of Experts; HSIC, Hilbert-Schmidt Independence Criterion; OLS, Ordinary Least Squares; FA, Factor Analysis; PCA, Principal Component Analysis; FDR, False Discovery Rate; GO, Gene Ontology; GOLD, Global Initiative for Chronic Obstructive Lung Disease; PRISm, Preserved Ratio Impaired Spirometry; SGRQ, St. George's Respiratory Questionnaire; mMRC, Modified Medical Research Council; LAA, Low Attenuation Area

**Funding Support:** This work was supported by NHLBI 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.

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

**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; Published online July 6, 2023. [https://doi.org/10.15326/jcopdf.2023.0399](https://doi.org/10.15326/jcopdf.2023.0399)

**Note:** This article has an online data supplement.
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:** To identify novel relationships between lung structural changes measured by chest computed tomography (CT) and blood transcriptome patterns measured by blood RNA sequencing.

**Methods:** CT scan images and blood RNA-seq gene expression from 1,223 subjects in the COPDGene study were jointly analyzed using deep learning to identify shared aspects of inflammation and lung structural changes that we refer to as 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 two distinct IEAs: IEA\textsubscript{emph} captures an emphysema-predominant process with a strong positive correlation to CT emphysema and a negative correlation to FEV\textsubscript{1} and Body Mass Index (BMI); IEA\textsubscript{airway} 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\textsubscript{emph} and IEA\textsubscript{airway}, respectively (adjusted p<0.001).

**Conclusions:** Integration of CT scans and blood RNA-seq data identified two IEAs that capture distinct inflammatory processes associated with emphysema and airway-predominant COPD.
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 CT imaging and blood 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 subjects. 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 identify 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 1,223 subjects in the COPDGene study (11) with CT scans and blood gene expression data. Our analysis identified two 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).

Subject Enrollment and Data Collection

COPDGene enrolled 10,198 subjects with a minimum 10 pack-years lifetime smoking history at 21 centers across the United States (NCT00608764, www.copdgene.org) (11). Subjects 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 6,717 subjects, and 10-year follow-up visits are currently being completed. Subjects underwent spirometry, questionnaire assessments, standardized inspiratory and expiratory chest CT imaging, and genome-wide 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 (IEAs)

Only subjects 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 DICOM image files using the following procedure. Each inspiratory chest CT scan was processed into 581 patches with a size of $32^3$ mm$^3$ following the practice in (14). For each patch, 128 features were generated using context-aware self-supervised representation learning (CSRL) (12), producing a matrix of dimension 581 x 128 for each CT 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 subject-level Image-Expression Axes (IEAs) using a Product of Experts.
(PoE) model (13). At this stage, we applied statistically independent constraints with the Hilbert-Schmidt independence criterion (HSIC) (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 (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 five-fold cross-validation, giving us five models. The final IEAs were given by taking the average value of the IEAs from the five models.

**Association of IEAs 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 (OLS) models for continuous measurements and logistic regression models for categorical measurements. We conducted a survival analysis (starting from the second visit in the supplement) with the Cox proportional hazards model (16). We applied the IEA model to 1,527 subjects from another subset of the COPDGene dataset to provide independent replication of our IEA associations. These subjects 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 IEAs 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. PCA Image Only Axes (PCA-Is): Disease axes constructed by applying Principal Component Analysis (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 performed differential gene expression analysis using limma and voom(18,19). Multiple comparisons were corrected with the Benjamini-Hochberg method to control the false discovery rate (FDR) at 10%(20). Gene Ontology (GO) pathway enrichment analysis was performed for pathways in the “Biological Process” category using the TopGO (v2.33.1) method (21). The threshold for statistical significance was an adjusted P-value < 0.001.

**Results**

One thousand two hundred and twenty-three subjects in the COPDGene study with complete CT scan and blood RNA-seq data were analyzed, and the flow diagram for the selection of subjects for analysis is shown in Supplemental Figure E1. The analyzed subjects were 50% female, 82% non-Hispanic white, and 18% African American, and the average age of the subjects was 67 years. The GOLD spirometric stage distribution of subjects was 42.5% GOLD 0, 44.0% in GOLD 1-4, and 13.5% with preserved ratio impaired spirometry (PRISm). For model training and validation, subjects were split into training and test sets, with no statistically significant differences in demographic or key clinical characteristics between these groups (Table 1).
IEA 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. 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 \( p = 1 \times 10^{-6} \) 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 \( \geq 0.96 \), Supplemental Table E1). With this threshold, 4,685 genes were included in the final model. The number of IEAs identified by the model was determined by the amount of gene expression variance explained, which was the highest with two IEAs (Figure 2).

IEA 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 (Table 2). 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 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 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.

Table 2 reveals that both IEAs showed a negative association with emphysema peel/core distribution ($Q_{perc15_{peel-core}}$), 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 $IE_{emph}$ exhibits a consistently positive association with $Q_{perc15_{peel-core}}$, regardless of the band used to define it. However, it is currently unclear whether the association between $IE_{airway}$ and $Q_{perc15_{peel-core}}$ reflects a real biological phenomenon at the extreme periphery of the lung, or if it is an artifact of segmentation.

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 3, Table 4, and Supplemental Tables E2 and E3). After adjusting for standard demographic variables, both $IE_{emph}$ and $IE_{airway}$ were significantly associated with FEV₁ %predicted, FEV₁/FVC, SGRQ total score, mMRC dyspnea score, 6-minute-walk distance, as well as neutrophil count. Additionally, $IE_{emph}$ was also associated with Frequent Exacerbator (History), Frequent Exacerbator (Future), all-cause mortality rate, monocyte proportion, and eosinophil count. On the other hand, $IE_{airway}$ is associated with white blood cell count, neutrophil proportion, lymphocyte count and lymphocyte proportion.

To provide independent replication of our IEA associations, the IEA model was applied to 1,527 subjects 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).

**COPD Subgroups Defined by IEAs and Comparison to Existing COPD Subtypes**

To further understand the clinical characteristics of COPD subgroups defined by IEAs, we divided the IEA space into four quadrants (Figure 3) and computed the average characteristics of each subgroup (Supplemental Table E8). As expected, subjects with low IEA\textsubscript{emph}/low IEA\textsubscript{airway} values had the least obstruction (mean FEV1 89.2% predicted), the highest percentage of GOLD spirometric grade 0 subjects, low emphysema, and the thinnest airways. Subjects with high IEA\textsubscript{emph}/low IEA\textsubscript{airway} values had characteristics consistent with emphysema-predominant COPD, namely high emphysema and low BMI, with about 70% of GOLD grade 4 subjects present in this group. Subjects with low IEA\textsubscript{emph}/high IEA\textsubscript{airway} values had an airway-predominant profile with thick airway walls, elevated BMI, and the greatest proportion of PRISm subjects. Subjects with high IEA\textsubscript{emph}/high IEA\textsubscript{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 one year) but did not change in FEV1.

The group with the highest mortality was high IEA\textsubscript{emph}/high IEA\textsubscript{airway} followed by high IEA\textsubscript{emph}/low IEA\textsubscript{airway}. The latter group also had the highest percentage of subjects with frequent exacerbations, both for retrospective (p<0.001, chi-square test of all groups) and prospective exacerbations (p=0.048).

To place IEA\textsubscript{emph} and IEA\textsubscript{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 factor-analysis derived disease axes (FAs) (17). In Figure 4, we observe that the highest values of IEA\textsubscript{emph} are found in the severe emphysema k-means subtype, and the highest values of IEA\textsubscript{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 (FAemph) and airway-predominant disease (FAairway), we compared the IEAs to the FAs and observed that IEAemph and FAemph showed a reasonably strong correlation (Pearson’s r = 0.58), but the IEAairway and FAairway axes showed only modest correlation (Pearson’s r = 0.28, see Supplemental Table E9).

Examination of the pattern of clinical associations for IEAairway and FAairway revealed that while airway axes were positively correlated to airway wall thickness, IEAairway is negatively correlated to emphysema whereas FAairway is positively correlated (Supplemental Table E10). To determine whether the IEAs provided additional information about COPD phenotypes (FEV1 % of predicted, FEV1/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).

Comparison to PCs 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 PCA to extract the top-2 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 six studied COPD phenotypes as well as prospective exacerbations and mortality (p < 0.001, Supplemental Tables E13 and E14). 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.

**IEAs 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 that PCA-Is, as demonstrated in Figure 5, 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 6,494 and 3,815 genes associated at an FDR of 10\% with IEA\textsubscript{emph} and IEA\textsubscript{airway}, respectively (Supplemental Tables E15 and E16). Gene Ontology pathway enrichment identified 29 and 13 enriched pathways (p-value < 0.001) for IEA\textsubscript{emph} and IEA\textsubscript{airway}, respectively (Supplemental Tables E17 and E18). The most significantly associated pathway for IEA\textsubscript{emph} was neutrophil degranulation, whereas IEA\textsubscript{airway} had the strongest enrichment for RNA processing. The most significant pathway results are shown in Table 5.

**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, and 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 two 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 associated (AAT) 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 FEV₁ 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 FEV₁ 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 lung 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 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 is 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. 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 two 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

The authors meet criteria for authorship as recommended by the International Committee of Medical Journal Editors. J.C., P.J.C, and K.B. designed the study. J.C performed the modeling and statistical analysis and wrote the initial manuscript. Z.X. conducted differential expression and usage analyses. L.S, and K.Y conducted image pre-processing and feature extraction. C.P.H, A.B, J. H., F.C.S, E.K.S and P.J.C assisted with analysis of 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.
References


### Table 1. Subject characteristics in training and test data.

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. $FEV_1 =$ Forced expiratory volume in 1 second; $FVC =$ Forced vital capacity; perc15 = 15th Percentile.
Hounsfield unit in Inspiratory CT scan; %Gas Trapping: %LAA 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. 

\[ Q_{perc15_{peel-core}} = 100 \times \log\left(\frac{perc_{15_{peel}}}{perc_{15_{core}}}\right), \]

where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary. \( \Delta FEV_1 \%_{predicted} \) and \( \Delta FEV_1/FVC \) are computed by subtracting the visit 3 values from the visit 2 values of \( FEV_1 \%_{predicted} \) or \( FEV_1/FVC \) and dividing them by the number of years between the two visits.
Table 2. Pearson’s correlation between image-expression axes (IEAs) and COPD-related characteristics and health outcomes. The symbols “*”, “**”, and “***” represent p<.05, p<.01, and p<.001, respectively.

FEV₁ = Forced expiratory volume in 1 second; FVC = Forced vital capacity; perc15 = 15th Percentile Hounsfield unit in Inspiratory CT scan; %Gas Trapping: %LAA 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; $Q_{perc15_{peel-core}} =$
100 \log_2(\text{perc15}_{\text{peel}} / \text{perc15}_{\text{core}}), where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary; \(\Delta FEV_1 \%\text{predicted}\) and \(\Delta FEV_1/FVC\) are computed by subtracting the visit 3 values from the visit 2 values of \(FEV_1 \%\text{predicted}\) or \(FEV_1/FVC\) and dividing it by the number of years between the two visits.
### Table 3. Multivariable associations of image-expression axes (IEAs) to continuous COPD-related characteristics and health outcomes.

The table reports the β coefficients and corresponding 95% confidence intervals for IEA\textsubscript{emph} and IEA\textsubscript{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. The symbols “*”, “**”, and “***” represent p<.05, p<.01 and p<.001, respectively. FE\textsubscript{1}V = Forced expiratory volume in 1 second; FVC = Forced vital capacity; perc\textsubscript{15} = 15th Percentile Hounsfield unit in Inspiratory CT.
scan; %Gas Trapping: %LAA 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; \( Q_{perc_{15\text{peel-core}}} = 100 \log (perc_{15\text{peel}} / perc_{15\text{core}}) \), where the peel region is defined to be <5mm from the lung boundary and the core region is >20mm from the lung boundary. \( \Delta FEV_1 \% \text{predicted} \) and \( \Delta FEV_1/FVC \) are computed by subtracting the visit 3 value from the visit 2 value of \( FEV_1 \% \text{predicted} \) or \( FEV_1/FVC \) and dividing it by the number of years between the two visits.

### Logistic Regression Model

<table>
<thead>
<tr>
<th>Logistic Regression Model</th>
<th>( \beta_{IEA_{\text{mph}}} ) (95%CI)</th>
<th>Odds Ratio (95%CI)</th>
<th>( \beta_{IEA_{\text{airway}}} ) (95%CI)</th>
<th>Odds Ratio (95%CI)</th>
</tr>
</thead>
</table>

---

Copyright Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation ©2023
Published online July 7, 2023  https://doi.org/10.15326/jcopdf.2023.0399
Table 4. Multivariable associations of image-expression axes (IEAs) to frequent exacerbations and mortality. The table reports the $\beta$ 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, smoking status as the covariates. The symbols “*”, “**”, and “***” represent $p<.05$, $p<.01$, and $p<.001$, respectively. Frequent exacerbator (history) indicated whether the subject had at least two self-reported exacerbations during the 12 months before the second visit. Frequent exacerbator (future) indicated whether the subject had at least two self-reported exacerbations during the past 12 months before the third visit.
Table 5. Top-10 Significant Gene Ontology enrichment terms for IEAs. Gene ontology (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 topGO (v2.33.1) that accounts for dependency in GO topology.
Figure 1. Overview of the machine learning workflow. CT images were processed as $32^3 \text{mm}^3$ patches from which 128 features were constructed using the Context-Aware Self-supervised Representation learning algorithm (CSRL) (12). These features were the input for a multilayer perceptron (MLP) that processed a $581 \times 128 \times 923$ tensor to output a $581 \times 2 \times 923$ subject-level data representation. The subject-level latent representation (IEAs) is given by summarizing the patch-level features into a matrix of $2 \times 923$. We introduce a linear layer ($2 \rightarrow 4,685$) that estimates gene expression for each subject, 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.
Figure 2. Total variance of the gene expression explained vs. number of IEAs. The figure on the left shows the plot of the 4,685 selected genes. The figure on the right shows the plot for all the genes. The figures show that when the number of IEAs is two, the total variance explained is maximized. We choose the number of IEAs to be two.
Figure 3. Visualization of subjects projected along each identified image-expression axis (IEA) 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 four subgroups based on dividing the IEA space into four quadrants. Lung CT scans and clinical characteristics are shown for one subject in each quadrant, where the red mask represents the emphysema regions (< -950 HU). The characteristics of the four quadrants are summarized in Supplemental...
Figure 4. Distribution of $\text{IEA}_{\text{emph}}$ and $\text{IEA}_{\text{airway}}$ values grouped by previously published COPD K-means clustering subtypes\(^{(22)}\). P-values are obtained using the Kruskal-Wallis test. The symbols “*”, “**”, and “***” represent $p<.05$, $p<.01$, and $p<.001$, respectively, for a t-test comparing to the Relatively Resistant Smokers Group. No symbol indicates non-significant results.

RRS: Relatively resistant smokers; UPE: Mild upper zone-predominant emphysema; AP: Airway-predominant; SE: Severe emphysema.
Figure 5. Histograms for the variances of genes explained ($R^2$) by the image-expression axes (IEAs) and PCA Image Only Axes (PCA-Is). The figure on the left shows the histogram for the 4,685 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.
Online Supplement

To access the online supplement: