BMJ Open Respiratory Research

Utility of peripheral protein biomarkers for the prediction of incident interstitial features: a multicentre retrospective cohort study

Samuel Ash ,[1,2] Tracy J Doyle ,[3] Bina Choi,[4] Ruben San Jose Estepar,[5] Victor Castro,[6] Nicholas Enzer,[4] Ravi Kalhan,[7] Gabrielle Liu,[8] Russell Bowler,[9] David O Wilson,[10] Raul San Jose Estepar,[11] Ivan O Rosas,[12] George R Washko[13]

To cite: Ash S, Doyle TJ, Choi B, *et al*. Utility of peripheral protein biomarkers for the prediction of incident interstitial features: a multicentre retrospective cohort study. *BMJ Open Respir Res* 2024;**11**:e002219. doi:10.1136/bmjresp-2023-002219

SA and TJD contributed equally.
RSJE and IOR contributed equally.

SA and TJD are joint first authors.
RSJE and IOR are joint senior authors.

Received 27 November 2023
Accepted 28 February 2024

For numbered affiliations see end of article.

Correspondence to
Dr Samuel Ash;
sash@southshorehealth.org

ABSTRACT

Introduction/rationale Protein biomarkers may help enable the prediction of incident interstitial features on chest CT.

Methods We identified which protein biomarkers in a cohort of smokers (COPDGene) differed between those with and without objectively measured interstitial features at baseline using a univariate screen (t-test false discovery rate, FDR p<0.001), and which of those were associated with interstitial features longitudinally (multivariable mixed effects model FDR p<0.05). To predict incident interstitial features, we trained four random forest classifiers in a two-thirds random subset of COPDGene: (1) imaging and demographic information, (2) univariate screen biomarkers, (3) multivariable confirmation biomarkers and (4) multivariable confirmation biomarkers available in a separate testing cohort (Pittsburgh Lung Screening Study (PLuSS)). We evaluated classifier performance in the remaining one-third of COPDGene, and, for the final model, also in PLuSS.

Results In COPDGene, 1305 biomarkers were available and 20 differed between those with and without interstitial features at baseline. Of these, 11 were associated with feature progression over a mean of 5.5 years of follow-up, and of these 4 were available in PLuSS, (angiopoietin-2, matrix metalloproteinase 7, macrophage inflammatory protein 1 alpha) over a mean of 8.8 years of follow-up. The area under the curve (AUC) of classifiers using demographics and imaging features in COPDGene and PLuSS were 0.69 and 0.59, respectively. In COPDGene, the AUC of the univariate screen classifier was 0.78 and of the multivariable confirmation classifier was 0.76. The AUC of the final classifier in COPDGene was 0.75 and in PLuSS was 0.76. The outcome for all of the models was the development of incident interstitial features.

Conclusions Multiple novel and previously identified proteomic biomarkers are associated with interstitial features on chest CT and may enable the prediction of incident interstitial diseases such as idiopathic pulmonary fibrosis.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Several protein biomarkers have been associated with interstitial lung disease prognosis, but less is known about their role in predicting the development of interstitial lung disease.

WHAT THIS STUDY ADDS

⇒ This study identifies peripheral protein biomarkers associated with the presence and progression of subtle changes on chest CT, which suggest early interstitial lung disease. These biomarkers may be used to predict incident interstitial feature development.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings may not only help with the prediction of incident interstitial lung disease development but also help identify new pathways for further research.

INTRODUCTION

Over the past several decades, it has been increasing recognised that subtle evidence of chronic lung injury is visible on CT scans of the chest.[1] More specifically, based on their shared clinical and genetic associations, these areas of higher attenuation tissue, often referred to visually as interstitial lung abnormalities (ILAs), likely represent early or subtle evidence of pulmonary fibrosis in some people.[2–4] Our group and others have demonstrated that these abnormalities can also be detected using a variety of automated, machine learning-based tools.[5 6] Because these imaging findings are similar to but not exactly equivalent to visually defined ILA, we have termed them quantitative interstitial features, and we have previously shown that they share the same clinical and genetic associations as ILA and idiopathic pulmonary fibrosis (IPF) such as lower lung function and the *MUC5B* promoter mutation, suggesting that they too may represent early evidence of fibrosis in some people.[5 7–10]

However, while the associations between interstitial features and clinical outcomes are well described, the prediction of interstitial

feature development is less so. Prediction of interstitial feature development is especially important because although interstitial features are considered a possible precursor to IPF, as noted above, interstitial features alone, even in the absence of advanced fibrosis, are associated with adverse clinical outcomes, and the only currently available pharmacologic interventions for pulmonary fibrosis slow its progression but do not reverse prior damage.[4 5 7 8] Of particular interest for the prediction of the development of interstitial features is the utility of peripheral protein biomarkers to predict incident disease, both because of their potential use for identifying high-risk clinical populations and because they may identify specific, targetable pathways that could be used to prevent disease development and progression.[9 10] In this study, we sought to identify peripheral protein biomarkers associated with interstitial features and use those biomarkers combined with baseline imaging and demographics to create machine learning models to predict their incident development.

METHODS

Study population

We performed the primary analyses using data from the COPDGene cohort and confirmatory analyses in the Pittsburgh Lung Screening Study (PLuSS) cohort. Both cohorts have been described in detail previously.[11–13]

Briefly, COPDGene is a multicentre, prospective, cohort study of over 10 300 ever smokers who at enrollment were aged 45–80, had at least a 10 pack-year smoking history and did not have prior bronchiectasis or interstitial lung diseases (ILDs) such as IPF. Initial (baseline/phase 1) visits occurred between 2006 and 2011, and 5-year follow-up (phase 2) visits occurred between 2013 and 2017. 10-year follow-up visits are currently ongoing and not included in these analyses. At both phase 1 and phase 2 visits, participants underwent inspiratory and expiratory chest CT scans, prebronchodilator and postbronchodilator spirometric testing, 6 min walk distance measurements, questionnaires and genotyping of the *MUC5B* polymorphism (rs35705950).[14 15] CT scans were obtained at inspiration (200 mAs) and after expiration (50 mA) with submillimeter slice reconstruction.[11]

PLuSS is a single-centre, prospective, cohort study of approximately 3800 ever smokers who at enrollment were aged 50–79, had at least a 12.5 pack-year smoking history and did not have a history of prior lung cancer. Participants underwent baseline inspiratory low-dose chest CT scans (40–60 mA with 2.5 mm reconstruction), spirometric testing and questionnaires between 2002 and 2005. PLuSS participants were subsequently followed with serial CT imaging as indicated based on the study lung cancer screening protocol, as well as with spirometry, annual telephone surveys and/or mailed questionnaires.[13] Although many participants in the PLuSS cohort had a several CT scans, in order to approximate the COPGene cohort, only participants first and last CT scan were included, and participants were excluded if they had less than 2 years of follow-up. Due to the use of previously obtained, anonymised data, patients were not involved in the design of this current study.

For both studies, only participants with complete baseline clinical data, at least one follow-up CT imaging study, and baseline protein biomarker data were included (figure 1).

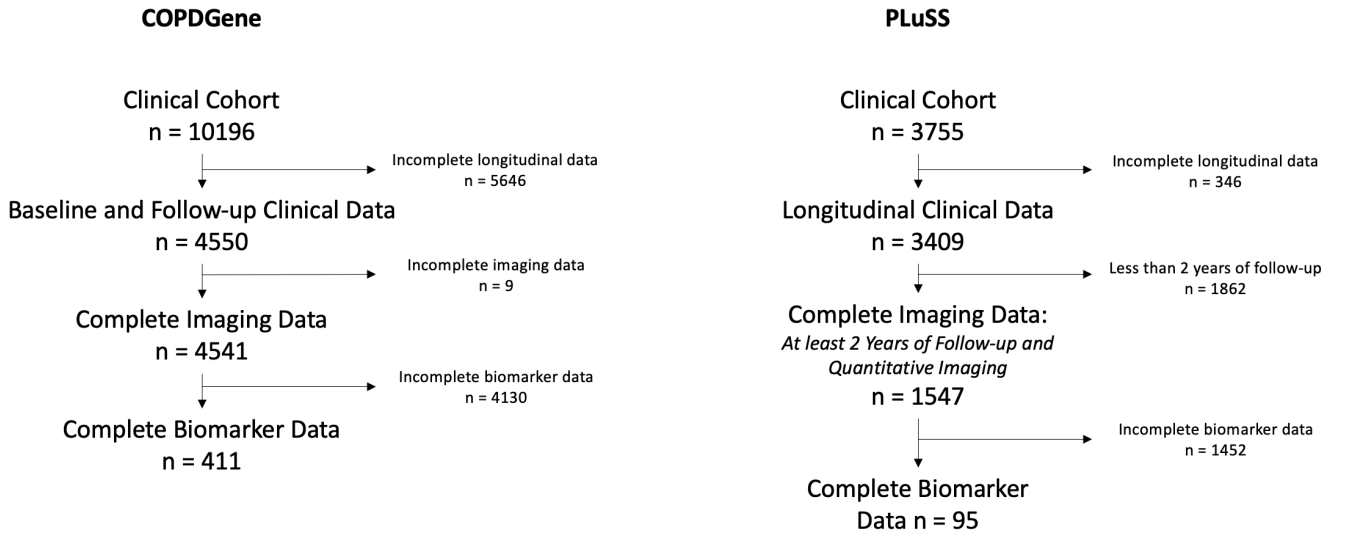

Figure 1 CONSORT diagram. CONSORT, Consolidated Standards of Reporting Trials; PLuSS, Pittsburgh Lung Screening Study.

Patient and public involvement

The public was not involved in the design of this specific study. All data resulting from this work will be made publicly available via dbGaP (https://www.ncbi.nlm.nih.gov/gap/).

Image and biomarker analysis

The percentage of lung occupied by interstitial features was measured using a local density classification approach, which uses a previously described, k-nearest neighbours classifier-based approach that uses the local histogram measurements combined with the distance from the pleural surface.[14 16 17] Peripheral protein biomarkers were measured in a subset of participants at the phase 1 and phase 2 visits in the COPDGene study using the SOMASCan 1.3K assay (SomaLogic Operating Company, Boulder, Colorado, USA). This technique has been described in detail previously. Briefly, it is an aptamer-based assay that enables the simultaneous measurement of a broad range of protein targets.[18] For this study, only the measurements at phase 1 (baseline) were used and a total of 1305 protein biomarkers were available. In the PLuSS study, peripheral protein biomarkers were measured at baseline using the Myriad Rules-Based Medicine (RBM) system (Luminex xMap technology, Myriad-RBM, Austin, Texas, USA), and 116 protein biomarkers were available. Due to their non-normal distribution, all protein biomarker values were log-transformed.[19]

Statistical analysis

All the analyses, apart from the validation of the final machine learning prediction model described below, were performed using data from the COPDGene cohort. To identify peripheral protein biomarkers associated with interstitial features, we first performed a univariate screen comparing those with interstitial features to those without interstitial features at the phase 1 visit. For the purposes of this analysis, participants were defined as having interstitial features if the percentage of their lung occupied by interstitial features was greater than the median percentage in the cohort. Student's t-tests were used to compare the biomarker levels in those with and without interstitial features. In order to limit the number of biomarkers selected to a smaller, potentially clinically relevant subset, only those with false discovery rate (FDR) p<0.001 were considered.[20–22] Those biomarkers found to be significant were then each used in separate, multivariable, mixed effects models in order to determine which were associated with longitudinal changes in interstitial features from phase 1 to phase 2. These models included all the participants in the COPDGene cohort and were each adjusted for age, gender, race, current smoking status, pack-years, body mass index and forced vital capacity, as well as random effects for subject, clinical centre and CT scanner model. Biomarkers with FDR p<0.05 for this multivariable confirmation step were considered significant.[22]

To determine the utility of peripheral protein biomarkers to predict incident interstitial features, we selected the subset of individuals who were in the lowest tertile of interstitial features at baseline, and defined incident interstitial feature development as moving to the highest tertile of interstitial features at follow-up. We then trained four random forest classifiers: the first using only clinical and imaging features associated with the development of pulmonary fibrosis (age, gender, smoking status, pack-years and baseline interstitial features) (termed the clinical/imaging model), the second using the clinical/imaging values plus the protein biomarkers identified in the univariate screen (univariate screen model), the third using the clinical/imaging values plus the protein biomarkers from the multivariable confirmation (multivariable confirmation model) and the fourth using the clinical/imaging values plus the protein biomarkers from the multivariable confirmation that were also available in PLuSS (limited multivariable confirmation model).[23] These models were trained in a two-thirds random subset of COPDGene. The first three models were evaluated in the remainingone-third of COPDGene (ie, the testing portion). The final model was evaluated in both the testing portion of COPDGene and in PLuSS. 10-fold cross-validation was used to tune model hyperparameters. Model performance was summarised using the area under the receiver operating characteristic curve (AUC), and feature importance was evaluated based on impurity (Gini importance).[24] All continuous predictors were normalised, all statistical tests were two sided unless otherwise stated, and all analyses were performed in R V.4.0.3, implemented using RStudio.[25 26]

RESULTS

Of the 10 196 participants in COPDGene, 4550 had complete clinical follow-up data, 4541 had complete longitudinal imaging data and 411 had complete protein biomarker data. Of the 3755 PLuSS participants, 3409 had complete longitudinal clinical data, 1547 had complete longitudinal imaging data, and 95 had complete protein biomarker data (table 1 and figure 1). At the baseline visit, participants in the COPDGene cohort were generally younger (mean age=62.5±8.6) than in the PLuSS cohort (mean age=64.4±73.1). This subset of COPDGene had a slight female predominance (n=219 (53.3%)) compared with the PLuSS cohort where the minority of participants were female (n=21 (22.1%)).

Of the 1305 protein biomarkers available in the COPDGene cohort, 20 were different between those with and without interstitial features at baseline. These included angiopoietin 2 (Ang2), apolipoprotein A-I (Apo-A1), matrix metalloproteinase 7 (MMP7), follicle stimulating hormone, macrophage inflammatory protein 1 alpha (MIP-1alpha), pulmonary and activation regulated chemokine (PARC), pleiotrophin, cathepsin B, retinoic acid receptor responder protein 2 (RARRES2), coiled-coil domain-containing protein 80 (CCDC80),

Table 1 Baseline characteristics of the cohort

COPDGene			
	Subgroup	**Visit 1**	**Visit 2**
n		411	411
Age (years) (mean (SD))		62.5 (8.6)	68.0 (8.6)
Gender (%)	Male	192 (46.7)	192 (46.7)
	Female	219 (53.3)	219 (53.3)
Race (%)	White	389 (94.6)	389 (94.6)
	Black	22 (5.4)	22 (5.4)
Smoking status (%)	Former smoker	282 (68.6)	312 (76.3)
	Current smoker	129 (31.4)	97 (23.7)
Pack-years (mean (SD))		41.9 (24.7)	42.7 (25.2)
Body mass index (kg/m^2) (mean (SD))		29.2 (5.7)	29.1 (6.2)
Time from baseline visit (years) (mean (SD))		–	5.5 (0.7)
Percentage of lung occupied by interstitial features (median (IQR))		4.4 (2.9–6.9)	4.0 (2.2–6.5)
Percentage of lung occupied by emphysema (median (IQR))		0.9 (0.2–6.9)	0.4 (0.1–4.8)
Percentage of lung occupied by normal parenchyma (median (IQR))		92.7 (83.9–95.8)	93.6 (83.6–96.5)
PLuSS			
	Subgroup	**First visit**	**Last visit**
n		95	95
Age (years) (mean (SD))		64.4 (6.6)	73.1 (6.2)
Gender (%)	Female	21 (22.1)	21 (22.1)
	Male	74 (77.9)	74 (77.9)
Race (%)	American Indian/Alaskan Native	0 (0.0)	0 (0.0)
	Asian	0 (0.0)	0 (0.0)
	Black	5 (5.3)	5 (5.3)
	Pacific Islander	0 (0.0)	0 (0.0)
	White	90 (94.7)	90 (94.7)
Smoking status (%)	Former smoker	43 (45.3)	Unavailable
	Current smoker	52 (54.7)	Unavailable
Pack-years (mean (SD))		76.0 (25.1)	Unavailable
Body mass index (kg/m^2) (mean (SD))		28.0 (5.3)	27.7 (5.1)
Time from baseline visit (years) (mean (SD))		–	8.8 (2.7)
Percentage of lung occupied by interstitial features (median (IQR))		14.4 (11.0–18.8)	14.8 (11.7–17.8)
Percentage of lung occupied by emphysema (median (IQR))		10.1 (6.9–12.6)	11.9 (7.7–13.8)
Percentage of lung occupied by normal parenchyma (median (IQR))		68.2 (62.7–72.6)	66.4 (62.8–71.1)

Table 2 Protein biomarkers that differ by percentage of interstitial features at baseline in COPDGene

Biomarker	Mean in those with less interstitial features	Mean in those with more interstitial features	FDR p value
Angiopoietin 2	5.2	5.28	7.25×10^{-5}
Apolipoprotein A-I	9.81	9.76	1.41×10^{-4}
Matrix Metalloproteinase 7	7.76	7.88	2.08×10^{-4}
Follicle stimulating hormone	7.43	7.77	6.69×10^{-5}
Macrophage inflammatory protein 1 alpha	6.69	6.8	$2.46*10^{-5}$
Pulmonary and activation regulated chemokine	8.8	8.93	2.46×10^{-5}
Pleiotrophin	7.36	7.43	2.49×10^{-4}
Cathepsin B	7.35	7.44	2.46×10^{-5}
Retinoic acid receptor responder protein 2	7.97	8.03	5.03×10^{-5}
Coiled-coil domain-containing protein 80	7.65	7.72	1.53×10^{-4}
Cystatin-M	8.77	8.68	2.56×10^{-4}
Carbonic anhydrase 6	8.47	8.32	4.40×10^{-4}
Growth/differentiation factor 15	7.23	7.35	3.10×10^{-5}
Macrophage metalloelastase	7.24	7.4	2.46×10^{-5}
Prothrombin	11.92	11.88	2.40×10^{-4}
Fatty-acid-binding protein	9.75	9.88	1.53×10^{-4}
Prostate-specific antigen	6.88	6.66	1.53×10^{-4}
Fibulin 3	7.46	7.51	5.35×10^{-4}
Leptin	8.81	9.18	2.13×10^{-9}
Galectin 9	7.13	7.23	2.02×10^{-4}

FDR, false discovery rate.

cystatin-M, carbonic anhydrase 6, growth/differentiation factor 15 (GDF-15), macrophage metalloelastase (MMP12), prothrombin, fatty-acid-binding protein, prostate-specific antigen, fibulin 3, leptin and galectin-9 (table 2).

Of these 20 protein biomarkers identified in the univariate screen, 11 were associated with longitudinal changes in interstitial features: Ang2, MMP7, MIP-1alpha, PARC, pleiotrophin, cathepsin B, RARRES2, GDF-15, MMP12, fibulin 3 and galectin-9 (table 3). Of these, four were available in the PLuSS cohort dataset: Ang2, MMP7, MIP-1alpha and PARC.

Regarding the incident development of interstitial features, for COPDGene, individuals were defined as having incident interstitial features at follow-up if they were in the lowest tertile of interstitial features at baseline (≤3.5%) and in the highest tertile of interstitial features at follow-up (>6.1%). Similarly for participants in the PLuSS cohort, individuals were defined as having incident interstitial features at follow-up if they were in the lowest tertile of interstitial features at baseline (≤12.3%) and in the highest tertile of interstitial features at follow-up (>16.4%). The random forrest classifier trained using only imaging and clinical features showed relatively poor discrimination for predicting incident interstitial features both in the testing subset of COPDGene and in PLuSS: AUC=0.69 and 0.59, respectively (figure 2). By contrast, the classifiers that included biomarker data all had relatively good discrimination for predicting incident interstitial features. For example, the classifier trained using clinical and imaging features plus all 20 protein biomarkers from the univariate screen had an AUC=0.78 in the testing subset of COPDGene, and the classifier trained using the clinical and imaging features plus the 11 protein biomarkers form the multivariable longitudinal associations had an AUC=0.76 in COPDGene. Finally, the classifier trained using the clinical and imaging features plus the 4 of those 11 protein biomarkers available in the PLuSS cohort had an AUC=0.75 in the testing subset of COPDGene and an AUC=0.76 in PLuSS (figure 3).

The relative feature importance for each of the four classifiers is shown in figure 4. Of note, while the imaging feature is consistently one of the most important features, several of the protein biomarkers are consistently among the more important features as well.

DISCUSSION

In this observational cohort study, we identified several peripheral protein biomarkers associated with the presence and progression of interstitial features, or subtle

Table 3 Longitudinal associations between protein biomarkers and interstitial features in COPDGene

		CI		FDR p value
Protein	**Change**	**Lower**	**Upper**	
Angiopoietin 2	2.55	1.466	4.436	1.86×10^{-2}
Apolipoprotein A-I	0.824	0.36	1.884	1×10^{0}
Matrix metalloproteinase 7	3.121	2.237	4.353	2.06×10^{-9}
Follicle stimulating hormone	0.755	0.6	0.952	3.47×10^{-1}
Macrophage inflammatory protein 1 alpha	2.873	1.928	4.281	4.67×10^{-6}
Pulmonary and activation regulated chemokine	3.045	2.068	4.484	3.92×10^{-7}
Pleiotrophin	0.392	0.326	0.471	3.94×10^{-21}
Cathepsin B	4.481	2.591	7.749	1.77×10^{-6}
Retinoic acid receptor responder protein 2	3.656	1.721	7.767	1.52×10^{-2}
Coiled-coil domain-containing protein 80	2.365	1.247	4.486	1.7×10^{-1}
Cystatin-M	0.74	0.466	1.175	1×10^{0}
Carbonic anhydrase 6	0.735	0.55	0.982	7.54×10^{-1}
Growth/differentiation factor 15	4.058	2.667	6.175	1.55×10^{-9}
Macrophage metalloelastase	1.751	1.25	2.454	2.31×10^{-2}
Prothrombin	0.376	0.149	0.95	7.76×10^{-1}
Fatty-acid-binding protein	0.929	0.599	1.441	1×10^{0}
Prostate-specific antigen	1.075	0.822	1.407	1×10^{0}
Fibulin 3	10.291	5.322	19.9	1.13×10^{-10}
Leptin	1.342	0.993	1.812	1×10^{0}
Galectin 9	2.239	1.475	3.399	3.16×10^{-3}

Longitudinal change in protein biomarker effect size is expressed as the change between visits in interstitial features per a 1 unit change in protein biomarker level between the visits with adjustments for age, gender, race, current smoking status, pack-years, body mass index and forced vital capacity, as well as random effects for subject, clinical centre and CT scanner model.

changes objectively measured on CT scans of the chest. In some people, these changes may represent early ILDs such as pulmonary fibrosis.[5] In addition, we demonstrated that these biomarkers can be used in conjunction with clinical and imaging features, such as the percentage of lung occupied by interstitial features on chest CT, to predict the incident development of new interstitial features over 5 years of follow-up.

One of the most interesting findings from this study is the performance of the machine learning classifier for predicting interstitial feature development in an entirely independent cohort and using only a limited number of protein biomarkers. This performance is particularly striking given the differences in clinical characteristics, imaging protocol and biomarker measurement system between the two cohorts: COPDGene and PLuSS.[11 13] While they are both research cohorts, COPDGene involves visits at 5-year time points, imaging using standard dose CT scans and protein biomarkers measured using SOMAlogic.[18] By contrast, PLuSS participants underwent low-dose CT scans as indicated by lung cancer screening protocols and their protein biomarkers were measured using the Myriad-RBM system. These findings suggest that this type of approach may be robust to differences in data generation. This is of particular interest given the eventual hope to apply this type of work to more heterogeneous, clinically acquired data.

The specific protein biomarkers identified as being associated with interstitial features in this study are also of interest. Reassuringly, several of the biomarkers identified have been previously shown to be associated with ILDs such as pulmonary fibrosis. For example, MMP7 has been shown to be associated with both advanced pulmonary fibrosis as well as similar, potentially early evidence of ILDs.[10 27 28] However, while several of the proteins have been shown to be important in animal models or in later stage disease, less is known about their role in early disease. For example, PARC, which not only was associated with interstitial feature progression, but also was an important protein biomarker for interstitial feature prediction based on feature importance. It has been shown to be associated with pulmonary fibrosis in laboratory models and to be associated with pulmonary fibrosis in patients with rheumatological diseases such as systemic sclerosis and rheumatoid arthritis, but its role in early fibrosis is less clear.[29 30] Similarly, MIP-1alpha has been shown to be important in pulmonary fibrosis, and, in fact, the use of a novel chemokine binding protein, evasin-1, has been shown in animal models to decrease bleomycin-induced pulmonary fibrosis.[31] Our findings,

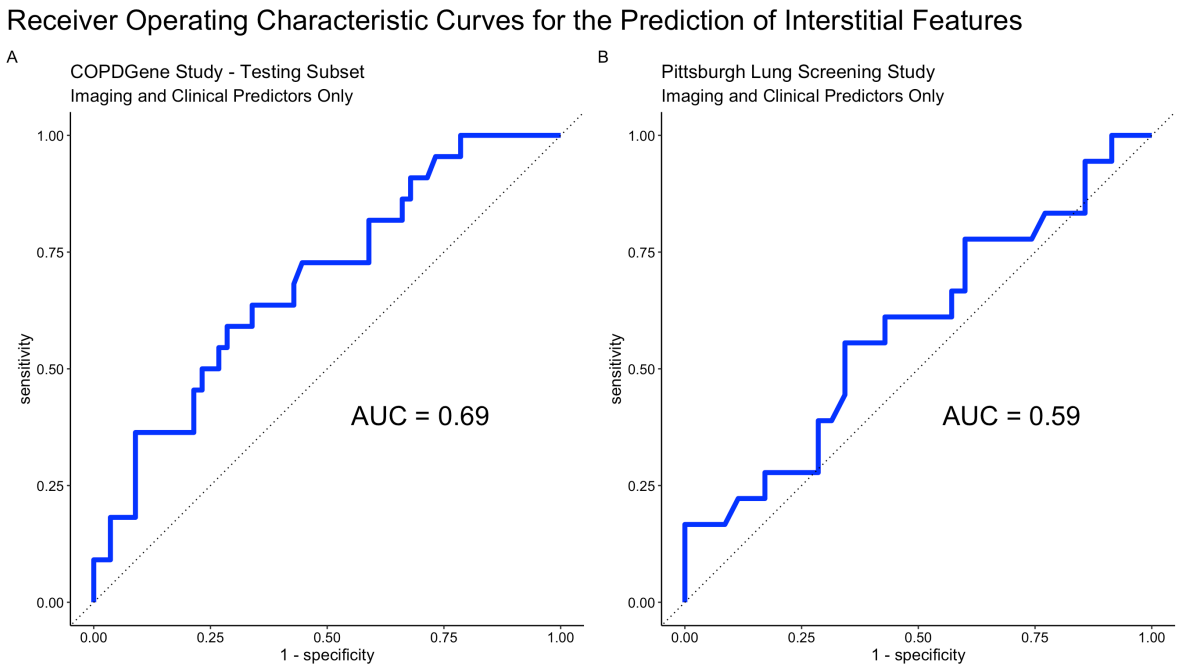

Figure 2 Receiver operating characteristic curves for random forest classifier trained using clinical and imaging features only. Receiver operating characteristic curves for (A) the random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) applied to the testing subset of COPDGene. (B) The random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) applied to all of the available complete data from PLuSS. AUC, area under the curve; PLuSS, Pittsburgh Lung Screening Study.

combined with this information, suggest that there may be a role for similar such therapies for the prevention of the progression or even the development of ILDs like pulmonary fibrosis. These results also add to the growing literature surrounding the use of precision medicine in multiple diseases such as pulmonary fibrosis, acute respiratory distress syndrome and severe COVID-19, all of which may share common biomarker risk factors in certain individuals.[10 32]

Finally, it should also be noted that even at the lowest amounts of interstitial features that is, among participants in the lowest tertile of interstitial features at baseline, the percentage of interstitial features still predicts incident disease. This suggests that truly any evidence of abnormality may indicate susceptibility. This is particularly important as we begin to consider therapeutics for those patients with more subtle imaging changes such as quantitative interstitial features as well as visually apparent ILAs. It may be that combining imaging a few select biomarkers may help enable predicting those at highest risk for progression and therefore those most likely to benefit from novel and existing therapies.

Our study has several limitations. For example, although both cohorts are quite large, the actual number of individuals with complete data, especially biomarker data is quite small, especially after subsetting, potentially making these findings more difficult to extend to a broader population. There were also many participants without sufficient longitudinal data, raising the concern for survival bias, and compared with our prior work using these cohorts the subset of individuals with complete data in this study were slightly older and more likely to be former rather than current smokers, raising concern for selection bias.[17 33] The lack of racial diversity in both cohorts is also of concern. The COPDGene cohort only included participants who identified as either White or Black. The PLuSS cohort included other racial groups in the larger study, but only white and black participants had complete data available for this current work, potentially introducing selection bias. Similarly, for this work, we elected to not separate participants by gender, potentially limiting the utility of certain biomarkers such as prostate-specific antigen. Prior work on systemic sclerosis-associated ILD has suggested that the biological profiles of the disease may differ between men and women.[34] Future work will be required to determine if biomarker predictors may vary by gender and/or sex in early pulmonary fibrosis, and how such differences may impact outcome prediction. Other limitations included the definitions of disease and its progression. As is the case with any new disease measurement, it is difficult to define what an abnormal amount of interstitial features is, both cross-sectionally and in terms of progression.[35] Because the aim of this study was to identify protein biomarkers that predicted incident disease, we defined new disease based on a relatively stringent threshold of moving from the lowest tertile of interstitial features to the highest tertile of interstitial features. Even with this definition,

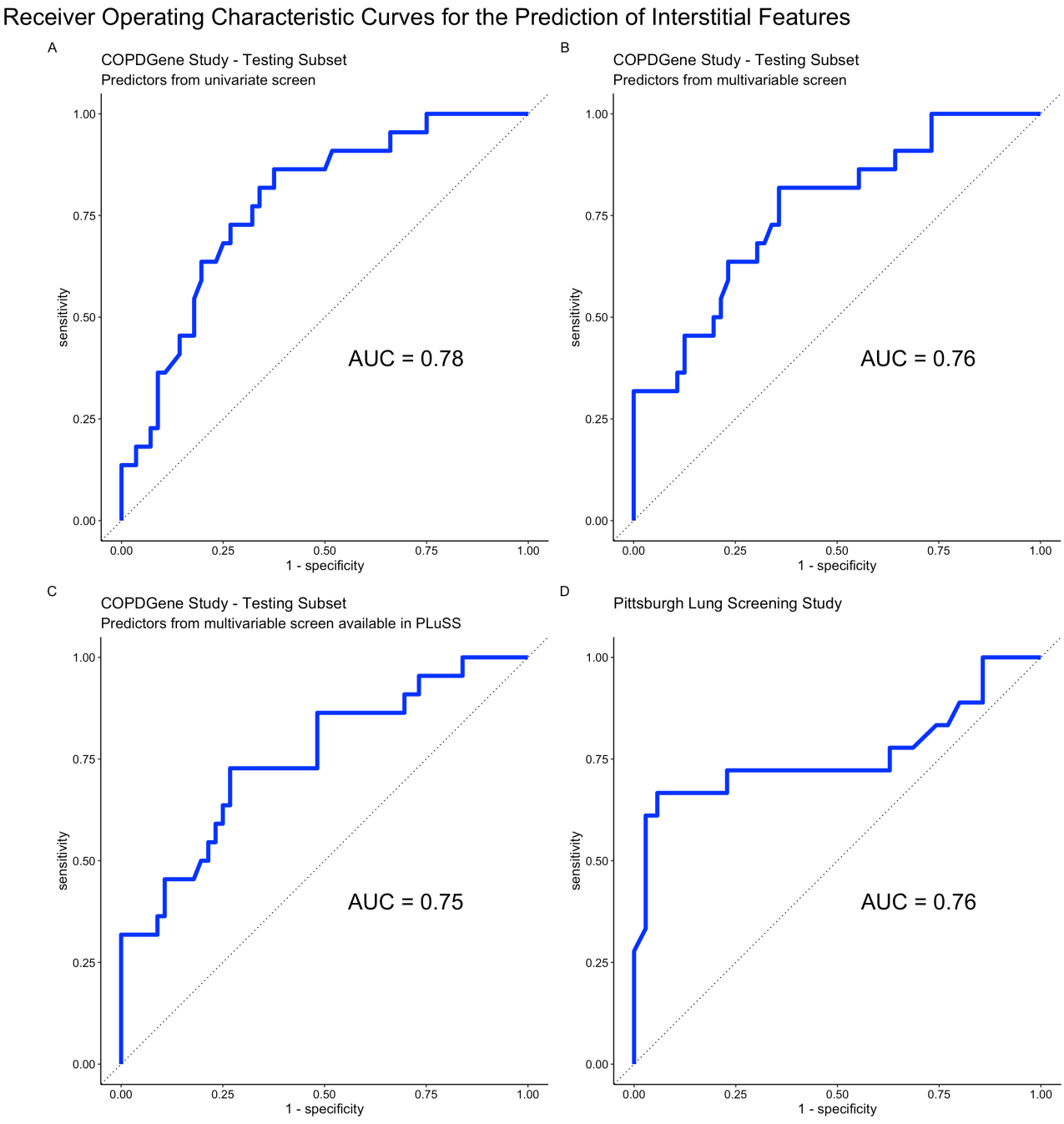

Figure 3 Receiver operating characteristic curves for random forest classifier trained using clinical and imaging features plus protein biomarkers. Receiver operating characteristic curves for (A) performance of the random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) plus the 20 protein biomarkers identified in the univariate screen, applied to the testing subset of COPDGene. (B) Performance of the random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) plus the 11 protein biomarkers found in the multivariable confirmation step, applied to the testing subset of COPDGene. (C) Performance of the random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) plus the four protein biomarkers found in the multivariable confirmation step in COPDGene that were available in PLuSS, applied to the testing subset of COPDGene. (D) Performance of the random forest classifier trained using clinical/imaging features (age, gender, smoking status, pack-years and baseline interstitial features) plus the four protein biomarkers found in the multivariable confirmation step in COPDGene that were available in PLuSS, applied to all available complete data in PLuSS. AUC, area under the curve; PLuSS, Pittsburgh Lung Screening Study.

the absolute change in interstitial features was relatively small. This, combined with our recent work on quantitative emphysema and interstitial progression in which a very small increase in fibrotic appearing parenchyma was associated with a significant increase in mortality, suggests that even small amounts of parenchymal change may be

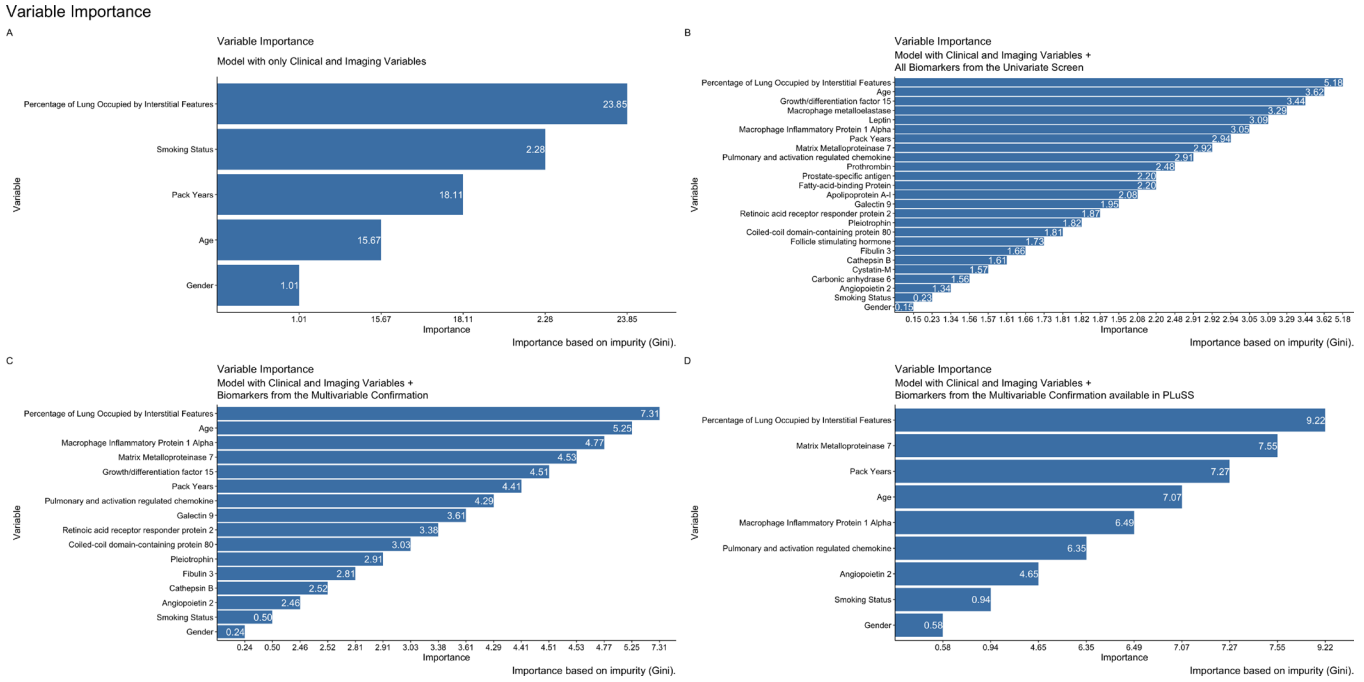

Figure 4 Variable importance for incident interstitial features prediction models. (A) Variable importance for the random forest classifier trained using only clinical/imaging features. (B) Variable importance for the random forest classifier trained using clinical/imaging features plus the 20 protein biomarkers identified in the univariate screen. (C) Variable importance for the random forest classifier trained using clinical/imaging features plus the 11 protein biomarkers found in the multivariable confirmation step. (D) Variable importance for the random forest classifier trained using clinical/imaging features plus the four protein biomarkers found in the multivariable confirmation step in COPDGene that were available in PLuSS. PLuSS, Pittsburgh Lung Screening Study.

clinically important.[36] However, it also raises the possibility of over diagnosing disease progression. Also, the overall decrease in interstitial features between visits in COPDGene suggests that other processes such as survival bias and changes in image acquisition over time are also important to investigate. Additional work is needed to better define minimum clinically important differences for these measurements and to determine other clinical and image acquisition-related factors that affect their measurement over time.[17 35]

With regard to limitations of the biomarker analyses in particular, it would be difficult to generate these data clinically, potentially limiting the clinical utility of these findings. However, while it is clearly impractical to do complete SOMALogic-type analyses for all potential patients at this current time, it may be possible to measure just a few biomarkers as shown in the final classifier. This goal, limiting the number of biomarkers selected, was the rationale for using a more stringent FDR threshold for the univariate selection step. Although the use of varying FDR thresholds can be considered depending on the overall goal of the study and the test application, this may have resulted in identifying fewer relevant biomarkers than the optimal approach.[21 22] Similarly, in order to limit the complexity of the final model created and potentially make these results more readily implemented clinically, genotype information was not included in these analyses and its integration may help further refine machine learning-based prediction models. Also, while a model that only included clinical and biomarker predictors was considered, because the diagnosis of interstitial features in this study required CT imaging and the definition of ILD clinically does as well, a clinical and protein biomarker-based model that does not include imaging would be unlikely to be of utility in either the research or the clinical setting.[37] Finally, there was a difference in the generation of data between the cohorts, especially with regard to CT scan protocol and cohort design, as well as how the biomarkers were measured, though the robustness of the findings in spite of these differences could also be viewed as a potential strength. For example, the raw values of interstitial features varied widely between the two cohort. This was primarily due to differences in CT protocol and radiation dose, and work is ongoing to overcome this issue.[38] Also, and as noted above, the interval between CT scans varied more for the PLuSS participants than the COPDGene participants. Future work will be needed to address these and the other aforementioned issues as well as investigating if other imaging measures such as lung volume, densitometry and airway measures improve the performance of imaging-based prediction models.

In summary, we identified a number of peripheral protein biomarkers associated with the presence and progression of interstitial features, which in some people may represent early ILDs such as pulmonary fibrosis.[5] In

addition, we demonstrated that these biomarkers can be used in conjunction with clinical and imaging features, such as the percentage of lung occupied by interstitial features on chest CT, to predict the incident development of new interstitial features over 5 years of follow-up. Although additional work is needed in clinical cohorts to replicate these findings, they may ultimately prove useful for identifying potential therapeutic targets to intervene specifically on early-stage disease, as well as for identifying those patients at the highest risk for pulmonary fibrosis before it becomes symptomatic and severe.

Author affiliations
[1]Department of Critical Care Medicine, South Shore Hospital, South Weymouth, Massachusetts, USA
[2]Tufts University School of Medicine, Boston, Massachusetts, USA
[3]Pulmonary and Critical Care Division, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[4]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[5]Department of Radiology, Brigham and Women's Hospital, Boston, Massachusetts, USA
[6]Boston University School of Medicine, Boston, Massachusetts, USA
[7]Division of Pulmonary/Critical Care, Northwestern University, Chicago, Illinois, USA
[8]Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[9]Medicine, National Jewish Health, Denver, Colorado, USA
[10]Medicine, Pulmonary Division, University of Pittsburgh, pittsburgh, Pennsylvania, USA
[11]Applied Chest Imaging Laboratory, Brigham and Women's Hospital, Boston, Massachusetts, USA
[12]Department of Medicine: Pulmonary, Critical Care and Sleep Medicine, Baylor College of Medicine, Houston, Texas, USA
[13]Pulmonary and Critical Care Medicine, Brigham and Women's Hospital/ Harvard Medical School, Boston, Massachusetts, USA

Acknowledgements In addition to the research support indicated on the title slide, the authors also greatly appreciate all of the research participants who participated in this study.

Contributors Study conception and design: SA, TJD, RuSJE, IOR and GRW. Data acquisition and analysis: SA, TJD, RaSJE, RuSJE, IOR, GRW, RB and DOW. Data interpretation: all authors. Initial manuscript draft: SA. Manuscript revision for critically important intellectual content: all authors. Final approval of the manuscript: all authors. Accountable for work: SA, TJD, IOR and GRW. Guarantor: SA.

Funding The COPDGene study (NCT00608764) is supported by NHLBI R01 HL089897 and R01 HL089856, as well as by the COPD Foundation through contributions made to an Industry Advisory Board composed of AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. The PLuSS study is supported by the University of Pittsburgh Lung Cancer SPORE: NCI P50-CA90440, University of Pittsburgh Cancer Institute and University of Pittsburgh Medical Center. Additional funding for this work includes National Institutes of Health grants: K08-HL145118 (SA), K23. HL119558/R03HL148484/ R01HL155522 (TJD), R01-HL116931 (RuSJE, GRW), R21-HL140422 (RaSJE, GRW), P01-HL114501 (GRW), and P30-CA047904 (DOW). As well as from the Department of Defense (DOD W81XWH1810772 (TJD, IOR, GRW, DOW)), Boehringer-Ingelheim Pharmaceuticals (GRW) and the Pulmonary Fibrosis Foundation (SA).

Competing interests SA reports equity/dividends from Quantitative Imaging Solutions and consulting for Vertex Pharmaceuticals, Verona Pharmaceuticals and Triangulate Knowledge, all unrelated to the current work. TJD has received grant support from Bristol Myers Squibb, consulting fees from Boehringer Ingelheim and L.E.K. consulting, and has been part of a clinical trial funded by Genentech, unrelated to the current work. BC reports consulting fees from Quantitative Imaging Solutions, unrelated to the current work. RuSJE reports consulting fees from Quantitative Imaging Solutions, unrelated to the current work. VC reports no competing interests. NE reports no competing interests. RK reports grants and personal fees from AstraZeneca, personal fees from CVS Caremark, personal fees from Aptus Health, grants and personal fees from GlaxoSmithKline, personal fees from Boston Scientific, personal fees from Boston Consulting Group, all outside the submitted work. GL reports no competing interests. RB reports no competing interests. DOW reports advisory board membership and shareholder of Online Disruptive Technologies, unrelated to the current work. RaSJE reports equity/ dividends from Quantitative Imaging Solutions, unrelated to the current work. IOR reports no competing interests. GRW reports grants from Boehringer Ingelheim, BTG Interventional Medicine and Janssen Pharmaceuticals; consultancies/advisory board participation for Boehringer Ingelheim, Janssen Pharmaceuticals, Pulmonx, Novartis, Philips, CSL Behring and Vertex; and equity/dividends from Quantitative Imaging Solutions, unrelated to the current work, all outside the submitted work. GRW's wife works for Biogen.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Brigham and Women's Hospital (IRB 2007P000544). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data resulting from this work will be made publicly available via dbGaP (https://www.ncbi.nlm.nih.gov/gap/).

ORCID iDs
Samuel Ash http://orcid.org/0000-0003-0224-6939
Tracy J Doyle http://orcid.org/0000-0002-0770-1059

REFERENCES

1 Ash SY, Washko GR. Interstitial lung abnormalities: risk and opportunity. *Lancet Respir Med* 2017;5:95–6.
2 Washko GR, Lynch DA, Matsuoka S, *et al*. Identification of early interstitial lung disease in Smokers from the Copdgene study. *Acad Radiol* 2010;17:48–53.
3 Hunninghake GM, Hatabu H, Okajima Y, *et al*. Muc5B promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* 2013;368:2192–200.
4 Putman RK, Hatabu H, Araki T, *et al*. Association between interstitial lung abnormalities and all-cause mortality. *JAMA* 2016;315:672–81.
5 Ash SY, Harmouche R, Putman RK, *et al*. Clinical and genetic associations of objectively identified interstitial changes in Smokers. *Chest* 2017;152:780–91.
6 Bermejo-Peláez D, Ash SY, Washko GR, *et al*. Classification of interstitial lung abnormality patterns with an ensemble of deep Convolutional neural networks. *Sci Rep* 2020;10:338.
7 Richeldi L, du Bois RM, Raghu G, *et al*. Efficacy and safety of Nintedanib in idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2071–82.
8 King TE Jr, Bradford WZ, Castro-Bernardini S, *et al*. A phase 3 trial of Pirfenidone in patients with idiopathic pulmonary fibrosis. *N Engl J Med* 2014;370:2083–92.
9 Podolanczuk AJ, Oelsner EC, Barr RG, *et al*. High Attenuation areas on chest computed tomography in community-dwelling adults: the MESA study. *Eur Respir J* 2016;48:1442–52.
10 Karampitsakos T, Juan-Guardela BM, Tzouvelekis A, *et al*. Precision medicine advances in idiopathic pulmonary fibrosis. *EBioMedicine* 2023;95:104766.
11 Regan EA, Hokanson JE, Murphy JR, *et al*. Genetic epidemiology of COPD (Copdgene) study design. *COPD* 2010;7:32–43.
12 Wilson DO, Weissfeld JL, Balkan A, *et al*. Association of radiographic emphysema and airflow obstruction with lung cancer. *Am J Respir Crit Care Med* 2008;178:738–44.
13 Wilson DO, Weissfeld JL, Fuhrman CR, *et al*. The Pittsburgh lung screening study (Pluss): outcomes within 3 years of a first computed tomography scan. *Am J Respir Crit Care Med* 2008;178:956–61.
14 Ash SY, Harmouche R, Ross JC, *et al*. The objective identification and Quantification of interstitial lung abnormalities in Smokers. *Acad Radiol* 2017;24:941–6.
15 Diaz AA, Strand M, Coxson HO, *et al*. Disease severity dependence of the longitudinal association between CT lung density and lung function in Smokers. *Chest* 2018;153:638–45.

16 Ross JC, San José Estépar R, Kindlmann G, *et al*. Automatic lung lobe Segmentation using particles, thin plate Splines, and maximum a Posteriori estimation. *Med Image Comput Comput Assist Interv* 2010;13(Pt 3):163–71.
17 Choi B, Adan N, Doyle TJ, *et al*. Quantitative interstitial abnormality progression and outcomes in the genetic epidemiology of COPD and Pittsburgh lung screening study cohorts. *Chest* 2023;163:164–75.
18 Gold L, Ayers D, Bertino J, *et al*. Aptamer-based Multiplexed Proteomic technology for biomarker discovery. *PLoS ONE* 2010;5:e15004.
19 Candia J, Cheung F, Kotliarov Y, *et al*. Assessment of variability in the Somascan assay. *Sci Rep* 2017;7:14248.
20 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995;57:289–300. 10.1111/j.2517-6161.1995.tb02031.x Available: https://rss.onlinelibrary.wiley.com/toc/25176161/57/1
21 Chen Z, Boehnke M, Wen X. Revisiting the genome-wide significance threshold for common variant GWAS. G3 2021;11:jkaa056.
22 Rogers AJ, Weiss ST. *Clinical and Translational Science*Second Edition. Sect III: Hum Genet, 2017.
23 Ley B, Collard HR. Epidemiology of idiopathic pulmonary fibrosis. *CLEP* 2013;5:483.
24 Nembrini S, König IR, Wright MN. The revival of the Gini importance *Bioinformatics* 2018;34:3711–8.
25 Team rs. Rstudio: integrated development for R. *Published Online First* 2015.
26 Team RC. R: A language and environment for statistical computing. 2020.
27 Armstrong HF, Podolanczuk AJ, Barr RG, *et al*. Serum matrix Metalloproteinase-7, respiratory symptoms, and mortality in community-dwelling adults: the multi-ethnic study of Atherosclerosis. *Am J Respir Crit Care Med* 2017;196:1311–7.
28 Bauer Y, White ES, de Bernard S, *et al*. MMP-7 is a predictive biomarker of disease progression in patients with idiopathic pulmonary fibrosis. *ERJ Open Res* 2017;3:00074-2016.
29 Atamas SP, Luzina IG, Choi J, *et al*. Pulmonary and activation-regulated Chemokine stimulates collagen production in lung fibroblasts. *Am J Respir Cell Mol Biol* 2003;29:743–9.
30 Kodera M, Hasegawa M, Komura K, *et al*. Serum pulmonary and activation-regulated Chemokine/Ccl18 levels in patients with systemic sclerosis: a sensitive indicator of active pulmonary fibrosis. *Arthritis Rheum* 2005;52:2889–96.
31 Russo RC, Alessandri AL, Garcia CC, *et al*. Therapeutic effects of Evasin-1, a Chemokine binding protein, in Bleomycin-induced pulmonary fibrosis. *Am J Respir Cell Mol Biol* 2011;45:72–80.
32 Ntatsoulis K, Karampitsakos T, Tsitoura E, *et al*. Commonalities between ARDS, pulmonary fibrosis and COVID-19: the potential of Autotaxin as a therapeutic target. *Front Immunol* 2021;12:687397.
33 Survival analysis for epidemiologic and medical research. 2008:1–26.
34 Volkmann ER, Tashkin DP, Silver R, *et al*. Sex differences in clinical outcomes and biological profiles in systemic sclerosis-associated interstitial lung disease: a post-hoc analysis of two randomised controlled trials. *Lancet Rheumatol* 2022;4:e668–78.
35 Jones PW, Beeh KM, Chapman KR, *et al*. Minimal clinically important differences in pharmacological trials. *Am J Respir Crit Care Med* 2014;189:250–5.
36 Ash SY, Choi B, Oh A, *et al*. Deep learning assessment of progression of emphysema and Fibrotic interstitial lung abnormality. *Am J Respir Crit Care Med* 2023;208:666–75.
37 Society ER. American Thoracic society/European respiratory society International Multidisciplinary consensus classification of the idiopathic interstitial Pneumonias. *Am J Respir Crit Care Med* 2002;165:277–304.
38 Vegas Sánchez-Ferrero G, Díaz AA, Ash SY, *et al*. Quantification of emphysema progression at CT using simultaneous volume, noise, and bias lung density correction. *Radiology* 2024;310.