



Published in final edited form as:

*Thorax*. 2019 December ; 74(12): 1131–1139. doi:10.1136/thoraxjnl-2018-212430.

## ***MUC5B* variant is associated with visually and quantitatively detected preclinical pulmonary fibrosis**

**Susan K. Mathai, MD<sup>1,2,\*</sup>, Stephen M. Humphries, PhD<sup>3</sup>, Jonathan A. Kropski, MD<sup>4</sup>, Timothy S. Blackwell, MD<sup>4</sup>, Julia Powers, MPH<sup>1</sup>, Avram D. Walts<sup>1</sup>, Cheryl Markin<sup>4</sup>, Julia Woodward<sup>1</sup>, Jonathan H. Chung, MD<sup>5</sup>, Kevin K. Brown, MD<sup>7</sup>, Mark P. Steele, MD<sup>1</sup>, James E. Loyd, MD<sup>4</sup>, Marvin I. Schwarz, MD<sup>1</sup>, Tasha E. Fingerlin, PhD<sup>6</sup>, Ivana V. Yang, MD<sup>1</sup>, David A. Lynch, MD<sup>3</sup>, David A. Schwartz, MD<sup>1,\*</sup>**

<sup>1</sup>University of Colorado School of Medicine, Department of Medicine 12631 E. 17<sup>th</sup> Ave., Mail Stop B178, Aurora, CO 80045, USA

<sup>2</sup>Baylor University Medical Center at Dallas, Center for Advanced Heart & Lung Disease 3410 Worth Street, Suite #250, Dallas, Texas 75246, USA

<sup>3</sup>National Jewish Health, Department of Radiology 1400 Jackson St, Denver, CO 80206, USA

<sup>4</sup>Vanderbilt University, Department of Medicine 1301 Medical Center Dr b817, Nashville, TN 37212, USA

<sup>5</sup>University of Chicago, Department of Radiology 5841 South Maryland Avenue, Chicago, IL 60637, USA

<sup>6</sup>National Jewish Health, Center for Genes, Environment, and Health Smith Building; A647, 1400 Jackson Street, Denver, CO 80206, USA

<sup>7</sup>National Jewish Health, Department of Medicine 1400 Jackson St, Denver, CO 80206, USA

### **Abstract**

\*Corresponding Authors: Susan K. Mathai, MD, Center for Advanced Heart & Lung Disease, Baylor University Medical Center at Dallas, 3410 Worth Street, Suite #250, Dallas, Texas 75246, USA, 214-820-7856 | susan.mathai@BSWHealth.org, David A. Schwartz, MD, Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, 12631 E. 17th Avenue, Mail Stop B178, Aurora, Colorado 80045, USA, 303-724-1783 | david.schwartz@ucdenver.edu.

Author contributions: S.K.M. compiled and analyzed anonymized visual CT reads, performed genotyping, statistical analyses, wrote first draft of manuscript, and revised the manuscript in collaboration with the other authors. S.M.H. created and implemented the quantitative HRCT algorithm described in this manuscript; he also performed %HAA quantification on CT scans in this study and contributed to first draft of the manuscript. J.P. led research coordination, data management, and study subject recruitment at the University of Colorado. A.W. performed nucleic acid extractions and sample management for this project. J.W. contributed to data management and study subject recruitment. I.V.Y. contributed to and advised overall study design, genetic analyses, and gene expression analyses. T.E.F. oversaw statistical analyses presented in this manuscript. K.K.B. and M.P.S. were integral to study subject recruitment. M.I.S. contributed to overall study design. D.A.L. and J.H.C. performed radiologic reviews for this study. C.M., J.A.K., T.S.B., and J.E.L. led patient recruitment, study design, and data and sample management at Vanderbilt University. D.A.S. led overall study design, subject recruitment, and contributed to each stage of data analyses and manuscript drafting. All authors contributed to manuscript revisions.

Competing interests: D.A.S. is the founder and chief scientific officer of Eleven P15, a company focused on the early diagnosis and treatment of pulmonary fibrosis. D.A.S. has an awarded patent (US Patent No: 8,673,565) for the treatment and diagnosis of fibrotic lung disease. D.A.L. and S.M.H. have a pending patent (Application US20170330320A1) for image analysis; S.M.H. reports consulting agreement with Boehringer Ingelheim.

**Background:** Relatives of Familial Interstitial Pneumonia (FIP) patients are at increased risk for pulmonary fibrosis. We assessed the prevalence and risk factors for preclinical fibrosis (PrePF) in first-degree relatives of FIP patients and determined the utility of deep learning in detecting PrePF on CT.

**Methods:** First-degree relatives of FIP patients over 40 years of age who believed themselves to be unaffected by pulmonary fibrosis underwent CT scans of the chest. Images were visually reviewed, and a deep learning algorithm was used to quantify lung fibrosis. Genotyping for common IPF risk variants in *MUC5B* and *TERT* was performed.

**Findings:** In 494 FIP relatives from 263 FIP families, the prevalence of PrePF on visual CT evaluation was 15.6% (95% CI [12.6,19.0]). Compared to visual CT evaluation, deep-learning quantitative CT analysis had 84% sensitivity (95% CI [0.72, 0.89]) and 86% sensitivity (95% CI [0.83, 0.89]) for discriminating subjects with visual PrePF diagnosis. PrePF subjects were older (65.9, SD: 10.1 years) than subjects without fibrosis (55.8± 8.7), more likely to be male (49% versus 37%), more likely to have smoked (44% versus 27%), and to have the *MUC5B* promoter variant rs35705950 (minor allele frequency 0.29 versus 0.21). *MUC5B* variant carriers had higher quantitative CT fibrosis scores (mean difference 0.36%), a difference that remains significant when controlling for age and sex.

**Interpretation:** PrePF is common in FIP relatives. Its prevalence increases with age and the presence of a common *MUC5B* promoter variant. Quantitative CT can detect these imaging abnormalities.

**Funding:** NIH-NHLBI (UH2/3-HL123442, R01-HL097163, R21/R33-HL120770, P01-HL092870, K23-HL136785, K08-HL130595, F32HL123240), U.S. DOD (W81XWH-17-1-0597).

## INTRODUCTION

Idiopathic pulmonary fibrosis (IPF), the most common idiopathic interstitial pneumonia, is a poorly understood disease characterized by progressive lung parenchymal scarring, impaired gas exchange, loss of lung function, physical debilitation, and shortened life-span. Median survival is approximately 3 years from the time of diagnosis [1], and the clinical course is unpredictable [1]. There are no curative therapies [2,3] other than lung transplantation.

Recent studies have identified genetic variants, both common and rare, associated with both familial and sporadic forms of pulmonary fibrosis [4–7]. A *MUC5B* promoter variant has been shown to be the most important variant associated with familial and sporadic disease and with interstitial lung abnormalities (ILAs) [4,5,8]. Numerous rare variants in *TERT* (and other telomerase-pathway genes) have been thought to be critical in familial diseases [6,9–11]; more recently a common variant in *TERT* [5] has been associated with sporadic and familial disease.

Better understanding and recognition of early pulmonary fibrosis is critical because medical therapies have been shown to slow progression, not to reverse existing fibrosis; intervention before irreversible fibrosis has become extensive has the potential to improve quality of life and decrease morbidity. While IPF affects approximately 5 million people worldwide [1], between 1.8 of the general population and 14% of the familial at-risk population 50 years

of age have radiologic findings of undiagnosed pulmonary fibrosis [8,12,13]. Large cohort studies indicate that ILAs, postulated to represent early pulmonary fibrosis, are associated with increased mortality and generally progress over time [12,13]. Members of families with 2 or more cases of pulmonary fibrosis (FIP, Familial Interstitial Pneumonia) have been identified as an “at-risk” population. In a previous study of FIP relatives, 14% had ILAs on high resolution computed tomography (HRCT), and 35% had an abnormal transbronchial biopsy indicating interstitial lung disease [14].

HRCT plays a key role in the diagnosis of the Idiopathic Interstitial Pneumonias (IIPs), including IPF. Currently, visual pattern diagnosis by thoracic radiologists, in conjunction with multidisciplinary clinical conference, is the gold standard for diagnosing IIPs [15]. However, visual assessment is imprecise and hampered by inter-observer variation [16]. Quantitative HRCT (qHRCT) evaluation provides measures of fibrosis extent that, in subjects diagnosed with IPF, correlate with degree of physiologic impairment at baseline, and may be more sensitive to subtle changes in disease status than routinely used physiological metrics [17,18]. The design and utility of qHRCT methods in the context of early forms of fibrotic ILD requires further study [19]. Deep learning methods have been increasingly used in imaging to identify and classify CT patterns [20], and may be valuable in detection of early lung fibrosis.

A key strength of deep learning algorithms, such as convolutional neural networks (CNNs), is that they simultaneously optimize feature extraction and calculation of classification rules. During training, CNNs “learn” to extract the most effective image features, including textural features at multiple scales, for the given classification task. This is in contrast to more traditional methods that rely on separate processes to engineer and select features, then develop classification rules. Engineered features are designed manually, often by using combinations of standard statistical or image processing calculations, and may not be the most effective features for a given classification task like the discrimination of pulmonary fibrosis.

This study aims to: (1) examine risk factors, including two common fibrosis-associated genetic variants in *MUC5B* and *TERT*, for undiagnosed pulmonary fibrosis (PrePF) in FIP first-degree relatives; and (2) determine the utility of a deep learning, texture-based qHRCT method in the detection of PrePF in this cohort.

## MATERIALS AND METHODS

### FIP Relatives Screening:

At the University of Colorado, National Jewish Health, and Vanderbilt University (COMIRB #15–1147; NJH IRB 1441a; Vanderbilt IRB #020343), non-Hispanic white (NHW) first-degree relatives of FIP patients, defined as those in families with two or more cases of pulmonary fibrosis (Figure S1), were contacted. After informed consent, first-degree relatives without a known prior diagnosis of pulmonary fibrosis and greater than 40 years of age were offered HRCT scans of the chest and peripheral blood draw. Those younger than 40 years of age or who reported on pre-scan questionnaires to be personally affected by pulmonary fibrosis were excluded (Figure 1).

**Visual CT Review:**

See Supplement for details. HRCT scans were interpreted by study radiologists using a standardized method [21]. “PrePF” was defined as the presence of “probable” or “definite” fibrotic ILD on HRCT in FIP relatives who had no known diagnosis of pulmonary fibrosis at the time of study enrollment (Figures 1,2).

**Quantitative CT:**

Inspiratory HRCT series with slice thickness 1.25mm and spacing 20.0mm were selected for quantitative analysis. This included 212 volumetric series with thin, contiguous sections (slice thickness and spacing both  $\leq 1.25$ mm) and 191 non-volumetric scans (56 with slice spacing  $>1.25$ mm and  $<10$ mm, 65 with slice spacing of 10mm and 70 with slice spacing = 20mm). Technically inadequate scans were omitted (Figure S2). In addition, 100 inspiratory volumetric HRCTs of normal, never-smoker control subjects from the COPDGene cohort were analyzed (Table S1) [22,23]. In an initial process, the lungs were segmented using a deep learning model that had been trained using CT of subjects with and without fibrosis. Details are available in the Supplement. Trained analysts verified lung segmentation visually and made edits, if necessary. Examples of the categorization of different parts of CT scans are shown in Figure 3. Some studies were acquired with contiguous thin axial sections while others used 1 or 2 cm intervals. Reconstruction kernel, a parameter that affects image sharpness and noise, was not standardized.

Fibrosis quantification on CT scans was performed using a second deep learning technique, called deepDTA, consisting of a convolutional neural network (CNN) algorithm trained with image regions of normal and abnormal lung identified by expert radiologists. Training data and an earlier algorithm version, called Data-driven Textural Analysis (DTA), were described previously [17]. Here, a more complex CNN architecture was employed that classifies image regions using pixel and texture features extracted by multiple convolutional layers at different scales. The CNN classifies image regions as either normal or fibrotic, with the fibrotic category trained using image regions labeled by a radiologist as reticular abnormality, honeycombing or traction bronchiectasis. Subject level HRCT fibrosis scores were computed as the percentage of total lung volume classified as fibrotic (Figure 3; Figure S3). A simpler previously described densitometric analysis of HRCTs, percent high attenuation area (%HAA), was also performed for comparison [24] (see supplement).

**Blood Processing, Genotyping, and Autoantibody Testing:** See supplement.

**Statistical Analysis:**

Analysis of the effect of specific alleles on PrePF risk was performed using minor allele frequency (MAF) for comparison of variant prevalence in the study groups; statistical significance was determined with a z-score test for proportions or a mixed effects logistic regression model when controlling for other variables (age, sex, smoking history, and family [random effect]) in dominant and log-additive models.

Distribution of qHRCT fibrosis scores was left skewed, as was %HAA, so these values were log transformed prior to analyses (Figures S4–7). Log of qHRCT fibrosis score (hereafter,

“fibrosis score”) and log (%HAA) were compared with visual scores using ANOVA and Tukey’s honest significant difference test. To determine the ability of qHRCT scores to predict visual diagnosis of PrePF, receiver-operating characteristic (ROC) analysis was performed. Optimal threshold for discriminating visual diagnosis of fibrotic ILD was determined with Youden’s method. Five-fold cross-validation was performed to test detection accuracy, sensitivity and specificity, and consistency of optimal threshold. Linear regression was performed to test association between the *MUC5B* genotype and qHRCT fibrosis score and log (%HAA).

A p-value of <0.05 was considered statistically significant for differences between groups as well as for associations between individual variables and outcomes in linear and logistic regression modeling. Statistical analyses were performed using RStudio (v.0.99.473).

## RESULTS

### Study cohort characteristics

1,090 familial interstitial pneumonia (FIP) first-degree relatives were contacted; 523 eligible subjects underwent HRCT screening (Figure 1). Of the 521 subjects, 26 were excluded due to technical inadequacy of images and one for an equivocal consensus read by study radiologists. The remaining 494 subjects from 263 families were included in the analyses. Subjects’ mean age was 57 years (SD: 9.6), 189 (38%) were male, and 148 (30%) were either current or former smokers. The minor allele (T) frequency of the *MUC5B* promoter variant rs35705950 was 0.22 in this cohort; 42% of the subjects in this cohort had one or two copies of the minor allele (Table 1). The minor allele (C) frequency of the *TERT* variant rs2736100 was 0.47 in the entire cohort; 69% of the subjects in the cohort having one or two copies of the minor allele (Table 1).

### Prevalence of preclinical pulmonary fibrosis (PrePF) in FIP relatives

Of the 494 HRCT scans, 399 showed no CT evidence of interstitial lung disease (ILD), and 93 showed evidence of ILD, either fibrotic (27 probable and 50 definite) or non-fibrotic (n=16). Therefore, among these 494 subjects who reported being personally unaffected by pulmonary fibrosis, the PrePF prevalence was 15.6% (n=77) (Figure 1).

The CT patterns noted in visually identified PrePF subjects (Table 2) show that possible, probable, or definite UIP pattern was the most commonly considered (n=59, 77% of all PrePF cases). NSIP was considered in 45 subjects (58% of all PrePF cases). The fibrotic changes were most commonly lower-lobe predominant and subpleural in nature, consistent with a UIP pattern (Table 2). Non-fibrotic ILD scans, on the other hand, generally had more diffuse, upper-lobe predominant abnormalities (Table S2–3).

There were 402 study subjects with HRCT scans that were technically adequate for quantitative assessment (Figure S2). 212 of the scans had both slice thickness and spacing  $\leq 1.25$ mm (thin, contiguous); of the remaining 191 scans, 56 had slice spacing  $> 1.25$ mm and  $< 10$ mm, 65 had slice spacing = 10mm, and 70 had slice spacing = 20mm. Volumetric HRCT scans on an additional 100 COPDGene subjects were included as normal controls (Table S1; Figure S3). HRCT CNN fibrosis score means were significantly different

( $p < 0.0001$ ) across groups defined by visual diagnosis (Figure 4). Comparison of means showed fibrosis score were significantly different comparing each group (all between-group comparisons  $p < 0.01$ ). Means of log (%HAA) scores were also significantly different across visual scoring groups ( $p < 0.0001$ ), and individual between-group comparisons showed log (%HAA) was significantly different in most comparisons ( $p < 0.0001$ ), except between the “probable” and “definite” visual scores ( $p = 0.35$ , Figure S7).

ROC analyses showed that fibrosis score discriminates subjects with visual diagnosis of PrePF (Figure 5B). Average area under the curve (AUC) in five-fold cross validation was 0.92 (range 0.91–0.93) and average accuracy, sensitivity, and specificity in the test partitions were 0.85 (range 0.81–0.88), 0.81 (range 0.71–0.92), and 0.86 (range 0.79–0.90), respectively. Optimal threshold for log fibrosis score was 0.60 (range 0.53 – 0.71), corresponding to 1.8% fibrotic area in examined lung. Utilizing a cutoff of 0.60 for log fibrosis score on the entire dataset, the sensitivity was 84% (95% CI [72, 92]), specificity was 86% (95% CI [83, 89]); and accuracy was 86%; while the positive predictive value of this test was only 46% (95% CI [36, 55]), the negative predictive value was 97% (95% CI [95, 99]) (Figure 5B–C).

Compared to the classification achieved with the CNN as described above, ROC analysis of log %HAA had lower mean AUC 0.80 (range 0.79–0.81) and average accuracy, sensitivity, and specificity of 0.67 (range 0.63–0.70), 0.82 (range 0.75–0.91), and 0.64 (range 0.62–0.70), respectively (Figure 5A). Mean optimal threshold for log %HAA ranged from 1.49–1.57.

Utilizing a cutoff of 1.49 for log %HAA, the sensitivity was 89% (95% CI [78,95]), specificity was 62% (95% CI [57,66]), and accuracy was 60%; while the positive predictive value of this test was only 24% (95% CI [19,30]), the negative predictive value of this test was 96% (95% CI [95,99]).

### Risk Factors for PrePF

Subjects with PrePF were older (mean age 65.9 years, SD 10.1) than those without fibrosis (mean age 55.8, SD 8.7;  $p = 6.36 \times 10^{-13}$ ) (Table 1, Figure S8); they were also more likely to have ever smoked (44% versus 27%,  $p = 0.004$ ), and to be male (49% versus 37%,  $p = 0.05$ ). However, there was no difference in breathlessness between the PrePF and subjects without fibrosis (mean score 0.5 versus 0.6,  $p = 0.24$ , Table 3). Quantitative fibrosis score was positively associated with breathlessness score ( $p = 0.007$ ), even after controlling for age (0.65), male sex ( $p = 0.52$ ), and smoking history ( $p = 0.59$ ). When fibrosis was defined by quantitative fibrosis score cutoff (0.60), there was a trend towards higher breathlessness score in scans demonstrating lung fibrosis (0.44 versus 0.65,  $p = 0.08$ ).

Screening for autoantibodies in this cohort revealed that there were no differences between PrePF and unaffected subjects in terms of overall seropositivity or specific antibodies’ testing in this cohort (Table S4). For quantitatively defined lung fibrosis, there was also no significant difference between groups, with similar overall seropositivity rates (11% versus 16%,  $p = 0.30$ ).



The *MUC5B* promoter variant rs35705950 was associated with the visual diagnosis of PrePF (present in 40% of those without fibrosis versus 53% with PrePF; MAF 0.29 versus 0.21, respectively, OR=2.14 (95% CI [1.00, 4.63], p=0.02, Table 1). After age 60, there was a statistically significant difference in the proportion of subjects with visually diagnosed PrePF when the cohort was stratified by *MUC5B* genotype (23.8% versus 39.8% prevalence, p=0.02); prior to age 60, PrePF prevalence is not significantly different by genotype (Figure 6).

*MUC5B* variant carriers, regardless of their visual CT diagnosis, had significantly higher qHRCT fibrosis scores (mean difference 0.36, p=0.006). The association between *MUC5B* genotype and fibrosis score was significant even when controlling for age, sex and smoking history in a linear regression (p=0.017, Table 4). Age (p<2.0×10<sup>-16</sup>) was significantly associated with fibrosis score, but male sex (p=0.26) was not; the association of smoking and fibrosis score was borderline (p=0.05). The simpler quantitative scoring method, log %HAA, was not significantly different in *MUC5B* variant carriers (p=0.4).

When the 341 subjects with a visual inspection negative for fibrosis were separated further by whether or not qHRCT score indicated fibrosis, 59 were identified to have lung fibrosis by the deep learning method and 282 were found to be unaffected. In those that were classified negative by both visual and computational methods, mean age was 54.7 (95% CI [53.8, 55.6]), 101 (35.8%) were male, and 271 had genotyping available (MAF of 0.21 for the *MUC5B* promoter variant). Of those that were classified negative visually but fibrotic by deep learning (n=59), the mean age was 61.2 (95% CI [58.4, 65.0]), 22 were male (37.3%), and all had genotyping available which revealed a MAF of 0.26 for the *MUC5B* promoter variant. Those that were identified as having lung fibrosis by deep learning were older (61.2 vs. 54.7 years; p=4.2×10<sup>-5</sup>) and were more likely to have the *MUC5B* variant (MAF=0.26 vs. 0.21; p=0.18); however the *MUC5B* promoter variant association in this sub-analysis did not reach statistical significance.

In contrast to the *MUC5B* variant, the common IPF-associated *TERT* variant (rs2736100) was not significantly associated with PrePF assessed either qualitatively (MAF 0.47 in PrePF versus 0.46 in unaffected, p = 0.77) or quantitatively (MAF 0.50 fibrotic versus 0.47 not fibrotic, p=0.40).

When these factors were examined for their contributions to risk of PrePF in our study cohort, we used a mixed effects logistic regression model to test the independent effects of age sex, smoking, and *MUC5B* or *TERT* genotypes while controlling for family. Age remained significantly associated with PrePF (OR 1.15, 95% CI [1.09, 1.21], p=6.74×10<sup>-7</sup>) and the *MUC5B* variant was more common in PrePF (OR 2.14, 95% CI [1.00, 4.63], p=0.05) (Table 1). The common *TERT* variant (rs2736100) associated with fibrotic idiopathic interstitial pneumonia [5] was not associated with PrePF in simple comparison of allele frequency (MAF was 0.45 in PrePF versus 0.45 in unaffected, p = 0.92) or in a log-additive model controlling for age, sex, smoking history, and family relatedness (p=0.38) (Table 1).

## Secondary Subgroup Analyses

Given the presence of non-fibrotic ILD (n=16, Figure 1) in the “No Fibrosis” cohort, secondary analyses were performed that: 1) excluded non-fibrotic ILDs (Table S5); and 2) compared all ILD (inclusive of non-fibrotic ILD) to those without any ILD (Table S6). When non-fibrotic ILDs were excluded from analyses, PrePF subjects were older ( $p=1.7\times 10^{-12}$ ), more commonly male ( $p=0.05$ ), more often had a smoking history ( $p=0.003$ ) and had a higher prevalence of the *MUC5B* promoter variant (MAF 0.29 versus 0.20,  $p=0.02$ ). However, when controlling for family relatedness and the other risk factors in a mixed effects logistic regression, age was associated with PrePF (OR 1.15, 95% CI [1.09, 1.21],  $p=8.8\times 10^{-7}$ ), and the *MUC5B* promoter polymorphism had a borderline association with PreP (OR 2.15, 95% CI [0.99, 4.69],  $p=0.05$ ) (Table S5). Another secondary analysis of the data was performed in which all subjects with CT findings of ILD (fibrotic or non-fibrotic) were compared to those without any evidence of ILD (Table S6). Those with CT evidence of any ILD were older (mean age 64.5 years, SD: 10.2) compared to those without any evidence of ILD (mean age 55.7 years, SD: 8.7,  $p=7.2\times 10^{-12}$ ), more likely to be male ( $p=0.02$ ), more likely to have smoked ( $p=0.0003$ ), and more likely to carry the *MUC5B* promoter variant (MAF 0.29 versus 0.21,  $p=0.01$ ). When controlling for family relatedness in a mixed effects logistic regression model, age (OR 1.11, 95% CI [1.07, 1.15],  $p=5.58\times 10^{-9}$ ) and the *MUC5B* promoter variant (OR 1.87, 95% CI [1.04, 3.36],  $p=0.04$ ) were significantly associated with risk of ILD; smoking history had a borderline association (OR 1.81, 95% CI [1.01, 3.25],  $p=0.05$ ).

## DISCUSSION

Interstitial lung abnormalities have been studied in FIP relatives [14]; our present study builds on these initial findings by presenting data from a larger cohort, focusing specifically on evidence of fibrotic radiologic abnormalities, and utilizing qHRCT analysis. PrePF is common among FIP first-degree relatives, and a texture-based qHRCT analysis is useful in identifying these abnormalities in this population, and key risk factors predict those at risk of this disease. PrePF subjects are older, more likely to be male, and more likely to have smoked than the unaffected subjects [1]. Additionally, the gain-of-function *MUC5B* promoter variant rs35705950, which is associated with established pulmonary fibrosis [4,5,7,25–29], is more common in PrePF subjects when compared to their unaffected family members. Given the high prevalence of findings suggestive of the UIP pattern on HRCT scan among subjects with PrePF and the association of IPF risk factors (age, gender, cigarette smoking, and *MUC5B* promoter variant) with PrePF, our findings suggest that PrePF subjects are at risk of developing progressive fibrosis and that quantitative CT imaging represents a sensitive means of detecting these radiographic abnormalities.

Even in a population such as FIP first-degree relatives that is at baseline enriched for the *MUC5B* variant compared to the general NHW population [4,8], the *MUC5B* variant was more common in those with PrePF. Study of this variant in larger at-risk populations is necessary to determine if the genotype could be used to target prospective screening, especially in those over the age of 60 (Figure 6). It is important to note, however, that the prevalence of PrePF was relatively high even in those without the *MUC5B* variant,



suggesting that the absence of this variant alone may not indicate that a particular individual in this at-risk cohort would not warrant screening. Notably, we examined another IPF-associated common variant in *TERT* in this cohort and did not find that variant to be associated with PrePF; it is possible that due to the high MAF of this variant in the general population that this study was underpowered to detect its relationship to risk of PrePF.

Deep learning method is capable of detecting and quantifying fibrotic ILD patterns on CT in this cohort. Prior studies using a similar method [17] utilized established IPF cases and correlated quantitative scores with pulmonary function testing, suggesting that a quantitative HRCT score reflect physiologic change in addition to CT change. The current study supports the use of quantitative HRCT analysis to detect PrePF in a cohort of high-risk subjects without known disease since it is associated with breathlessness and the *MUC5B* promoter variant [4,5,30], a known IPF risk factor. However, the negative predictive value (97%) of using a quantitative fibrosis score cutoff was noted to be much higher than its positive predictive value (46%), suggesting that it may be particularly useful in terms of identifying higher risk scans that may require more careful visual inspection by radiologists. Recent studies illustrating that interstitial lung abnormalities are underreported in real-world settings suggest that technology-aided evaluation of routine chest imaging could improve timeliness of patient referral and evaluation [31].

This deep learning method based on textural analysis appears to be superior to %HAA, a simpler densitometry-based method that has been applied to quantitative CT assessment of ILDs [24]. Compared to the deep learning fibrosis score, the %HAA method was less accurate, had a lower positive predictive value, and was not associated with the *MUC5B* risk variant. While the %HAA method of HRCT analysis may capture some forms of PrePF, more advanced methods of quantifying subtle fibrosis are needed to quantify these features consistently. A limitation of deep learning is the need for a significant amount of labeled training data. The CNN used for the present study was trained using an independent dataset comprised of subjects enrolled in a clinical trial for IPF [17]. These subjects had more advanced lung fibrosis than those in this cohort, and their HRCT technical parameters were more consistent

Though this study was performed on a cohort of FIP first-degree relatives, we hypothesize that the findings could be relevant to first-degree relatives of sporadic IPF patients. Given that genome-wide studies have shown that FIP and sporadic IPF are indistinguishable in terms of common risk variants [5], a hypothesis that should be tested is that the genetic and genomic markers identified through the study of PrePF in FIP families could be applicable to first-degree relatives of sporadic IPF. Additional genetic variants, both common and rare, associated with fibrotic ILD could be examined in this cohort to determine how they contribute to risk of PrePF. Due to lack of power especially for common variants, larger cohorts would be required to determine additive effects of multiple genetic variants.

We hypothesize that PrePF could represent an early form of IPF. Prior studies suggest that interstitial lung abnormalities are associated with progressive loss of lung function and increased mortality [12,13], suggesting that the abnormalities we observe here, like ILAs studied in other cohorts [8,32,33], may have clinical consequences. However, longitudinal

observation of these subjects is required to determine at what rate and with what frequency PrePF progresses among FIP relatives in particular and whether it behaves like IPF. In addition, our ability to determine whether PrePF progresses to clinical IPF (versus other progressive fibrotic lung diseases) is limited by the fact that we do not have verified data regarding the subjects' previous environmental and occupational exposures—a clinical diagnosis of IPF would necessitate exclusion of environmental or occupational exposures associated with fibrosis with extensive interviewing. Given the differing age distributions of those with and without PrePF, it is likely that a substantial proportion of study subjects that had HRCTs without evidence of fibrosis at this one point in time may develop pulmonary fibrosis as they age. Further characterization of PrePF is necessary before we can determine how these findings should be applied to counseling and potential screening of FIP patient relatives.

Currently, the *MUC5B* promoter variant and quantitative methods of CT analysis are not utilized to assist in the clinical detection of PrePF. Future studies will further phenotype PrePF in this population. Other genetic variants (rare and common) associated with pulmonary fibrosis will be examined to determine the relative importance of different risk alleles in this population. Longitudinal study is required to determine the ability of the deep learning method of HRCT analysis to detect parenchymal changes that may precede fibrosis identified by standard visual examination.

In conclusion, PrePF is common in FIP relatives and associated with age as well as the *MUC5B* promoter variant. Quantitative HRCT scoring utilizing deep learning is capable of detecting PrePF and is associated with the *MUC5B* promoter variant, breathlessness symptoms, and visual diagnosis of fibrotic ILD.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## REFERENCES

1. Ley B, Collard HR. Epidemiology of Idiopathic Pulmonary Fibrosis. *Clin Epidemiol* 2013;5:483–92. doi:10.1055/s-2007-1006319 [PubMed: 24348069]
2. 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. doi:10.1056/NEJMoa1402584 [PubMed: 24836310]
3. King TE, 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. doi:10.1056/NEJMoa1402582 [PubMed: 24836312]
4. Seibold MA, Wise A, Speer M, et al. A common *MUC5B* promoter polymorphism and pulmonary fibrosis. *N Engl J Med* 2011;364:1503–12. [PubMed: 21506741]
5. Fingerlin TE, Murphy E, Zhang W, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* 2013;45:613–20. doi:10.1038/ng.2609 [PubMed: 23583980]
6. Borie R, Tabèze L, Thabut G, et al. Prevalence and characteristics of TERT and TERC mutations in suspected genetic pulmonary fibrosis. *Eur Respir J* 2016;48:1721–31. doi:10.1183/13993003.02115-2015 [PubMed: 27836952]

7. Wei R, Li C, Zhang M, et al. Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes. *Transl Res* 2014;163:494–502. doi:10.1016/j.trsl.2013.12.006 [PubMed: 24434656]
8. Hunninghake GM, Hatabu H, Okajima Y, et al. MUC5B promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* 2013;368:2192–200. doi:10.1056/NEJMoa1216076 [PubMed: 23692170]
9. Lee H-L, Ryu JH, Wittmer MH, et al. Familial idiopathic pulmonary fibrosis: clinical features and outcome. *Chest* 2005;127:2034–41. doi:10.1378/chest.127.6.2034 [PubMed: 15947317]
10. de Leon AD, Cronkhite JT, Katzenstein AL a, et al. Telomere lengths, pulmonary fibrosis and telomerase (TERT) Mutations. *PLoS One* 2010;5:e10680. doi:10.1371/journal.pone.0010680 [PubMed: 20502709]
11. Tsakiri KD, Cronkhite JT, Kuan PJ, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci U S A* 2007;104:7552–7. doi:10.1073/pnas.0701009104 [PubMed: 17460043]
12. Putman RK, Hatabu H, Araki T, et al. Association Between Interstitial Lung Abnormalities and All-Cause Mortality. *JAMA* 2016;315:672. doi:10.1001/jama.2016.0518 [PubMed: 26881370]
13. Araki T, Putman RK, Hatabu H, et al. Development and Progression of Interstitial Lung Abnormalities in the Framingham Heart Study. *Am J Respir Crit Care Med* 2016;:Epub ahead of print.
14. Kropski JA, Pritchett JM, Zoz DF, et al. Extensive phenotyping of individuals at risk for familial interstitial pneumonia reveals clues to the pathogenesis of interstitial lung disease. *Am J Respir Crit Care Med* 2015;191:417–26. doi:10.1164/rccm.201406-1162OC [PubMed: 25389906]
15. Castillo D, Walsh S, Hansell DM, et al. Validation of multidisciplinary diagnosis in IPF. *Lancet Respir Med* 2018;6:88–9. doi:10.1016/S2213-2600(18)30023-7 [PubMed: 29413085]
16. Watadani T, Sakai F, Johkoh T, et al. Interobserver Variability in the CT Assessment of Honeycombing in the Lungs. *Radiology* 2013;266:936–44. doi:10.1148/radiol.12112516 [PubMed: 23220902]
17. Humphries SM, Yagihashi K, Huckleberry J, et al. Idiopathic Pulmonary Fibrosis: Data-driven Textural Analysis of Extent of Fibrosis at Baseline and 15-Month Follow-up. *Radiology* 2017;285:270–8. doi:10.1148/radiol.2017161177 [PubMed: 28493789]
18. Kim HJ, Brown MS, Chong D, et al. Comparison of the Quantitative CT Imaging Biomarkers of Idiopathic Pulmonary Fibrosis at Baseline and Early Change with an Interval of 7Months. *Acad Radiol* 2015;22:70–80. doi:10.1016/j.acra.2014.08.004 [PubMed: 25262954]
19. Kliment CR, Araki T, Doyle TJ, et al. A comparison of visual and quantitative methods to identify interstitial lung abnormalities. *BMC Pulm Med* 2015;15:1–9. doi:10.1186/s12890-015-0124-x [PubMed: 25578181]
20. Kim GB, Jung K-H, Lee Y, et al. Comparison of Shallow and Deep Learning Methods on Classifying the Regional Pattern of Diffuse Lung Disease. *J Digit Imaging Published Online First: 17 10 2017*. doi:10.1007/s10278-017-0028-9
21. 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. doi:10.1016/j.acra.2009.07.016 [PubMed: 19781963]
22. Regan E a, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010;7:32–43. doi:10.3109/15412550903499522 [PubMed: 20214461]
23. Zach JA, Newell JD, Schroeder J, et al. Quantitative computed tomography of the lungs and airways in healthy nonsmoking adults. *Invest Radiol* 2012;47:596–602. doi:10.1097/RLI.0b013e318262292e [PubMed: 22836310]
24. Ash SY, Harmouche R, Vallejo DLL, et al. Densitometric and local histogram based analysis of computed tomography images in patients with idiopathic pulmonary fibrosis. *Respir Res* 2017;18:45. doi:10.1186/s12931-017-0527-8 [PubMed: 28264721]
25. Borie R, Crestani B, Dieude P, et al. The MUC5B variant is associated with idiopathic pulmonary fibrosis but not with systemic sclerosis interstitial lung disease in the European Caucasian population. *PLoS One* 2013;8:e70621. doi:10.1371/journal.pone.0070621 [PubMed: 23940607]

26. Noth I, Zhang Y, Ma S-F, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *lancet Respir Med* 2013;1:309–17. doi:10.1016/S2213-2600(13)70045-6 [PubMed: 24429156]
27. Peljto AL, Zhang Y, Fingerlin TE, et al. Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *Jama* 2013;309:2232–9. doi:10.1001/jama.2013.5827 [PubMed: 23695349]
28. Stock CJ, Sato H, Fonseca C, et al. Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis. *Thorax* 2013;68:436–41. doi:10.1136/thoraxjnl-2012-201786 [PubMed: 23321605]
29. Putman RK, Rosas IO, Hunninghake GM. Genetics and early detection in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2014;189:770–8. doi:10.1164/rccm.201312-2219PP [PubMed: 24547893]
30. Peljto AL, Zhang Y, Fingerlin TE, et al. Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* 2013;309:2232–9. doi:10.1001/jama.2013.5827 [PubMed: 23695349]
31. Oldham JM, Adegunsoye A, Khera S, et al. Underreporting of Interstitial Lung Abnormalities on Lung Cancer Screening Computed Tomography. *Ann Am Thorac Soc* 2018;15:764–6. doi:10.1513/AnnalsATS.201801-053RL [PubMed: 29490147]
32. Washko GR, Hunninghake GM, Fernandez IE, et al. Lung volumes and emphysema in smokers with interstitial lung abnormalities. *N Engl J Med* 2011;364:897–906. doi:10.1056/NEJMoa1007285 [PubMed: 21388308]
33. Sverzellati N, Guerci L, Randi G, et al. Interstitial lung diseases in a lung cancer screening trial. *Eur Respir J* 2011;38:392–400. doi:10.1183/09031936.00201809 [PubMed: 21233262]

**Key Messages:****What is the key question?**

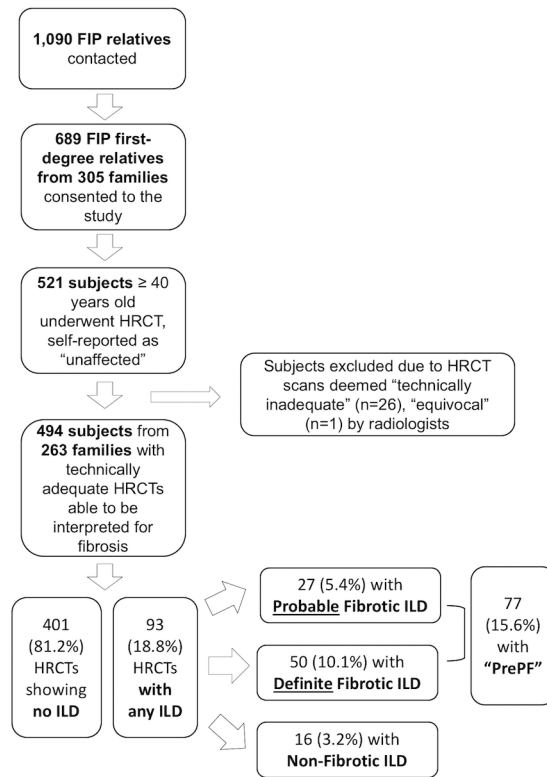
What are the risk factors for undiagnosed pulmonary fibrosis in first-degree relatives of Familial Interstitial Pneumonia Patients and can deep learning methods be utilized to detect it?

**What is the bottom line?**

Undiagnosed pulmonary fibrosis in first-degree relatives of Familial Interstitial Pneumonia Patients is common, associated with the *MUC5B* promoter variant, and deep learning methods can be utilized to detect it.

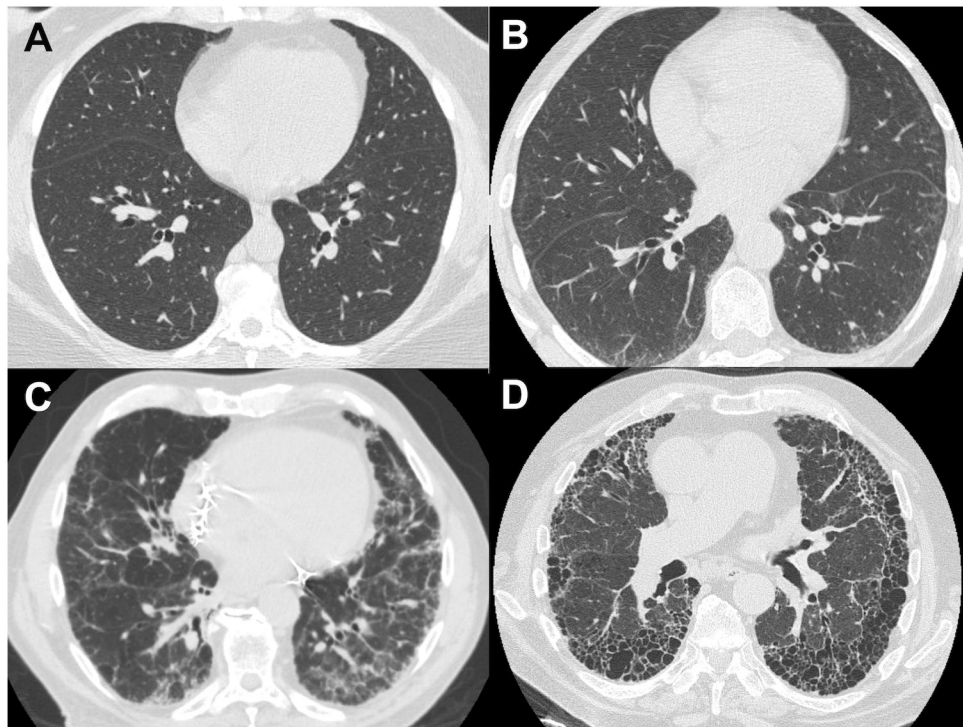
**Why read on?**

Our manuscript describes in detail the radiologic findings of early pulmonary fibrosis in at-risk subjects, genetic and clinical risk factors for it, and the application of a deep-learning algorithm on CT imaging of these subjects.



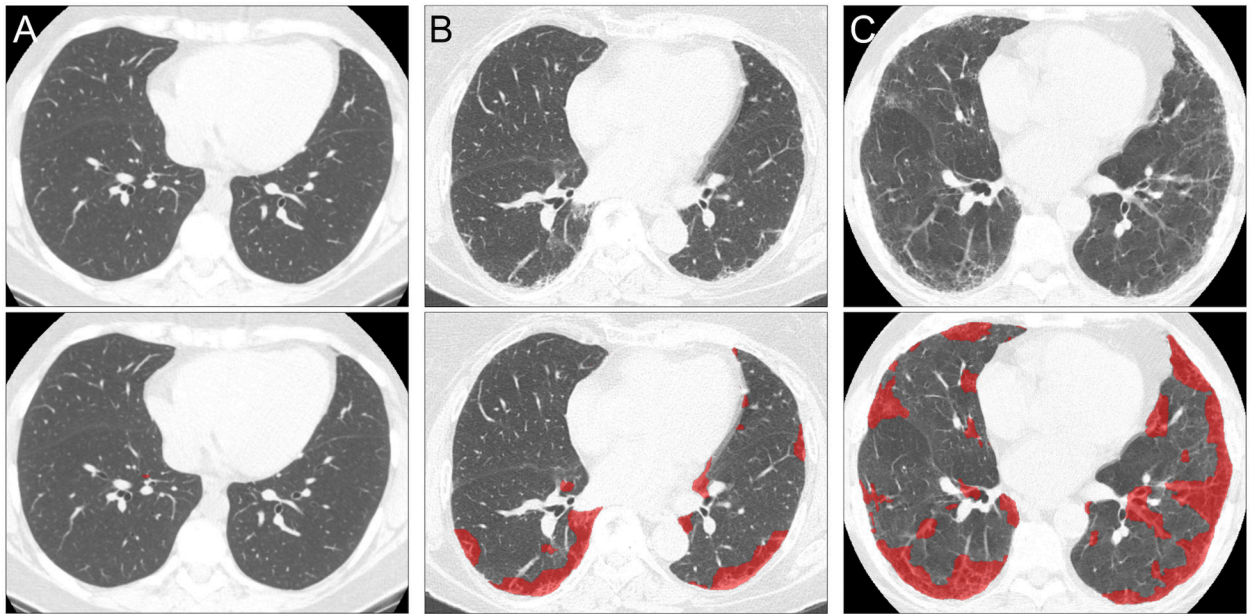
**Figure 1. Enrollment and Screening Flowchart**  
 Description of enrollment process and results for study subjects.



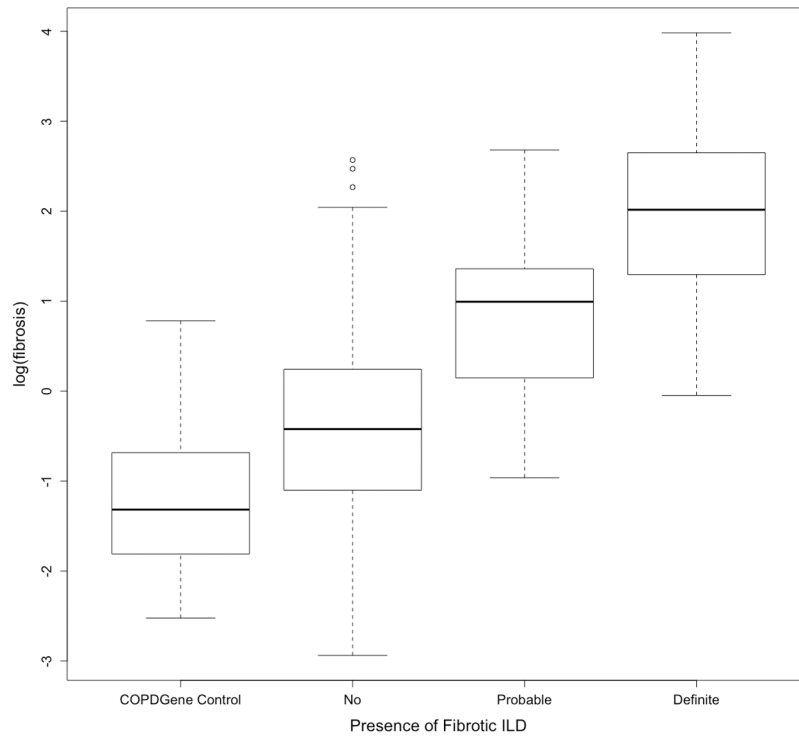


**Figure 2. Representative Images from Cohort Subjects**

A. High-resolution CT (HRCT) image of the chest from a study subject whose scan was read as normal, without signs of interstitial lung disease or fibrosis. B. HRCT image from subject who was categorized as having "Probable Fibrotic ILD." C. Representative HRCT image from subject who was characterized as having "Definite Fibrotic ILD." D. HRCT image from a case of previously diagnosed, established Idiopathic Pulmonary Fibrosis (IPF) in one of the study families.

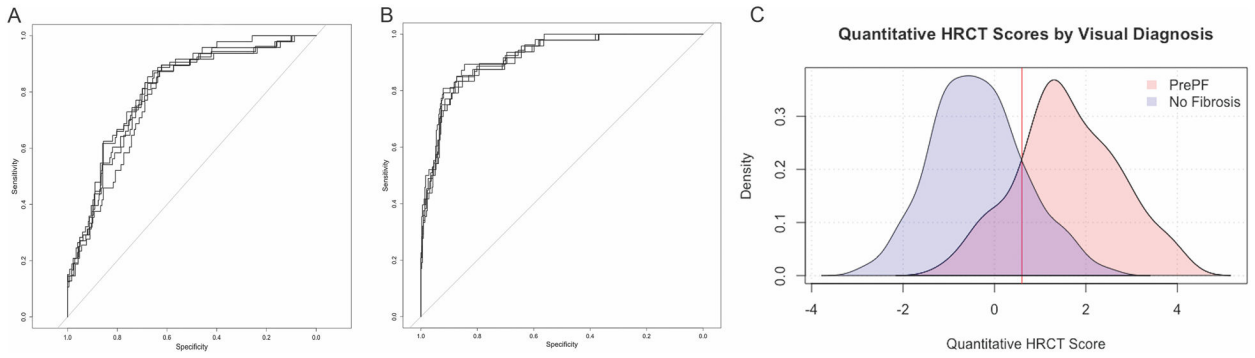


**Figure 3. Categorization of Regions of HRCT Images using Quantitative Methodology**  
Representative axial HRCT images visually assessed as “No Fibrosis” (A), “Probable Fibrotic ILD” (B) and “Definite Fibrotic ILD” (C). Below each is the corresponding quantitative HRCT results for the above scan; regions classified as fibrotic are shown in red. (A) “No Fibrosis” fibrosis extent 0.10% ( $\log(\text{fibrosis score}) = -2.30$ ); (B) “Probable Fibrotic ILD” fibrosis extent 12.46% ( $\log(\text{fibrosis score}) = 2.52$ ); (C) “Definite Fibrotic ILD” fibrosis extent 24.05% ( $\log(\text{fibrosis score}) = 3.18$ ).



**Figure 4. Fibrosis Score by Visual Diagnosis**

Boxplots of fibrosis scores based on quantitative HRCT assessment for each visual diagnosis category. Fibrosis score means were significantly different (ANOVA,  $p < 0.0001$ ) across groups defined by visual diagnosis. Comparison of fibrosis score between groups showed significant differences for all individual comparisons ( $p < 0.01$  for all).



**Figure 5. Receiver Operating Characteristic (ROC) Curves for Quantitative Imaging Measures of Fibrosis and PrePF**

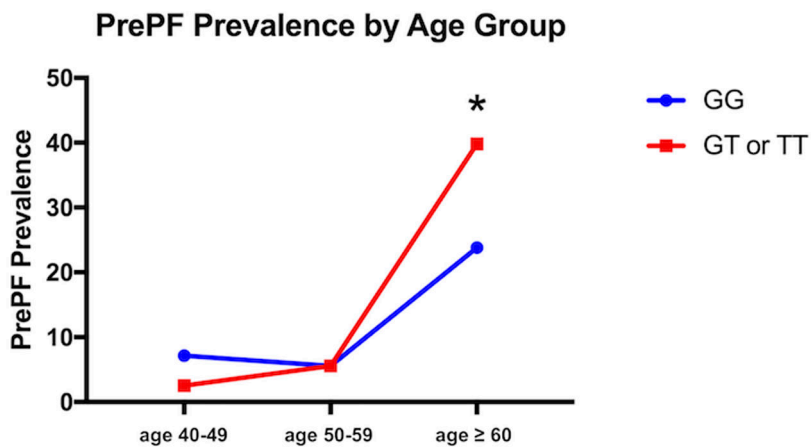
A. ROC curves for visual diagnosis compared to log %HAA. For this quantitative method, mean AUC was 0.80 (range 0.79–0.81). B. ROC Curves for visual diagnosis compared to fibrosis scores. ROC analysis showed that fibrosis score discriminates subjects with visual diagnosis of PrePF. Average area under the curve (AUC) in five-fold cross validation was 0.92 (range 0.91–0.93) and average accuracy, sensitivity, and specificity in the test partitions were 0.85 (range 0.81–0.88), 0.81 (range 0.71–0.92), and 0.86 (range 0.79–0.90), respectively. Optimal threshold for log fibrosis score was 0.60 (range 0.53 – 0.71), corresponding to 1.8% fibrotic area in examined lung. (C) Density plots of fibrosis scores for visually diagnosed PrePF (pink) and No Fibrosis (blue) scans—the fibrosis score optimal threshold is indicated with the red line (0.60).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 6. Prevalence of PrePF in FIP Siblings Cohort by Age and *MUC5B* Genotype**  
PrePF prevalence in this FIP siblings cohort increases by age, as shown in this graph. By age > 60 years, the prevalence of PrePF differed significantly based on *MUC5B* genotype (\* $p=0.02$ ). Subjects with the variant are depicted by the red line, while those without it are depicted with the blue line.

**Table 1.**

Screening Cohort Subject Characteristics.

	No Fibrosis (n=417)	PrePF (n=77)	p-value	OR [95% CI], controlling for family**	Mixed effects logistic regression p-value
Age, mean (SD), years	55.8 (8.7)	65.9 (10.1)	$2.35 \times 10^{-12}$	1.15 [1.09, 1.21]	$6.74 \times 10^{-7}$
Male, %	37%	49%	0.05	1.86 [0.91, 3.80]	0.09
Ever smoker, %	27%	44%	0.004	1.52 [0.74, 3.14]	0.26
<i>MUC5B</i> Promoter Variant (rs35705950), MAF (% subjects with variant) *	0.21 (40%)	0.29 (53%)	0.02	2.14 [1.00, 4.63]	0.05***
<i>TERT</i> Common Variant (rs2736100), MAF (% subjects with variant) *	0.45 (69%)	0.45 (67%)	0.92	0.69 [0.302, 1.58]	0.38****

\* DNA available on a total of 489 subjects (402 No Fibrosis and 75 PrePF subjects).

\*\* Odds ratios reported in this table were calculated comparing PrePF to no lung fibrosis in a mixed effects logistic regression model including age (as a continuous variable), male sex, ever smoker (yes/no), and *MUC5B* promoter variant (rs35705950) genotype. In the final row, the OR and p-value reported here are for the model with only the *TERT* common variant included (without the *MUC5B* variant).

\*\*\* In the reported model, rs35705950 was coded as a dominant allele; in log-additive genetic model, p=0.05, as well.

\*\*\*\* In this analysis, rs2736100 was coded as a dominant allele.



**Table 2.**

Visually identified patterns of CT Abnormalities in Scans with Probable or Definite Fibrotic ILD

<b>Total with Fibrotic ILD</b>	<b>77</b>
<b>Cranio-caudal distribution</b>	
Upper	11 (14%)
Middle	5 (7%)
Lower	54 (70%)
Diffuse	4 (5%)
<i>Not noted</i>	<i>3 (4%)</i>
<b>Axial distribution</b>	
Subpleural	67 (87%)
Diffuse	5 (6.5%)
Peribronchovascular	2 (2.5%)
<i>Not noted</i>	<i>3 (4%)</i>
<b>Honeycombing?</b>	<b>12 (15.6%)</b>
<b>CT pattern *</b>	
<b>UIP</b>	<b>59 (77.7%)</b>
Possible	41 (70%)
Probable	9 (15%)
Definite	9 (15%)
<b>NSIP</b>	<b>45 (58.4%)</b>
Possible	41 (91%)
Probable	2 (4%)
Definite	2 (4%)
<b>Sarcoidosis</b>	<b>3 (3.9%)</b>
<b>Hypersensitivity Pneumonitis (Possible)</b>	<b>14 (18.2%)</b>

\* Because a confident single diagnosis was relatively uncommon, most cases included consideration of several patterns. For this reason, the percentages add up to more than 100%.

**Table 3.**

## Dyspnea Questionnaire Data

<b>A. Breathlessness Responses for Cohort:</b>	<b>Yes</b>	<b>No</b>	<b>No answer</b>
Are you troubled by shortness of breath when hurrying on the level or walking up a slight hill?	121	344	31
Do you have to walk slower than people of your age on the level because of breathlessness?	46	422	28
Do you ever have to stop for breath when walking at your own pace on the level?	32	442	22
Do you ever have to stop for breath after walking about 100 yards (or after a few minutes)?	36	439	21
Are you too breathless to leave the house or breathless dressing or undressing?	7	470	19

<b>B. Breathlessness Responses by Visual CT Diagnosis:</b>	<b>PrePF (n=77)</b>	<b>No Fibrosis (n=419)</b>
Are you troubled by shortness of breath when hurrying on the level or walking up a slight hill?*	43 no 26 yes (37%) 8 no answer	301 no 95 yes (24%) 23 no answer
Do you have to walk slower than people of your age on the level because of breathlessness?	60 no 9 yes (13%) 8 no answer	362 no 37 yes (9.3%) 20 no answer
Do you ever have to stop for breath when walking at your own pace on the level?	64 no 6 yes (8.6%) 7 no answer	378 no 26 yes (6.4%) 15 no answer
Do you ever have to stop for breath after walking about 100 yards (or after a few minutes)?	66 no 6 yes (8.3%) 5 no answer	373 no 30 yes (7.4%) 16 no answer
Are you too breathless to leave the house or breathless dressing or undressing?	71 no 1 yes (1.4%) 5 no answer	399 no 6 yes (1.4%) 14 no answer

\* For this question,  $p=0.02$  for proportion responding "Yes" of those who answered question; all other questions have no significant difference between groups.

**Table 4.**

Subject Characteristics Based on Quantitative Fibrosis Score.

	Fibrosis score negative (n=292)	Fibrosis score positive (n=110)	Unadjusted p- value*	Adjusted, p-value
Age, mean (SD), years	54.8 (7.7)	64.1 (10.6)	$2.97 \times 10^{-14}$	$<2.0 \times 10^{-16}$
Male, n (%)	104 (36%)	49 (45%)	0.10	0.26
Ever smoker, n (%)	79 (27%)	40 (37%)	0.05	0.16
<i>MUC5B</i> Promoter Variant (rs35705950), MAF (% subjects with variant)**	0.21(38%)	0.27 (51%)	0.006	0.017

Clinical characteristics and genotype breakdown of subjects with quantitative HRCT analyses. The cutoff of 0.60 for the logarithm of fibrosis score is based on analyses presented in the text.

\* Unadjusted p-value compares characteristic between groups refer to t-tests (age) and Pearson's chi-square test (proportions). Adjusted p-values refer to p-values derived from regression of fibrosis score on age, male sex, smoking history, and *MUC5B* promoter variant.

\*\* In the reported model, rs35705950 coded as a dominant allele given small number of TT subjects. (MAF = minor allele frequency)