



UNIVERSITÀ DI PARMA

ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Deep Learning-based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-Resolution Computed Tomography

This is the peer reviewed version of the following article:

Original

Deep Learning-based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-Resolution Computed Tomography / Walsh, S. L. F.; Mackintosh, J. A.; Calandriello, L.; Silva, M.; Sverzellati, N.; Larici, A. R.; Humphries, S. M.; Lynch, D. A.; Jo, H. E.; Glaspole, I.; Grainge, C.; Goh, N.; Hopkins, P. M. A.; Moodley, Y.; Reynolds, P. N.; Zappala, C.; Keir, G.; Cooper, W. A.; Mahar, A. M.; Ellis, S.; Wells, A. U.; Corte, T. J.. - In: AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE. - ISSN 1535-4970. - 206:7(2022), pp. 883-891. [10.1164/rccm.202112-2684OC]

Availability:

This version is available at: 11381/2934167 since: 2022-11-21T16:51:54Z

Publisher:

NLM (Medline)

Published

DOI:10.1164/rccm.202112-2684OC

Terms of use:

Anyone can freely access the full text of works made available as "Open Access". Works made available

Publisher copyright

note finali coverage

(Article begins on next page)

Deep Learning-based Outcome Prediction in Progressive Fibrotic Lung Disease Using High-resolution Computed Tomography

Simon LF Walsh¹, John A Mackintosh², Lucio Calandriello³, Mario Silva⁴, Nicola Sverzellati⁴, Anna Rita Larici³, Stephen M Humphries⁵, David A Lynch⁵, Helen E Jo⁶, Ian Glaspole⁷, Christopher Grainge⁸, Nicole Goh^{9,10,11}, Peter M A Hopkins^{2,12}, Yuben Moodley¹³, Paul N Reynolds¹⁴, Christopher Zappala¹⁵, Gregory Keir¹⁶, Wendy A Cooper^{17,18}, Annabelle M Mahar¹⁷, Samantha Ellis¹⁹, Athol U Wells^{1,20}, Tamera J Corte⁶

¹National Heart and Lung Institute, Imperial College London, London, United Kingdom

²Queensland Lung Transplant Service, The Prince Charles Hospital, Queensland,

Australia, ³Dipartimento di Diagnostica per immagini, Radioterapia, Oncologia ed

Ematologia, Fondazione Policlinico Universitario A. Gemelli, IRCCS, Rome, Italy

⁴Scienze Radiologiche, Department of Medicine and Surgery (DiMeC), University of

Parma, Italy, ⁵Department of Radiology, National Jewish Health, Denver, CO, United

States, ⁶Respiratory Medicine, Royal Prince Alfred Hospital, New South Wales,

Australia, ⁷Department of Allergy and Respiratory Medicine, Alfred Hospital, Victoria,

Australia, ⁸Department of Respiratory Medicine, John Hunter Hospital, New South

Wales, Australia, ⁹Department of Respiratory and Sleep Medicine, Austin Health,

Melbourne, Victoria, Australia, ¹⁰Institute for Breathing and Sleep, Melbourne, Victoria,

Australia, ¹¹University of Melbourne, Melbourne, Victoria, Australia, ¹²Faculty of

Medicine, University of Queensland, Brisbane, Queensland, Australia, ¹³School of

Medicine & Pharmacology, University of Western Australia, Australia, ¹⁴Royal

Adelaide Hospital Chest Clinic, South Australia, Australia, ¹⁵Royal Brisbane and

Women's Hospital, Queensland, Australia, ¹⁶Department of Respiratory Medicine,

Princess Alexandra Hospital, Queensland, Australia, ¹⁷Tissue Pathology and Diagnostic Oncology, New South Wales Health Pathology, Royal Prince Alfred Hospital, Sydney, Australia, ¹⁸School of Medicine, University of Sydney, Sydney, Australia, ¹⁹Department of Radiology, Alfred Health, Melbourne, Australia, ²⁰Interstitial Lung Disease Unit, Royal Brompton Hospital, London, United Kingdom

Corresponding author: Simon L.F Walsh, MD FFRRCSI

Address: National Heart and Lung Institute,
Imperial College,
Guy Scadding Building,
Dovehouse St, Chelsea,
London SW3 6LY
United Kingdom

Email s.walsh@imperial.ac.uk

Author contributions:

Study concept	SLFW, JM, HEJ, AUW, TJC
Patient data collection	JAM, HEJ, LC, MS, NS
SOFIA development	SLFW
Scoring of patient cases	LC, MS
Data analysis and critique	SLFW, AUW, JAM, TJC
Writing of the manuscript	SLFW, AUW, JAM, TJC
Editing and approving the manuscript	SLFW, JAM, LC, MS, NS, SMH, DAL, HEJ, IG, CG, NG, PMAH, YM, PNR, CZ, GK, WAC, ANM, SE, AUW, TJC

Short running head: Deep learning-based prognostication in fibrotic lung disease

Funding: SLFW is funded by a National Institute of Health Research Clinician Scientist Fellowship CS-2018-18-ST2-004

Descriptor: 9.23 Interstitial Lung Disease

Full word count: 2769

This article has an online data supplement, which is accessible from this issue's table of content online at www.atsjournals.org

At a Glance

What is the current scientific knowledge on this subject?

Deep learning has been successfully applied to diagnosis in patients with suspected fibrotic lung disease, providing radiologists with decision support when expertise is limited. However, it is currently not possible to reliably predict progressive fibrotic lung disease in an individual patient.

What does this study add to the field?

We demonstrate the prognostic utility of a deep learning algorithm, validated in the identification of UIP-like features in a large population of patients with suspected IPF, drawn from a national IPF registry. In principle, this tool could be used to identify patients at risk of developing progressive fibrotic lung disease using baseline HRCT imaging of the chest.

ABSTRACT

RATIONALE

Reliable outcome prediction in patients with fibrotic lung disease using baseline high-resolution computed tomography (HRCT) data remains challenging.

OBJECTIVES

To evaluate the prognostic accuracy of a deep learning algorithm (SOFIA), trained and validated in the identification of UIP-like features on HRCT (UIP probability), in a large cohort of well characterised patients with progressive fibrotic lung disease, drawn from a national registry.

METHODS

SOFIA and radiologist-UIP probabilities were converted to PLOPED-based UIP probability categories (UIP not included in the differential: 0-4%, low probability of UIP: 5–29%, intermediate probability of UIP: 30–69%, high probability of UIP: 70–94%, and pathognomonic for UIP:95-100%) and their prognostic utility assessed using Cox proportional hazards modelling.

MEASUREMENTS AND MAIN RESULTS

On multivariable analysis adjusting for age, gender, guideline based radiologic diagnosis and disease severity (using total ILD extent on HRCT, %predicted FVC, DLco or the CPI), only SOFIA-UIP probability PLOPED categories predicted survival. SOFIA-PLPED UIP probability categories remained prognostically significant in patients considered indeterminate (n=83) by expert radiologist consensus (HR1.73, $p<0.0001$, 95%CI 1.40-2.14). In patients undergoing surgical lung biopsy (SLB)

(n=86), after adjusting for guideline-based histologic pattern and total ILD extent on HRCT, only SOFIA-PIOPED probabilities were predictive of mortality (HR1.75, $p<0.0001$, 95%CI 1.37-2.25).

CONCLUSIONS

Deep learning-based UIP probability on HRCT provides enhanced outcome prediction in patients with progressive fibrotic lung disease when compared to expert radiologist evaluation or guideline-based histologic pattern. In principle this tool may be useful in multidisciplinary characterisation of fibrotic lung disease. The utility of this technology as a decision support system when ILD expertise is unavailable requires further investigation.

Abstract word count: 257

Introduction

On initial evaluation, a confident diagnosis of idiopathic pulmonary fibrosis (IPF) identifies patients who, on average, have a worse outcome than those with other forms of ILD, and require anti-fibrotic therapy (1). High resolution computed tomography (HRCT) of the chest plays a pivotal role in diagnosing IPF and the current IPF Guideline HRCT classification is anchored to the likelihood of underlying usual interstitial pneumonia (UIP), when IPF is suspected (2). However, a limitation with this classification is that it requires that patients are assigned a single category based on the predominant HRCT pattern; it does not account for background UIP features when the predominant HRCT pattern suggests an alternative diagnosis. In principle, a more rigorous evaluation of the likelihood of UIP features across all four guideline-based HRCT categories may provide additional prognostic information in patients with fibrotic lung disease. This is supported by recent data from the INBUILD study, which reported that patients with UIP-like fibrosis in non-IPF in the placebo arm had the same rate of forced vital capacity (FVC) decline as untreated IPF (3).

Deep learning is a branch of machine learning which can autonomously detect features in CT images and map them to simple classifications such as outcome (4, 5). In this study, we test the prognostic utility of a deep learning algorithm (SOFIA), developed, and validated in the identification of UIP-like features on HRCT, in patients drawn from the Australian IPF registry (AIPFR) (6). Patients were enrolled in the AIPFR in the belief that they had IPF, but subsequent clinical evaluation excluded IPF in a subgroup, based on the presence of a connective tissue disease or hypersensitivity pneumonitis. Therefore, the registry population

includes patients with both idiopathic diseases, and a patient subset that matches patients enrolled in recent non-IPF anti-fibrotic trials (3).

Methods

Australian IPF Registry patient population

Patients enrolled in the Australian IPF Registry (AIPFR) with HRCT imaging suitable for SOFIA analysis were eligible for this study. The study has ethical approval from the Sydney Local Health District (protocol no. X14-0264). Details of the AIPFR, which commenced in 2012, have been published previously (6). In brief, registrants were referred by their treating physician with a clinical diagnosis of IPF. For each patient in this first stage of recruitment, clinical, radiologic, and where available, histologic data were reviewed centrally by a panel of three expert radiologists, three histopathologists and three expert ILD physicians and assigned an IPF diagnosis (IPF, probable IPF, possible IPF, alternative diagnosis) based on the 2011 ATS/ERS/JRS/ALAT IPF guideline statement (2). Diagnostic disagreement was resolved by panel discussion. Baseline and longitudinal data were collected for the duration of a participant's enrolment. For the current study, the follow-up period was transplant-free survival, calculated from the date of the patient's HRCT. Disease progression at 12 months was defined as any of the following occurring within 12 months of HRCT acquisition date: (A) a decline in FVC percentage predicted of 10% or more or DLCO percentage predicted of 15% or more that was sustained at 18 months, (B) death or (C) transplantation.

Semiquantitative HRCT evaluation

A detailed description of HRCT pattern definitions and the HRCT scoring method can be found in the online repository. Briefly, two thoracic radiologists (LC, MS, 10- and 12-years' experience) scored HRCTs for each patient on total ILD extent, the extent of four interstitial patterns (ground glass opacification, reticulation, honeycombing and consolidation), emphysema and the severity and extent of traction bronchiectasis. Each HRCT was also assigned a diagnostic probability (censored at 5% and summing to 100%) for each of radiologic diagnosis categories specified by the 2018 Clinical Practice Guideline for IPF e.g., UIP:75%, probable UIP: 25%, indeterminate for UIP: 0% and alternative diagnosis: 0%.

SOFIA-based image evaluation

SOFIA (Systematic Objective Fibrotic Imaging Analysis Algorithm) is a deep convolutional neural network loosely based on the Inception-ResNet-v2 architecture proposed by Szegedy, which combines Inception modules with residual connections. Development and validation of this algorithm has been published previously (4, 7) . Briefly, SOFIA was trained on a database of 420,096 unique 4-HRCT slice montages from 1157 fibrotic lung disease specific HRCTs derived from two tertiary referral centres for ILD and validated against the performance of 92 thoracic radiologists on a test cohort of 150 HRCTs from a third institution (8). The algorithm's input is four HRCT slice montage and its output a set of continuous numbers from 0 to 1 each representing a probability of each of the UIP diagnosis categories, whose sum is 1.0 (e.g., definite UIP 0.985, probable UIP: 0.011, indeterminate: 0.002, alternative diagnosis 0.002) (Figure 1). SOFIA generates up to 500 unique montages per HRCT

scan and its final prediction for a single HRCT is the average probability assigned for each diagnostic category, for these montages (Figure 2).

Statistical analysis

Statistical analyses were performed using STATA version 16 (StataCorp, College Station, Texas) and the Python package, SciPy version 0.19.1. Data are given as means with standard deviations (SD), medians with interquartile range (IQR) or as the number of patients and percentage where appropriate. P values <0.05 were considered statistically significant.

Radiologist-based UIP likelihoods and SOFIA-based UIP probabilities for definite UIP were first examined as continuous variables and then standardized by converting them to diagnostic probability categories using the PIOPED diagnostic criteria (UIP not included in the differential: 0-4%, low probability of UIP: 5–29%, intermediate probability of UIP: 30–69%, high probability of UIP: 70–94%, and pathognomonic for UIP:95-100%). The PIOPED criteria were originally developed for categorical estimation of probability of pulmonary embolus but have also recently been used to evaluate diagnostic agreement in ILD (9, 10).

Cox proportional hazards modelling was used to determine crude and adjusted hazards ratios. Transplant-free survival was the outcome and the survival period for each patient was calculated from the date of the registry HRCT to 20th November 2020. We tested the assumptions of proportional hazards by visual inspection of the log-log plot of survival, comparison of the Kaplan-Meier observed survival curves with the Cox predicted curves for the same variable and graphical and formal analysis of Schoenfeld residuals (analysis not shown). Results are reported as HR, 95% CIs, and p values. Logistic regression was performed to

investigate associations between SOFIA-based UIP probabilities and disease progression at 12 months. Results are reported as ORs, 95% CIs, and p values.

Results

2018 Clinical Practice Guideline-based prognostic separation

A total of 515/868 patients had baseline HRCTs for analysis. Of these, 504 were amenable to SOFIA-based analysis (Table 1). Frequency of SOFIA assigned UIP diagnosis categories were as follows: UIP:164, probable UIP:214, indeterminate for UIP:55 and alternative diagnosis:71. Interobserver agreement between SOFIA and consensed radiologist based UIP diagnosis categories was fair (Kw 0.39). Mean probabilities of first-choice diagnoses based on SOFIA probability scores are shown in the Supplementary Appendix, Table A1.

SOFIA-based UIP probabilities

On bivariable analysis guideline-based diagnosis categories determined by both SOFIA and radiologist's consensus were predictive of transplant free survival (Table 2). However, on multivariate analysis, adjusting for total ILD extent, neither of these two assessments of UIP diagnostic category were predictive of transplant free survival (HR 1.02, p=0.809, 95%CI 0.89-1.15 and HR 1.04, p=0.454, 95%CI 0.94-1.15 respectively).

Only SOFIA-UIP probabilities (% probability of definite UIP on HRCT expressed in 5% increments, n=504, 0.29±0.33, 0.0-0.99) were predictive of transplant free survival on bivariable analysis with consensed radiologists-UIP probabilities (HR 1.07, p<0.0001, 95%CI 1.05-1.09) and remained predictive of when adjusting for total ILD extent (HR 1.06, p<0.0001, 95%CI 1.04-1.08).

SOFIA and radiologist-UIP probabilities were converted to PIOPED-based UIP probability categories (Figure 3). Only SOFIA-PIOPED UIP probability categories were predictive of transplant free survival on bivariable analysis with radiologist-PIOPED UIP probability categories (HR1.45, $p < 0.0001$, 95%CI 1.33-1.59) and remained predictive of transplant free survival when adjusted for total ILD extent (HR1.31, $p < 0.0001$, 95%CI 1.19-1.44) (Table 3). On multivariable analysis adjusting for age, gender and total ILD extent, only SOFIA-PIOPED UIP probability categories predicted transplant free survival (Table 4) and were maintained when adjusted for disease severity using %predicted FVC, DLco, CPI and GAP stage (Supplementary Appendix, Table A3-A4). The prognostic utility of SOFIA-PIOPED UIP probability categories was also maintained for subgroup analysis of each guideline-based diagnosis category as assigned by radiologists (Table 5). In particular, 21/83 HRCTs considered indeterminate for UIP by expert radiologist consensus were re-classified as having an intermediate probability, high probability or pathognomonic of UIP by the algorithm (Figure 4).

INBUILD stratification

All 504 patients with an HRCT amenable to SOFIA analysis had a registry multidisciplinary team diagnosis. Using the consensus radiologic diagnosis provided by the study radiologists, these 504 patients were stratified into two groups:

1. *Group 1, UIP-like fibrotic pattern, n=331*: HRCT showing “UIP” or “probable UIP” or surgical lung biopsy showing “possible”, “probable” or “definite UIP”
2. *Group 2, Other fibrotic patterns, n=173*: remaining patients

SOFIA-PIOPED UIP probability categories and total ILD extent were the only CT variables that predicted transplant free survival in group 1 (n=331, hereafter called 'UIP-like fibrotic patterns') subgroup and in group 2 (n=177, hereafter called 'other fibrotic patterns') (Table 6). The prognostic utility of SOFIA-PIOPED UIP probability categories was maintained in these subgroups, adjusting for disease severity using %predicted FVC, DLco and the CPI (Supplementary Appendix, Table A4).

Predicting disease progression at 12 months

Rapidly progressive or stable disease could be calculated in 463 patients (progressive, n=98, stable, n=365). Increasing SOFIA-PIOPED probability was associated with a 1.58-fold increased risk of progression at 12 months when adjusted for total ILD extent (OR 1.58, $p < 0.0001$, 95%CI 1.28-1.85). These associations were maintained on subgroup analysis in patients with 'UIP-like disease' on HRCT (n=308, OR 1.48, $p = 0.001$, 95%CI 1.19-1.85) and patients with 'other fibrotic patterns' on HRCT (n=150, OR 1.67, $p = 0.014$, 95%CI 1.11-2.51).

Subgroup analysis in patients undergoing surgical lung biopsy.

A total of 86/868 patients underwent surgical lung biopsy (not UIP:6, possible UIP:11, probable UIP:16, UIP:53). On subgroup analysis in these patients, adjusting for guideline-based histologic pattern and total ILD extent on HRCT, only SOFIA-PIOPED probabilities were predictive of mortality (Table 7a). Increasing SOFIA-PIOPED probability category was associated with a 2.37-fold increased likelihood of progressive disease at 12 months (Table 7b).

Discussion

In this study, we demonstrate the prognostic utility of a deep learning algorithm, previously trained, and validated in the identification of UIP-like features, in a large cohort of patients with fibrotic lung disease, drawn from a national IPF registry (4). A key strength of this study is the application of the algorithm to a new imaging dataset consisting of HRCT scans performed at multiple institutions using different CT scanners and HRCT protocols. The algorithm provided prognostic power uniformly across all patient sub-groups, stratified based on CT pattern, underlying cause (i.e., idiopathic versus non-idiopathic disease) and in patients who underwent surgical lung biopsy (SLB).

In IPF, one value of making a confident diagnosis is that it enables physicians to identify, using baseline information, patients who are likely to progress (11). HRCT plays a central role in the initial evaluation of patients with fibrotic lung disease and in the correct clinical setting, guideline-based CT classification is linked to the likelihood of underlying UIP; a definite or probable UIP pattern is associated with a poor outcome. However, one difficulty with this classification is that subtle or limited UIP-like features may be overlooked if the dominant HRCT pattern suggests an alternative diagnosis such as hypersensitivity pneumonitis. In principle, a more rigorous evaluation of UIP-like features across all four guideline categories may improve prognostic discrimination. Support for this hypothesis comes from INBUILD, where patients with UIP-like fibrosis in non-IPF have similar outcomes as untreated IPF (3, 12).

Traditional visual-based HRCT assessment has several well-documented limitations, including high levels of interobserver variability and poor reproducibility.

Also, the human brain has a natural tendency to take ambiguous visual information and organise it into a single recognizable pattern (8, 13, 14). Familiar patterns stand out prominently in the foreground while less familiar or incongruent patterns recede into the background. Our data supports this hypothesis; radiologists tended to default to a binary categorisation of UIP likelihood in most cases with 72.8% of HRCTs being assigned either a 0% or 100% diagnostic likelihood of UIP. The result is that prognostic information is lost in cases with intermediate likelihoods of UIP. In contrast, computer-based assessment is not subject to these perceptual biases and can capture the full range of UIP probabilities as a continuous variable, regardless of the dominant HRCT pattern.

Although SOFIA appears to provide uniformity of prognostic power across all four guideline-based CT categories, three sub-group analyses warrant discussion. First, although patients were initially placed in the AIPFR with a clinical diagnosis of IPF, subsequent evaluation established that in addition to a large sub-group of patients where no primary cause could be identified, there was a smaller sub-group in which IPF was excluded by rigorous multidisciplinary review, with the presence of a connective tissue disease or hypersensitivity pneumonitis. Therefore, our study population covers both idiopathic diseases, and a patient sub-group that matches those enrolled in recent non-IPF anti-fibrotic therapy trials (3). SOFIA provided similar prognostic power regardless of clinical diagnosis. Second, the algorithm provided prognostic separation of patients with indeterminate HRCT appearances, highlighting the utility of the algorithm in patients where visual HRCT assessment was considered unhelpful (Figure 3). Lastly, in a subgroup of 86 patients who underwent surgical lung biopsy, histologic UIP classification provided no additional prognostic information once algorithmic predictions were accounted for, although it

should be noted that this subgroup was mostly made up of UIP or probable/possible UIP cases (80/86). Furthermore, since surgical lung biopsy also provides important diagnostic information, SOFIA should not be considered a replacement for histologic evaluation at this time.

Based on recent studies, in many countries, anti-fibrotic treatment is only approved for disorders other than IPF when traditional therapy has failed. This means delaying intervention until progression has been observed. Although UIP-like disease is progressive in most patients, there are no data available that reliably predict outcome in patients with other fibrotic patterns on HRCT (5). Furthermore, predicting progressive disease is especially problematic when baseline disease extent is less severe. The results of our study suggest that deep learning-based algorithms such as SOFIA, have the potential to address this difficulty at least partially, by providing accurate, reproducible outcome prediction in patients with fibrotic lung disease, using their baseline imaging data.

It is important to highlight how SOFIA differs from quantitative CT (QCT) (15-17). Traditional QCT tools rely on 'feature engineering'; the computer is trained by human experts to quantify pre-specified HRCT patterns. A limitation of this supervised approach is that the training process is confined to image features that are known *a priori*; it misses the opportunity to identify novel or visually inaccessible patterns of disease (16). Deep learning algorithms such as SOFIA overcome this difficulty by automatically learning the most predictive features directly from the images and mapping these features to the desired output (18). The increased prognostic accuracy observed using SOFIA-based UIP probabilities over and above features traditionally evaluated by radiologists, suggests that additional prognostic

signal is being captured by the algorithm including signal that is undetectable to the human eye. This may include features of lung senescence which is a known driver in the development of IPF (19, 20) . Likewise, subtle vascular volume abnormalities and lung volume shrinkage may also be incorporated into SOFIA's output predictions. Finally, in the past, combining variables from different domains to create multidimensional prognostic models in interstitial lung disease has proven more fruitful than focusing on the stand-alone value of variables in isolation (21, 22). This suggests that the clinical utility of SOFIA might be improved by integrating its outputs with lung function or -omic-based biomarkers.

Our study has several limitations. The relative opacity of neural networks, upon which deep learning is based, has meant that this technology is occasionally viewed as a “black box” (5). Decoding the image features that deep neural networks use to make predictions will be crucial for biomarker development in patients with established fibrotic lung disease. Algorithm interpretability will also be necessary to appraise biomarker plausibility before successful integration into clinical practice. Also, although our data were generated from a large national IPF registry, prospective clinical utility studies which demonstrate clear patient benefit over current best practice will be needed before this technology can be implemented in clinical practice.

In conclusion, we have demonstrated the prognostic utility of a deep learning algorithm in patients with progressive fibrotic lung disease enrolled in a national IPF registry. In principle, the algorithm's output, the probability of UIP on HRCT, could be incorporated in multidisciplinary characterisation of fibrotic lung disease by providing

enhanced outcome prediction as well as by providing decision support to centres where ILD expertise is unavailable.

References

1. Richeldi L, Collard HR, Jones MG. Idiopathic pulmonary fibrosis. *Lancet*. 2017;389(10082):1941-52.
2. Raghu G, Remy-Jardin M, Myers JL, Richeldi L, Ryerson CJ, Lederer DJ, et al. Diagnosis of Idiopathic Pulmonary Fibrosis. An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline. *Am J Respir Crit Care Med*. 2018;198(5):e44-e68.
3. Flaherty KR, Wells AU, Cottin V, Devaraj A, Walsh SLF, Inoue Y, et al. Nintedanib in Progressive Fibrosing Interstitial Lung Diseases. *New England Journal of Medicine*. 2019.
4. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. *Lancet Respir Med*. 2018;6(11):837-45.
5. Walsh SLF, Humphries SM, Wells AU, Brown KK. Imaging research in fibrotic lung disease; applying deep learning to unsolved problems. *Lancet Respir Med*. 2020.
6. Jo HE, Glaspole I, Grainge C, Goh N, Hopkins PM, Moodley Y, et al. Baseline characteristics of idiopathic pulmonary fibrosis: analysis from the Australian Idiopathic Pulmonary Fibrosis Registry. *Eur Respir J*. 2017;49(2).
7. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:160207261v2*. 2016.

8. Walsh SL, Calandriello L, Sverzellati N, Wells AU, Hansell DM, Consort UIPO. Interobserver agreement for the ATS/ERS/JRS/ALAT criteria for a UIP pattern on CT. *Thorax*. 2016;71(1):45-51.
9. Pioped Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA*. 1990;263(20):2753-9.
10. Ryerson CJ, Corte TJ, Lee JS, Richeldi L, Walsh SLF, Myers JL, et al. A Standardized Diagnostic Ontology for Fibrotic Interstitial Lung Disease. An International Working Group Perspective. *Am J Respir Crit Care Med*. 2017;196(10):1249-54.
11. Walsh SLF, Lederer DJ, Ryerson CJ, Kolb M, Maher TM, Nusser R, et al. Diagnostic Likelihood Thresholds That Define a Working Diagnosis of Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2019;200(9):1146-53.
12. Brown KK, Martinez FJ, Walsh SLF, Thannickal VJ, Prasse A, Schlenker-Herceg R, et al. The natural history of progressive fibrosing interstitial lung diseases. *Eur Respir J*. 2020;55(6).
13. Watadani T, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, et al. Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology*. 2013;266(3):936-44.
14. Walsh SLF, Kolb M. Radiological diagnosis of interstitial lung disease: is it all about pattern recognition? *Eur Respir J*. 2018;52(2).

15. Humphries SM, Swigris JJ, Brown KK, Strand M, Gong Q, Sundry JS, et al. Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis. *Eur Respir J*. 2018;52(3).
16. Salisbury ML, Lynch DA, van Beek EJ, Kazerooni EA, Guo J, Xia M, et al. Idiopathic Pulmonary Fibrosis: The Association between the Adaptive Multiple Features Method and Fibrosis Outcomes. *Am J Respir Crit Care Med*. 2017;195(7):921-9.
17. Kim HG, Tashkin DP, Clements PJ, Li G, Brown MS, Elashoff R, et al. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. *Clin Exp Rheumatol*. 2010;28(5 Suppl 62):S26-35.
18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-44.
19. Schafer MJ, White TA, Iijima K, Haak AJ, Ligresti G, Atkinson EJ, et al. Cellular senescence mediates fibrotic pulmonary disease. *Nat Commun*. 2017;8:14532.
20. Faner R, Rojas M, Macnee W, Agusti A. Abnormal lung aging in chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*. 2012;186(4):306-13.
21. Goh NS, Desai SR, Veeraraghavan S, Hansell DM, Copley SJ, Maher TM, et al. Interstitial lung disease in systemic sclerosis: a simple staging system. *Am J Respir Crit Care Med*. 2008;177(11):1248-54.

22. Ley B, Ryerson CJ, Vittinghoff E, Ryu JH, Tomassetti S, Lee JS, et al. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med.* 2012;156(10):684-91.

Variable	
Age (Median (IQR))	69.7 (64.9 – 76.1)
Gender (Female/Male)	168 (32.6%) / 347(67.4%)
Lung function	
FVC% (Median (IQR))	79.6 (66.1 – 93.5)
DLCO% (Median (IQR))	46.9 (37.9 – 58.7)
CPI (Median (IQR))	46.1 (36.8 – 54.8)
Anti-Fibrotic Therapy (No/Yes)	337 (65.4%)/178 (34.6%)
Smoking History (Ever/Never)	333 (64.7%)/182(35.3%)
IPF Diagnosis Category (ATS 2011) * (n=515)	
Definite IPF	249 (48.3%)
Probable IPF	26 (5.0%)
Possible IPF	110 (21.4%)
Inconsistent IPF	126 (24.5%)
No consensus reached	4 (0.8%)
UIP Diagnosis Category (ATS 2018) ** (Radiologist consensus, n=515)	
Definite UIP	136 (26.4%)
Probable UIP	167 (32.4%)
Indeterminate UIP	84 (16.3%)
Alternative diagnosis	128 (24.9%)
UIP Diagnosis Category (ATS 2018) (SOFIA-analysis, n=504 #)	
Definite UIP	164 (32.5%)
Probable UIP	214 (42.5%)
Indeterminate UIP	55 (10.9%)
Alternative diagnosis	71 (14.1%)
Histologic Diagnosis (ATS 2011) * (n=86)	
Definite UIP	53 (61.6%)
Probable UIP	16 (18.6%)
Possible UIP	11 (12.8%)
Alternative to UIP	6 (7.0%)

Table 1. Baseline Demographics and Centralised MDM Review Characteristics.

*Multidisciplinary team meeting diagnoses were made based on the 2011 IPF guideline statement. **Thoracic radiologist scores (LC, MS). #11 HRCTs were not amenable to SOFIA analysis.

Variable (n=504)	HR	P Value	CI 95%
SOFIA 2018 diagnoses	1.23	0.003	1.07-1.40
Radiologists 2018 diagnosis**	1.14	0.015	1.03-1.27

Table 2. Bivariable Cox proportionate hazards model including SOFIA-based guideline diagnoses and consensed radiologist's guideline diagnoses. **Weighted kappa 0.64

Variable (n=504)	HR	P Value	CI 95%
SOFIA PIOPED UIP probability categories	1.31	<0.0001	1.19-1.44
Radiologists PIOPED UIP probability categories**	1.07	0.067	0.99-1.16
Total ILD extent (1% increments)	1.02	<0.0001	1.02-1.03

Table 3. Multivariable Cox proportionate hazards model including SOFIA PIOPED UIP probability categories and radiologists PIOPED UIP probability categories, adjusting for total ILD extent on HRCT. **weighted kappa 0.79

Variable (n=504)	HR	P Value	CI 95%
Age	1.01	0.042	1.00-1.03
Gender	1.37	0.007	1.09-1.72
SOFIA PIOPED UIP probability categories	1.29	<0.0001	1.17-1.41
Radiologists PIOPED UIP probability categories	1.08	0.052	0.99-1.16
Total ILD extent (1% increments)	1.02	<0.0001	1.02-1.03

Table 4. Multivariable analysis Cox proportionate hazards model adjusting for age, gender and total ILD extent.

Radiologic diagnosis category*	HR	P Value	CI 95%
UIP (n=135)	1.34	<0.0001	1.15-1.56
Probable UIP (n=162)	1.47	<0.0001	1.25-1.72
Indeterminate (n=83)	1.73	<0.0001	1.40-2.14
Alternative diagnosis (n=124)	1.44	0.003	1.13-1.83

Table 5. Subgroup analysis of SOFIA PIOPED UIP probabilities in radiologic diagnosis subgroups as assigned by thoracic radiologists.

Variable	HR	P Value	CI 95%
<i>UIP-like fibrotic patterns (n=331/338*)</i>			
SOFIA PIOPED UIP likelihood	1.34	<0.0001	1.22-1.48
Total ILD extent (1% increments)	1.02	<0.0001	1.01-1.03
<i>Other fibrotic patterns (n=173/177*)</i>			
SOFIA PIOPED UIP likelihood	1.34	<0.0001	1.12-1.61
Total ILD extent (1% increments)	1.02	<0.0001	1.01-1.04

Table 6. Cox proportionate hazards models adjusting for disease severity based on total ILD extent in INBUILD subgroups. *Limited to patients who had an HRCT amenable to SOFIA analysis.

Variable (n=83*)	HR	P Value	CI 95%
SOFIA PIOPED UIP probability categories	1.75	<0.0001	1.37-2.25
Guideline histological pattern	1.29	0.109	0.94-1.78
Total ILD extent (1% increments)	1.01	0.237	0.99-1.02

Table 7a. Cox proportionate hazards model of SOFIA PIOPED UIP probabilities patients who underwent surgical lung biopsy (n=86). *In three patients who underwent SLB, the HRCT was not amenable to SOFIA analysis.

Variable (n=83*)	OR	P Value	CI 95%
SOFIA PIOPED UIP probability categories	2.37	0.005	1.30-4.35
Guideline histological pattern	1.51	0.309	0.68-3.35
Total ILD extent (1% increments)	1.07	0.003	1.02-1.12

Table 7b. Associations between progressive disease at 12 months and SOFIA PIOPED UIP probabilities patients who underwent surgical lung biopsy (n=86). OR=odds ratio. *In three patients who underwent SLB, the HRCT was not amenable to SOFIA analysis.

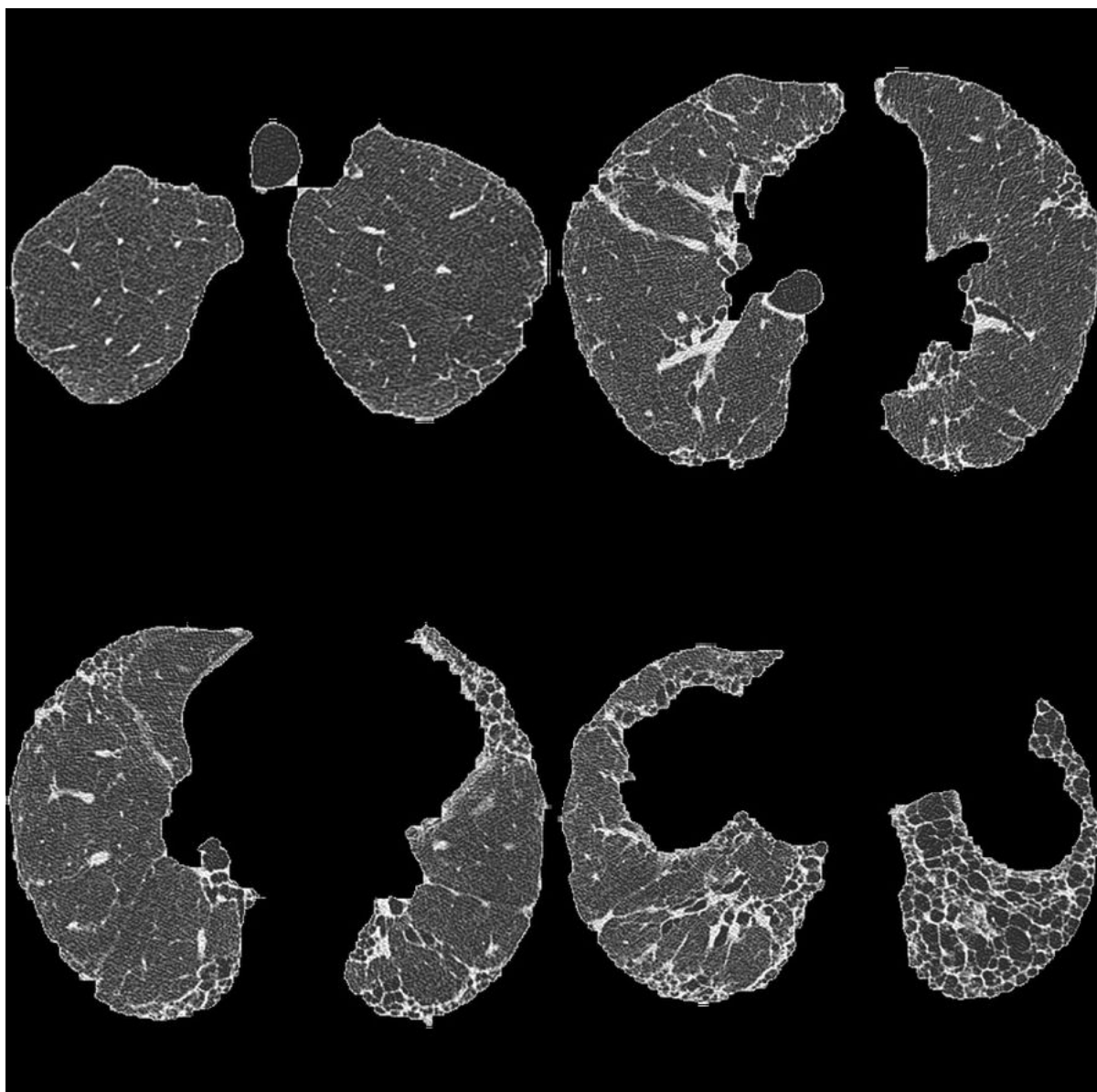


Figure 1. An example of a 4-slice montage created from an HRCT showing typical UIP. (SOFIA analysis: UIP:0.9972, Probable UIP: 0.0022, Indeterminate for UIP: 0.0008, alternative diagnosis: 0.000)

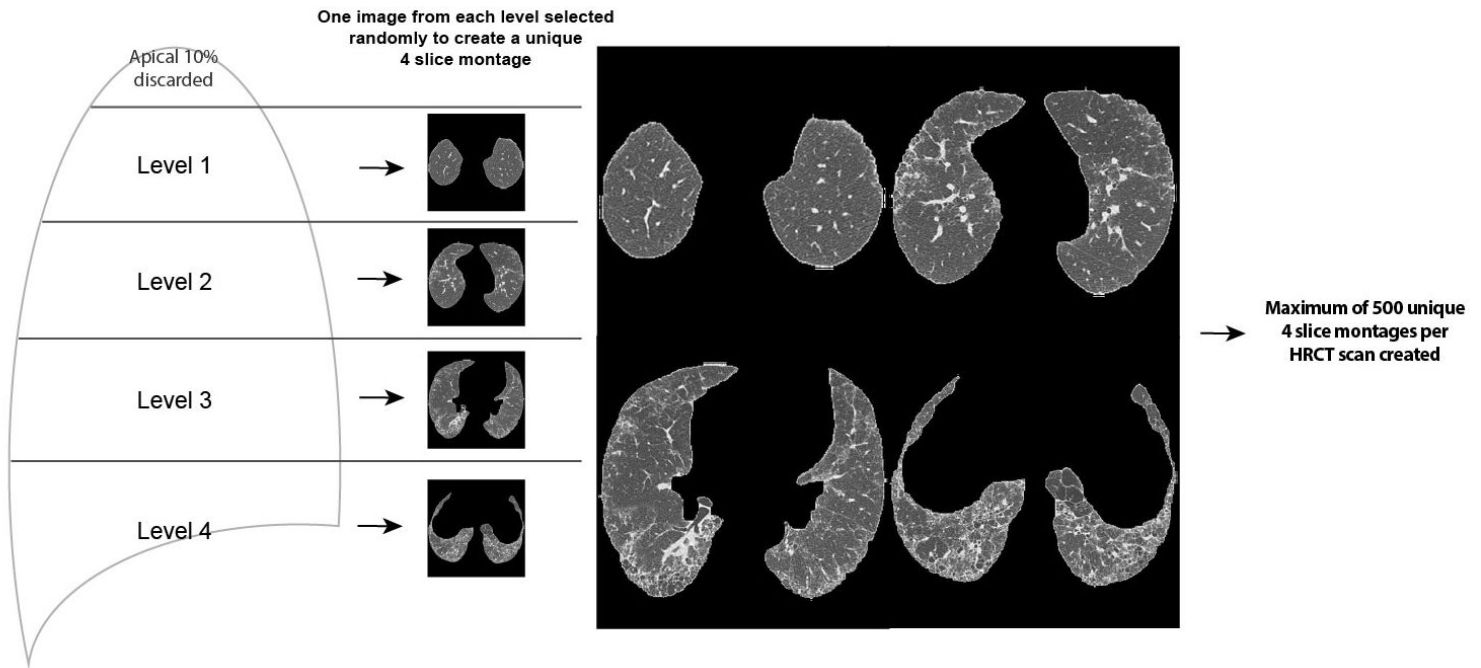


Figure 2. For each HRCT, the lungs are segmented, and four axial slice montages are created by randomly selected a slice from each lung quarter length (excluding the apical 10%). The resampling procedure is designed to ensure that all montages were unique. A maximum of 500 montages were created for each HRCT.

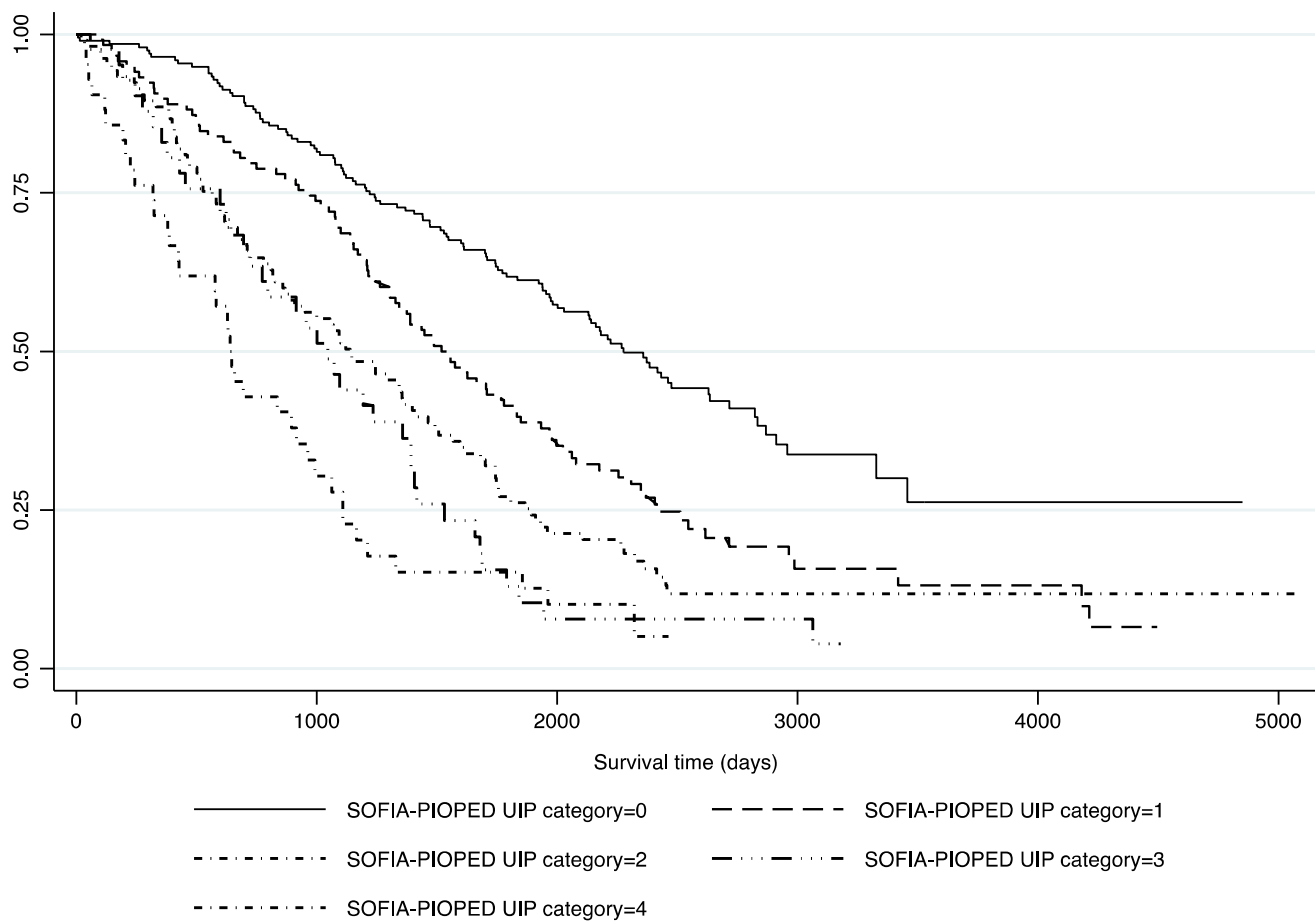


Figure 3 Kaplan-Meier of survival differences between patients assigned to SOFIA-PIOPED UIP categories

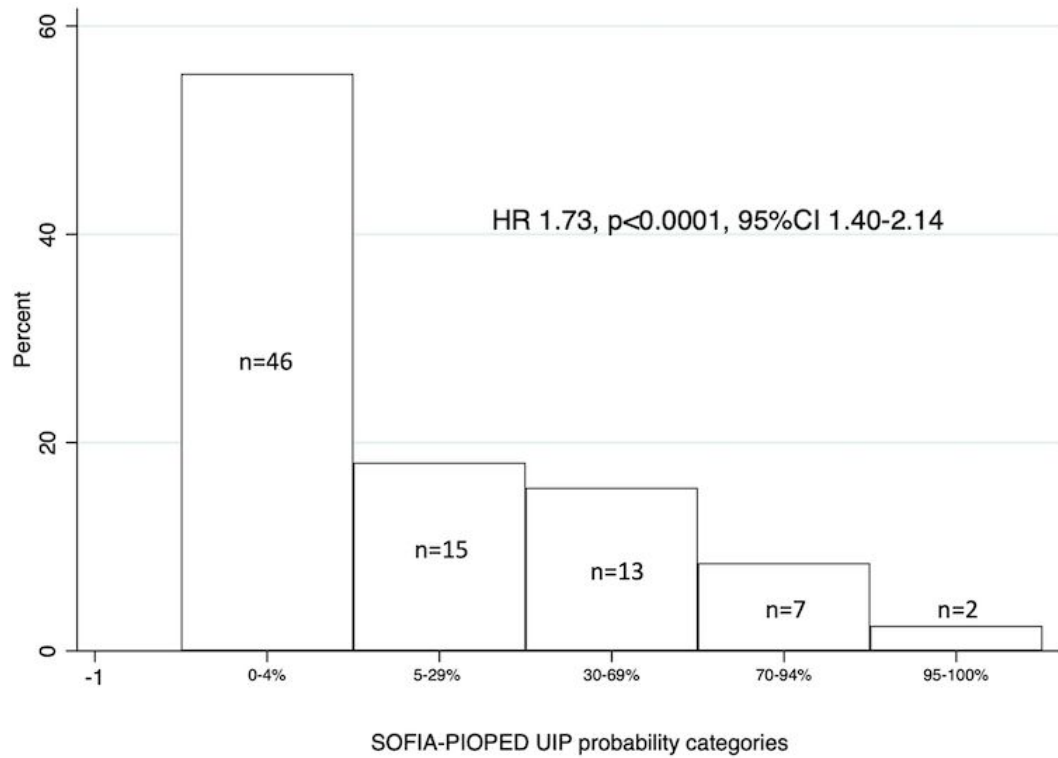


Figure 4. Histogram showing frequency of SOFIA-PIOPED probability categories in patients with indeterminate HRCT appearances based on expert thoracic radiologist consensus.

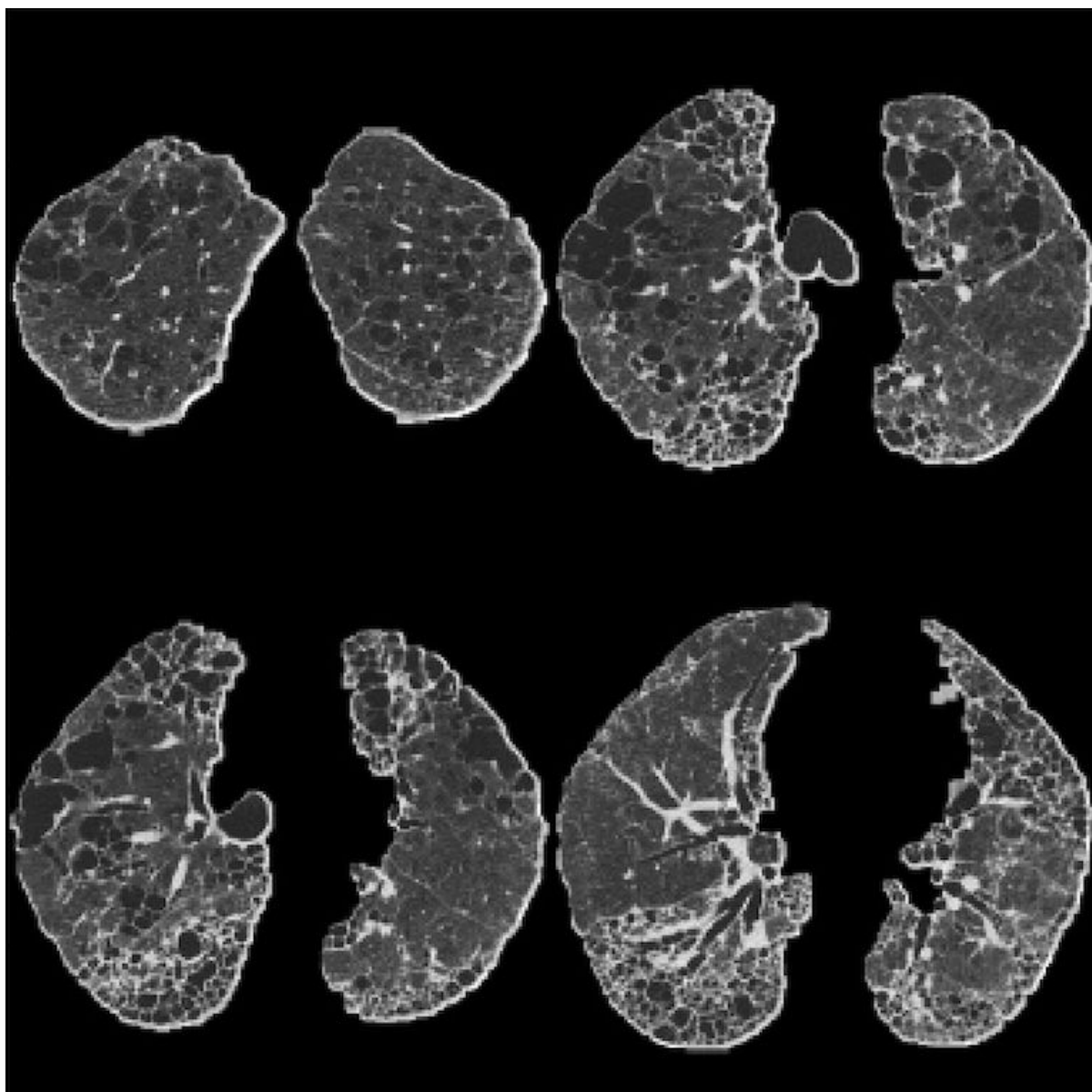


Figure 5a. A single four-slice HRCT montage taken from a patient with a UIP pattern. SOFIA-UIP probabilities for this case were UIP:0.999559, probable UIP:0.000107, indeterminate for UIP: 0.000025, alternative diagnosis: 0.000310.

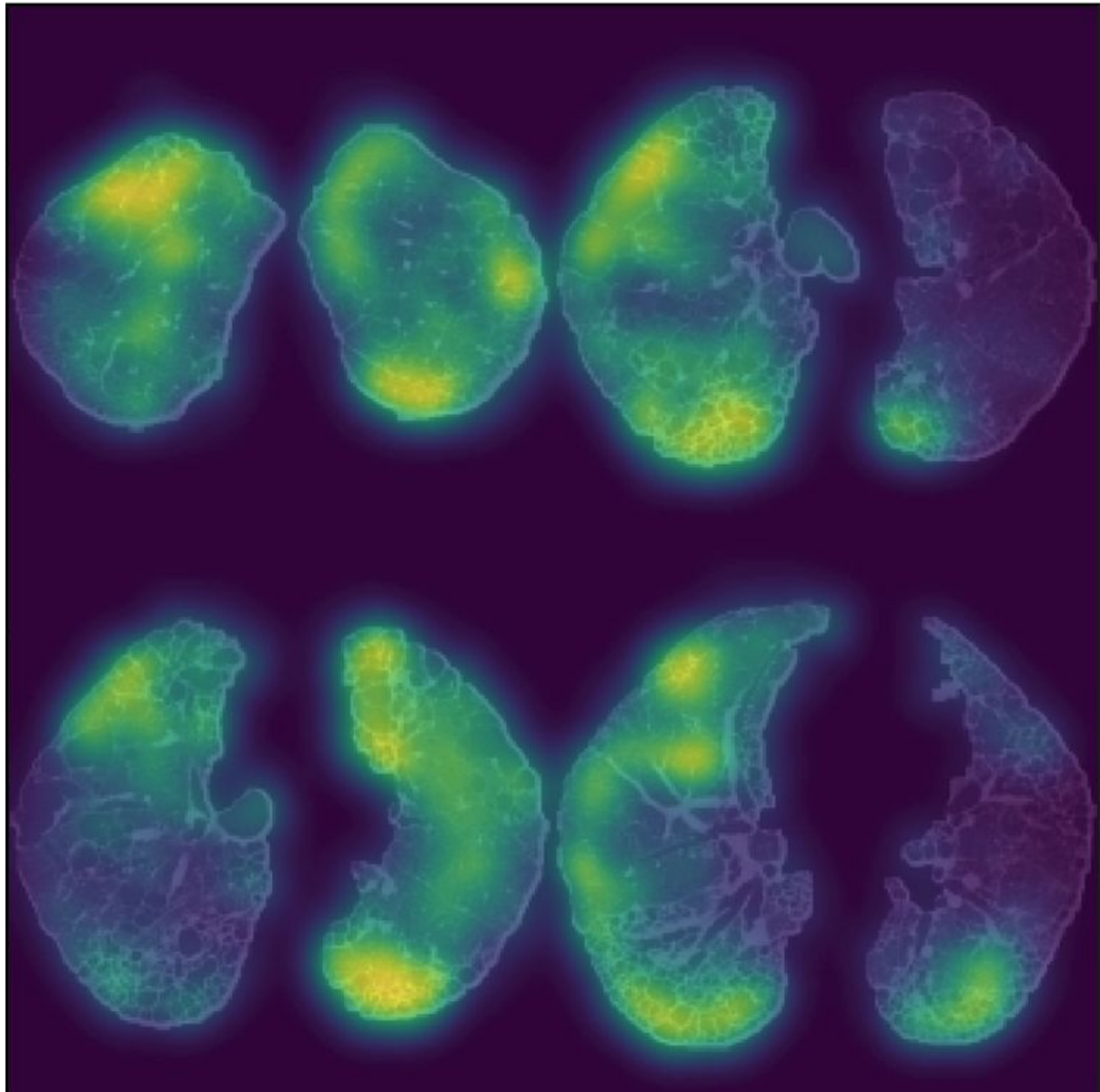


Figure 5b. Saliency map generated by SOFIA highlighting pixels within figure 4a leading to a diagnosis of UIP. The map demonstrates that regions of peripheral honeycombing in (depicted as hotspots) contributed most to the algorithm's diagnosis. A Gaussian smoothing filter was applied to reduce image noise

Online Data Supplement

Semiquantitative HRCT evaluation

Each HRCT scan was scored independently by two thoracic radiologists (LC, MS, 10- and 12-years' experience) who were blinded to all clinical information. HRCTs were scored on a lobar basis. The total extent of interstitial lung disease (ILD) was initially estimated to the nearest 5%, then subclassified into four patterns: ground glass opacification, reticulation, honeycombing, consolidation, and emphysema, using definitions from the Fleischner Society glossary of terms for thoracic imaging. Parenchymal pattern scores for each lobe were generated by multiplying the total lobar ILD extent by the individual lobar parenchymal pattern extents and divided by 100. The individual lobar percentages of each parenchymal pattern were summed for each radiologist and a total extent score for each pattern, for each HRCT. Traction bronchiectasis, as defined in the Fleischner Society glossary of terms, was assigned a severity score (none:0, mild:1, moderate:2, severe:3) for each lobe and these scores were summed to give a total traction bronchiectasis severity score for each HRCT, for each radiologist. Average total ILD extents scores, total parenchymal pattern scores and total traction bronchiectasis severity scores were generated for each HRCT from the individual radiologists' scores.

Each radiologist provided a 0-100% probability score for each of the four ATS/ERS/JRS/LATS 2018 guideline categories (definite UIP, probable UIP, indeterminate for UIP, alternative diagnosis), summing to 100% e.g., UIP:75%, probable UIP 25%, indeterminate for UIP: 0%, alternative diagnosis:0%. Average diagnosis category probabilities for each HRCT were generated from the individual radiologists' scores. The final first-choice diagnosis for each HRCT was taken as the

diagnosis category with the highest probability. Consensus was reached for cases where the probability of two diagnosis categories were equal e.g., UIP:50%, probable UIP 50%, indeterminate for UIP: 0%, alternative diagnosis:0%.

Radiologic diagnosis category	Mean probability (SD)
UIP (n=164)	0.72 ± 0.21
Probable UIP (n=214)	0.63 ± 0.13
Indeterminate (n=55)	0.57 ± 0.17
Alternative diagnosis (n=71)	0.63 ± 0.20

Table A1. Mean probability of first-choice diagnosis based on SOFIA probability scores

	HR	P Value	CI 95%
SOFIA PIOPED UIP probability categories	1.52	<0.0001	1.38-1.67
%Predicted FVC (n=356) *	0.08	<0.0001	0.04-0.16
SOFIA PIOPED UIP probability categories	1.33	<0.0001	1.20-1.48
%Predicted DLco (n=313) *	0.02	<0.0001	0.01-0.05
SOFIA PIOPED UIP probability categories	1.32	<0.0001	1.19-1.47
CPI (n=309) *	1.06	<0.0001	1.05-1.07

Table A2. Cox proportionate hazards models adjusting for disease severity based on lung function. *Limited to patient's lung function performed within 6 months of HRCT in either direction.

Variable	C-index	HR	P Value	CI 95%
Model 1				
GAP stage	0.71	1.57	<0.0001	1.24-1.98
CPI		1.05	<0.0001	1.04-1.06
Model 2				
SOFIA PIOPED UIP probability categories	0.64	1.48	<0.0001	1.37-1.60
Model 3				
GAP stage	0.73	1.40	0.005	1.11-1.77
CPI		1.05	<0.0001	1.04-1.06
SOFIA PIOPED UIP probability categories		1.29	<0.0001	1.17-1.41

Table A3. Cox proportionate hazards models with C-indices for models including key disease severity variables and SOFIA PIOPED UIP probability categories.

Variable	HR	P Value	CI 95%
<i>UIP-like fibrotic patterns</i>			
SOFIA PIOPED UIP probability categories	1.47	<0.0001	1.33-1.63
%Predicted FVC (n=317/338*)	0.11	<0.0001	0.06-0.22
SOFIA PIOPED UIP probability categories	1.28	<0.0001	1.14-1.43
%Predicted DLco (n=280/338*)	0.02	<0.0001	0.01-0.05
SOFIA PIOPED UIP probability categories	1.30	<0.0001	1.17-1.45
CPI (n=278/338*)	1.05	<0.0001	1.03-1.06
<i>Other fibrotic patterns</i>			
SOFIA PIOPED UIP probability categories	1.56	<0.0001	1.31-1.87
%Predicted FVC (n=162/177*)	0.12	<0.0001	0.04-0.36
SOFIA PIOPED UIP probability categories	1.35	0.003	1.11-1.63
%Predicted DLco (n=135/177*)	0.02	<0.0001	0.01-0.10
SOFIA PIOPED UIP probability categories	1.36	0.002	1.13-1.65
CPI (n=133/177*)	1.06	<0.0001	1.04-1.08

Table A4. Cox proportionate hazards models adjusting for disease severity based on lung function in INBUILD subgroups. *Limited to patient's where lung function was performed within 6 months of HRCT in either direction.