EUROPEAN RESPIRATORY journal
OFFICIAL SCIENTIFIC JOURNAL OF THE ERS

Mortality prediction in IPF: evaluation of automated CT analysis with conventional severity measures

Journal:	*European Respiratory Journal*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Jacob, Joseph; Royal Brompton and Harefield NHS Foundation Trust, Bartholmai, Brian; Mayo Clinic, Radiology Rajagopalan, Srinivasan Kokosi, Maria; Royal Brompton Hospital, Respiratory Nair, Arjun; Guys and St Thomas NHS Foundation Trust Karwoski, Ronald; Mayo Clinic, Department of Physiology and Biomedical Engineering Walsh, Simon; King's College Hospital Foundation Trust, Wells, Athol; Interstitial Lung Disease Unit, Royal Brompton Hospital Hansell, David; Royal Brompton Hospital, Radiology
Key Words:	idiopathic pulmonary fibrosis, quantitative CT of ILD, pulmonary vessels

Mortality prediction in IPF: evaluation of automated CT analysis with conventional severity measures

Joseph Jacob[1], Brian J. Bartholmai[2], Srinivasan Rajagopalan[3], Maria Kokosi[4], Arjun Nair[1], Ronald Karwoski[3], Simon L.F. Walsh[1], Athol U. Wells[4], David M. Hansell[1].

[1]Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK.

[2] Division of Radiology, Mayo Clinic Rochester, Rochester, Minnesota, USA.

[3]Department of Physiology and Biomedical Engineering, Mayo Clinic Rochester, Rochester, Minnesota, USA.

[4]Interstitial Lung Disease Unit, Royal Brompton Hospital, Royal Brompton and Harefield NHS Foundation Trust, London, UK.

Corresponding author: Joseph Jacob, 10 Wolsey road, Northwood, Middlesex, UK HA6 2HW.

Email address: akelajacob@gmail.com

TAKE HOME MESSAGE:

CALIPER-derived parameters such as pulmonary vessel volume, are more accurate prognostically than visual CT scores.

Authors contributions

JJ, MK, AN, SLFW, AUW, DMH were involved in either the acquisition, or analysis or interpretation of data for the study.

JJ, AUW and DMH were also involved in the conception and design of the study.

BJB, RK and SR invented and developed CALIPER. They were involved in processing the raw CT scans and in generation of figures but were not involved with the analysis or interpretation of the data in the study.

All authors revised the work for important intellectual content and gave final approval for the version to be published. All authors agree to be accountable for the all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics committee approval

Approval for this study of clinically indicated CT and pulmonary function data was obtained from the Institutional Ethics Committee of the Royal Brompton Hospital and Mayo Clinic and informed patient consent was not required.

ABSTRACT

Computer-based CT analysis can provide objective quantitation of disease in idiopathic pulmonary fibrosis (IPF). A computer algorithm, CALIPER, was compared with conventional CT and pulmonary function measures of disease severity for mortality prediction.

283 consecutive patients with an IPF multidisciplinary diagnosis had variables evaluated against mortality: visual and CALIPER features (total interstitial lung disease extent, honeycombing, reticular pattern, ground glass opacities, pulmonary vessel volume, emphysema, traction bronchiectasis and consolidation) and pulmonary function tests (FEV1, FVC, DLco, Kco and a composite physiologic index). A combination of mortality predictors was compared with the gender, age, physiology model.

On univariate analyses, all visual and CALIPER-derived interstitial features and functional indices were predictive of mortality to a 0·01 level of significance. On multivariate analysis, visual CT parameters were discarded. Independent predictors of mortality were: composite physiologic index (HR=1·05, CI=1·02-1·07, p<0·001) and two CALIPER parameters: pulmonary vessel volume (HR=1·23, CI=1·08-1·40, p=0·001) and honeycombing (HR=1·18, CI=1·06-1·32, p=0·002). A three group staging system derived from this model was powerfully predictive of mortality (HR=2.23, CI=1.85-2.69, p<0.0001).

CALIPER-derived parameters, in particular the pulmonary vessel volume, are more accurate prognostically than traditional visual CT scores. Quantitative tools such as CALIPER have the potential to improve staging systems in IPF.

KEY WORDS: quantitative CT of ILD, pulmonary vessels, idiopathic pulmonary fibrosis, CALIPER

INTRODUCTION:

Accurate prognostication is central to the management of patients with idiopathic pulmonary fibrosis (IPF). In addition to informing a patient of their probable life expectancy,[1] accurate prediction of a patients likely clinical course allows institution of appropriate management which today may include anti-fibrotic medication, referral for transplantation,[2] or palliative care pathways.

However prognostication in IPF is fraught with difficulties. An array of prognostic indicators have been used over the years, but have met with varying degrees of success in IPF. Pulmonary function tests such as DLco are perhaps the most sensitive markers of disease severity but are associated with a measurement variation of between 10-15% per test. [3] Visual CT evaluation has to contend with interobserver variation. [4, 5] Composite indices have also been proposed but are yet to be fully validated. [6] As a result, new tools that may be more accurate in predicting a patient's prognosis are required. A recent recommendation from the Fleischner society,[7] has emphasized computer-based quantitative CT analysis as a potential outcome measure in IPF.

Establishing whether a quantitative tool has the ability to substitute as a marker of disease outcome requires evaluation of the tool against other markers of baseline disease severity. A sophisticated quantitative CT algorithm (CALIPER) has been shown to have better correlations with pulmonary function tests than semi-quantitative visual CT evaluation. [8]

The aims of our study were to compare CALIPER, visual CT scoring and pulmonary function tests (PFTs) against survival in IPF. Optimal stratification was explored using CALIPER variables, visual variables, PFTs and the GAP score.

METHODS

A retrospective analysis of an interstitial lung disease database identified all consecutive, newly attending patients receiving a multidisciplinary team diagnosis of IPF according to published guidelines,[9] over a four and a half year period (January 2007 to July 2011). Patients with a departmental, non-contrast, supine, volumetric CT were included in the study cohort (as per CONSORT in Figure 1). CT, echocardiography and pulmonary function test protocols are included in the online appendix as are details of CALIPER CT evaluation. The DICOM images for the CT scans were transferred to the Biomedical Imaging Resource (BIR), Mayo Clinic, Rochester, USA for blinded CALIPER processing. Approval for this study of clinically indicated CT and pulmonary function data was obtained from the Institutional Ethics Committee of the Royal Brompton Hospital and the Institutional Review Board of the Mayo Clinic.

Visual CT evaluation:

Each CT scan was evaluated independently by two radiologists (AN, SLW) with 5 and 7 years thoracic imaging experience respectively, blinded to all clinical information. An initial training dataset of 15 non-study cases was used to help to identify pre-existing biases. The scores of the test cases were reviewed and the most widely discrepant results discussed with a third radiologist (JJ).

CTs were scored on a lobar basis using a continuous scale. The total interstitial lung disease (ILD) extent was initially estimated to the nearest 5%, then sub-classified into four patterns: reticular pattern, ground glass opacification, honeycombing and

consolidation, using definitions from the Fleischner Society glossary of terms for thoracic imaging. [10] To derive a lobar percentage for each parenchymal pattern, the total lobar ILD extent was multiplied by individual lobar parenchymal pattern extents and divided by 100. Furthermore, the percentage (to the nearest 5%) of each lobe that contained mosaicism (decreased attenuation component) or emphysema was recorded. The individual lobar percentages of each parenchymal pattern were summed for each radiologist and divided by six to create an averaged lobar score per pattern, per scorer per case.

Traction bronchiectasis, as defined in the Fleischner society glossary of terms,[10] was assigned with a categorical "severity" score that took into account the average degree of airway dilatation within areas of fibrosis as well as the extent of dilatation throughout the lobe and was given a gestalt score of: none=0, mild=1, moderate=2, severe=3. An index of pulmonary hypertension (main pulmonary artery:ascending aorta ratio) was assessed by a single scorer using electronic caliper diameter measurements of the ascending aorta and pulmonary artery diameters at the level of the pulmonary artery bifurcation. [11] Consensus formulation for visual scores is outlined in the online appendix.

CALIPER CT evaluation:

Data processing: Initial data processing steps involved extraction of the lung from the surrounding thoracic structures and segmentation into upper, middle and lower zones. Lung segmentation was performed with an adaptive density-based morphological approach,[12] whilst airway segmentation involved iterative three-

dimensional region growing, density thresholding (thresholds including -950HU and -960HU) and connected components analysis. Segmentation of pulmonary vessels, prior to their extraction, was achieved using an optimized multi-scale tubular structure enhancement filter based on the eigenvalues of the Hessian matrix. The filters calculated the 2nd-order derivatives that occurred in the regions that surrounded each pulmonary voxel. The eigenvalues of the Hessian matrix that were constructed from the derivatives were then analyzed, and from these values, it was possible to determine the likelihood that an underlying voxel was connected to a dense tubular structure and therefore represented a vessel.[13, 14] The pulmonary vessel volume (PVV) score quantified the volumes of pulmonary arteries and veins excluding vessels at the lung hilum as a percentage of lung volume (Figure 2).

Parenchymal tissue type classification was applied to 15x15x15 voxel volume units using texture analysis, computer vision-based image understanding of volumetric histogram signature mapping features and 3D morphology.[13] The CALIPER tool was trained by sub-specialty thoracic radiologist consensus assessment of pathologically confirmed datasets.[13, 15]

Pattern evaluation: CALIPER evaluation of CT data involved algorithmic identification and volumetric quantification of every voxel volume unit into one of eight radiological parenchymal features: normal lung, three grades of decreased lung attenuation (grade 1=mild, 2=moderate 3=marked), ground glass opacification, reticular pattern, honeycombing and the pulmonary vessels (Figure 2). Volumes for all eight parenchymal features were converted into a percentage using the total lung

volume also measured by CALIPER. Total ILD extent represented the sum of ground glass, reticular and honeycomb percentages. The sum of grade 2 and 3 decreased lung attenuation represented emphysema. [8]

Statistical analysis:

Data are given as medians, means with standard deviations, or numbers of patients with percentages where appropriate. Interobserver variation for visual scores was assessed using the single determination standard deviation. Linear regression analyses were performed to explore relationships between the PVV and both CALIPER ILD extent and RVSP. Univariate and multivariate Cox regression analyses were used to investigate relationships within and between the three data sets: CALIPER CT evaluation, visual CT evaluation and pulmonary function tests. In all study analyses, a p-value of <0.01 was considered significant.

The hazard ratios for those parameters that were independent predictors of mortality on multivariate Cox regression analysis were used to generate a formula that represented an estimate of mortality for each patient. The hazard ratios for CALIPER parameters alone that were independent predictors of mortality on multivariate Cox regression analysis were also used to generate a separate formula that represented an estimate of mortality for each patient.

The mortality estimates derived from the hazard ratio scores were converted into categorical scores and compared to mortality estimates derived from the Gender, Age, Physiology (GAP) index staging system,[16] using univariate and bivariate Cox

regression analyses, Kaplan Meier survival plots and the log-rank test. Robustness of results were confirmed using bootstrapping and resampling of the dataset up to 1000 times. Goodness of fit of the survival models was calculated using Harrells concordance index. Assumptions of linearity and proportional hazards were tested by visual inspection of Martingale residuals and scaled Schoenfeld residuals. Statistical analyses were performed with STATA (version 12, StatCorp, College Station, TX, USA).

RESULTS

Demographic data

The study population consisted of 283 consecutive patients with a multidisciplinary diagnosis of IPF. The median age at presentation was 67 years, the mean follow up time was 30 months (±21·5 months) and 210 patients (74%) died during the study period. Data on vital status was completed on 98.6% of cases with 4 patients censored (1.4%). Demographic data and average visual score, CALIPER score and pulmonary function test data are provided in Table 1. Interobserver variation values for the visual scores are provided in the supplementary appendix.

Mortality Analyses

All visual and CALIPER-derived interstitial features were predictive of mortality. The CALIPER-derived PVV (shown in Figure 2) was highly significant on univariate analysis as were all pulmonary function tests, the right ventricular systolic pressure (RVSP) and the GAP score (Table 2). When RVSP and CALIPER ILD extent were each placed alongside PVV in a bivariate mortality analysis, only PVV remained independently predictive of mortality. On linear regression analysis, major co-linearity was demonstrated between PVV and CALIPER ILD extent (R^2=0.76, p<0.0001). [8] A lesser degree of co-linearity was demonstrated between PVV and RVSP (R^2=0.20, p<0.0001).

On stepwise proportional hazards analysis, the independent CALIPER-derived predictors of mortality were: honeycombing and the PVV (Table 3). On multivariate analysis of pulmonary function indices, the CPI was the strongest predictor of

mortality. A multivariate Cox regression analysis combining CALIPER scores, visual scores, pulmonary function tests and the GAP score demonstrated the CPI (which has been previously validated by CALIPER [8]) to be the variable that best quantified the severity of ILD. The two remaining CT variables independently predictive of mortality were derived by CALIPER: honeycombing and PVV. No visually scored CT parameters were independently associated with mortality (Table 3). When the GAP score was substituted for the CPI in the final multivariate model, though it was retained as an independent predictor of mortality, it was not as strong as the CPI in the model (Table 3).

Comparisons of composite variables

A further mortality evaluation compared the GAP index staging system [16] to two mortality estimates derived from the hazard ratios of two multivariate models. One formula was derived from the hazard ratios of the final three independent predictors of mortality (CALIPER-CPI score) as follows:

CALIPER-CPI score = (CALIPER PVV x 23·0904) + (CALIPER honeycombing x 18·3795) + (CPI x 4·5065)

The second formula was derived from the hazard ratios of the two variables that on multivariate analysis of CALIPER scores were independent predictors of mortality (CALIPER-only score):

CALIPER-only score = (CALIPER PVV x 52·9004) + (CALIPER honeycombing x 12·0524)

Both the CALIPER-CPI and CALIPER-only scores were converted into categorical scores by aligning the individual scores in ascending numerical order and dividing the respective cohorts into three equally sized groups (n=83). Univariate Cox regression analyses demonstrated that the CALIPER-CPI and CALIPER-only categories were not only of similar prognostic strength to the GAP index staging system, but demonstrated improved goodness of fit as models (Table 4).

On bivariate Cox regression analysis, the GAP index staging system did not retain significance against the CALIPER-CPI categories, a result confirmed with bootstrapping of 1000 samples (Table 4). The GAP index staging system was also shown to be a weaker predictor of mortality than the CALIPER-only categories, again confirmed with bootstrapping of 1000 samples. Kaplan Meier survival curves demonstrated similar separation of the groups using the CALIPER-CPI and CALIPER-only categories when compared to the GAP index staging system (log-rank test $p<0.0001$ for the GAP index staging system, CALIPER-CPI and CALIPER-only categories) (Figure 3).

When the CALIPER-CPI and CALIPER-only scores were adjusted such that patient numbers in each of the groups were identical to the patient numbers in the GAP staging system groups, the relationships on univariate analysis and bivariate analysis with bootstrapping did not change. Goodness of fit of the two models was similar to that of the GAP index staging system (Harrells C index=0.66 for the CALIPER-CPI score and 0.65 for the CALIPER-only score). Kaplan Meier survival curves again demonstrated similar separation of the groups using the new adjusted CALIPER-CPI

and CALIPER-only categories when compared to the GAP index staging system (log-rank test $p<0.0001$ for the adjusted CALIPER-CPI and CALIPER-only categories) (Figure 3).

DISCUSSION

Our study has demonstrated that computer-derived quantitative CT parameters are better predictors of mortality in IPF than visually scored parameters. When patients are stratified on the basis of CALIPER variables and the CPI, mortality prediction is improved when compared to stratification using the GAP index staging system. Central to the strength of mortality prediction using CALIPER was a single variable, the pulmonary vessel volume (PVV). The PVV is a novel CALIPER variable with no visually scored equivalent. Accordingly, the PVV may represent a new parameter in the evaluation of patients with IPF.

DLco has long been considered the parameter that best reflects disease severity at baseline in IPF,[17] but is handicapped by measurement noise ranging between 5-15%.[3, 18] Consequently, interest is increasing in exploring other potential markers of disease severity or worsening in IPF such as peripheral blood[19] and imaging biomarkers.[7] Given the rapid technological advances of computer based quantitative CT,[13] an exploration of their potential accuracy in assessing IPF is timely.

In the current study, on univariate analysis, visually scored reticular pattern, honeycombing and traction bronchiectasis were predictive of mortality confirming the conclusions of a previous study which evaluated the prognostic significance of the same patterns in patients with a histopathologically proven UIP pattern.[20]

The final combined visual, CALIPER and CPI multivariate model identified three independent predictors of survival. One of these was the CPI, and the finding

confirms previous reports highlighting the strength of signal of the CPI in predicting outcome in fibrosing lung disease.[21, 22] Similarly the prognostic signal associated with honeycombing when scored by CALIPER confirms results from visual scoring of the same pattern in IPF. [23]

The co-linearity demonstrated between PVV and CALIPER ILD extent and between PVV and RVSP suggests that the PVV might represent a variable that simultaneously captures disease within the interstitial and vascular compartments. It is noteworthy that CALIPER software was originally designed to segment out and discard vascular structures (rather than quantify them) and in so doing optimize the classification of parenchymal patterns. In cases with more severe fibrosis on CT, vessel segmentation inevitably captures a minor degree of peripheral reticular pattern (Figure 2). There is the potential to refine vessel delineation by CALIPER, or a similar quantitative tool, and thus improve the prognostic signal of this new CT parameter.

The basis of the signal provided by pulmonary vessel volume is obscure. PVV has not previously been studied as a prognostic marker in IPF or considered as a prognostic indicator independent of pulmonary hypertension. Recent IPF studies assessing the complex interactions between angiogenic and angiostatic mediators in disease pathogenesis, [24] have primarily considered the vasculature through the prism of overt pulmonary hypertension.[25]

Blood perfusion in areas of fibrosis has been shown to be reduced,[26] but conversely increased in spared lung adjacent to areas of fibrosis.[27, 28] It follows that the strong

correlations between ILD extent and vessels may reflect regional, subclinically elevated local pulmonary arterial pressures within mildly fibrotic lung, or capillary bed destruction in more advanced disease, which may produce a preferential diversion of blood flow to relatively spared or non-fibrotic lung. The vascular capacitance of spared lung (the upper and middle lobes in patients with IPF, a predominantly basal disease) may result in an increase in vessel volume in more advanced disease. The identification of greater numbers of vessels, of a size that could be detected by CALIPER, could therefore act as a surrogate marker for the extent and severity of parenchymal disease in IPF.

Another possible explanation for the relationship between PVV and ILD extent relates to the increased negative intra-thoracic pressure that non-compliant fibrotic lungs need to generate during inspiration. The transmission of high negative pressures through the pleural space into the parenchyma could in turn be exerted on the vasculature, resulting in dilatation throughout the lung and an increase in capacitance. A third possible mechanism relates to pleuroparenchymal and/or bronchial-pulmonary artery anastomoses described histopathologically in patients with fibrosing lung disease.[29] While the clinical importance of shunting within the lungs is yet to be established,[30] the development of shunts could theoretically increase the PVV as fibrosis progresses.

The evaluation of CALIPER in our previous study highlighted the close correlations between parenchymal patterns scored using an automated computer system and pulmonary function indices. [8] The results suggested that a tool such as CALIPER,

when used alone, could be a viable alternative to pulmonary function testing in predicting outcome in IPF. Our mortality analyses have also underlined the importance of integrating structural and functional parameters for prognostication in IPF. The CPI when combined with CALIPER variables produced a stronger model with which to predict mortality than was achieved by CALIPER or functional indices alone.

The GAP score is a multidimensional continuous score that aims to predict mortality in IPF by utilizing commonly measured clinical and physiological variables.[16] The GAP score was shown to be a strong univariate predictor of mortality in our study, and was enhanced when combined with CALIPER variables in a multivariate model. However, the model was not as powerful as the combination of CALIPER variables with CPI. The findings highlight a relative weakness of the GAP score consequent upon its relatively coarse nine-point gradation, whereas the continuous nature of the CPI allows the CPI to be more discriminatory.

The GAP staging system represents a categorical version of the GAP score.[16] A consequence of stratification according to the GAP staging system is that only 50% of patients in the current study constituted GAP groups 1 and 3 (patients with mild and severe disease respectively). Yet it is precisely such patients that require identification in IPF cohorts. Those identified as likely to have limited disease can be monitored using a watch and wait policy, whilst those with more severe disease might be referred earlier for transplantation.[2] When stratification with the CALIPER-CPI categories was performed, despite the arbitrary division of the patients into

evenly sized groups, prognostication and goodness of fit was improved when compared to the categorical GAP staging system.

In conclusion we have shown that quantitative computer-derived CT variables in IPF are superior predictors of mortality than any visually scored CT parameter. Stratification using CALIPER variables and CPI provides a stronger mortality signal than stratification using the GAP index. But one CALIPER variable in particular, the pulmonary vessel volume, has the strongest link with mortality and could be a new index in the evaluation of IPF.

Declaration of Interests

BJB, RK, SR report a grant from the Royal Brompton Hospital during the conduct of the study; another from Imbio, LLC, was outside the submitted work; and all have a patent: SYSTEMS AND METHODS FOR ANALYZING IN VIVO TISSUE VOLUMES USING MEDICAL IMAGING DATA licensed to Imbio, LLC.

AUW receives personal fees for participating in advisory boards and speaking at symposia from Boehringer Ingleheim, Intermune, Roche and Bayer, and for participating in advisory boards from Gilead, MSD and speaker fees from Chiesi.

SLW reports personal fees from Boehringer Ingleheim, outside the submitted work.

DMH has received a grant from Intermune for creating an educational website and consultancy and receives personal consultancy fees from AstraZeneca, Boehringer Ingleheim, Intermune, Roche, Sanofi, Glaxo Smith Klein.

Funding source

There is no funding source for the current study. Joseph Jacob had full access to all the data in the study and had final responsibility for the decision to submit for publication.

REFERENCES:

1. Schoenheit G, Becattelli I, Cohen AH. Living with idiopathic pulmonary fibrosis: an in-depth qualitative survey of European patients. *Chron Respir Dis*. 2011 November 1, 2011;**8**(4):225-31.

2. Lamas DJ, Kawut SM, Bagiella E, Philip N, Arcasoy SM, Lederer DJ. Delayed Access and Survival in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2011 2011/10/01;**184**(7):842-7.

3. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, Coates A, van der Grinten CPM, Gustafsson P, Hankinson J, Jensen R, Johnson DC, MacIntyre N, McKay R, Miller MR, Navajas D, Pedersen OF, Wanger J. Interpretative strategies for lung function tests. *Eur Respir J*. 2005;**26**(5):948-68.

4. Lynch DA, David GJ, Safrin S, Starko KM, Hormel P, Brown KK, Raghu G, King TE, Jr., Bradford WZ, Schwartz DA, Richard WW. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *Am J Respir Crit Care Med*. 2005;**172**:488-93.

5. Watadani T, Sakai F, Johkoh T, Noma S, Akira M, Fujimoto K, Bankier AA, Lee KS, Muller NL, Song JW, Park JS, Lynch DA, Hansell DM, Remy-Jardin M, Franquet T, Sugiyama Y. Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology*. 2013;**266**(3):936-44.

6. Ley B, Collard HR. Risk prediction in idiopathic pulmonary fibrosis. *AmJRespirCrit Care Med*. 2012;**185**(1):6-7.

7. Hansell DM, Goldin JG, King Jr TE, Lynch DA, Richeldi L, Wells AU. CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment

trials: a Position Paper from the Fleischner society. *The Lancet Respiratory Medicine*. 2015.

8. Jacob JB, B.; Rajagopalan, S.; Kokosi, M.; Nair, A.; Karwoski, R.; Walsh, S.; Wells, A.U.; Hansell, D.M. Automated quantitative CT versus visual CT scoring in idiopathic pulmonary fibrosis: validation against pulmoanry function. *J Thorac Imaging*. 2016:In press.

9. Raghu G, Collard HR, Egan JJ, Martinez FJ, Behr J, Brown KK, Colby TV, Cordier JF, Flaherty KR, Lasky JA, Lynch DA, Ryu JH, Swigris JJ, Wells AU, Ancochea J, Bouros D, Carvalho C, Costabel U, Ebina M, Hansell DM, Johkoh T, Kim DS, King TE, Jr., Kondoh Y, Myers J, Muller NL, Nicholson AG, Richeldi L, Selman M, Dudden RF, Griss BS, Protzko SL, Schunemann HJ. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med*. 2011;**183**(6):788-824.

10. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology*. 2008;**246**(3):697-722.

11. Ng CS, Wells AU, Padley SP. A CT sign of chronic pulmonary arterial hypertension: the ratio of main pulmonary artery to aortic diameter. *J Thorac Imaging*. 1999;**14**(4):270-8.

12. Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Trans Med Imaging*. 2001;**20**(6):490-8.

13. Bartholmai BJ, Raghunath S, Karwoski RA, Moua T, Rajagopalan S, Maldonado F, Decker PA, Robb RA. Quantitative CT imaging of interstitial lung diseases. *J Thorac Imaging*. 2013;**28**(5):298-307.

14. Shikata H, McLennan G, Hoffman EA, Sonka M. Segmentation of pulmonary vascular trees from thoracic 3D CT images. *Int J Biomed Imaging*. 2009:11.

15. Maldonado F, Moua T, Rajagopalan S, Karwoski RA, Raghunath S, Decker PA, Hartman TE, Bartholmai BJ, Robb RA, Ryu JH. Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. *Eur Respir J*. 2014;**43**(1):204-12.

16. Ley B, Ryerson CJ, Vittinghoff E, Ryu JH, Tomassetti S, Lee JS, Poletti V, Buccioli M, Elicker BM, Jones KD, King TE, Jr., Collard HR. A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Ann Intern Med*. 2012;**156**(10):684-91.

17. Lynch DA, Godwin JD, Safrin S, Starko KM, Hormel P, Brown KK, Raghu G, King TE, Jr., Bradford WZ, Schwartz DA, Richard WW. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *AmJRespirCrit Care Med*. 2005;**172**(4):488-93.

18. Hathaway EHT, D.P.; Simmons, M.S. Intraindividual variability in serial measurements of DLCO and alveolar volume over one year in eight healthy subjects using three independent measuring systems. *Am Rev Respir Dis*. 1989;**140**:1818-22.

19. Ley B, Brown KK, Collard HR. Molecular biomarkers in idiopathic pulmonary fibrosis. *American Journal of Physiology - Lung Cellular and Molecular Physiology*. 2014;**307**(9):L681-L91.

20. Sumikawa H, Johkoh T, Colby TV, Ichikado K, Suga M, Taniguchi H, Kondoh Y, Ogura T, Arakawa H, Fujimoto K, Inoue A, Mihara N, Honda O, Tomiyama N, Nakamura H, Muller NL. Computed tomography findings in pathological usual interstitial pneumonia: relationship to survival. *Am J Respir Crit Care Med*. 2008;**177**(4):433-9.

21. Wells AU, Desai SR, Rubens MB, Goh NS, Cramer D, Nicholson AG, Colby TV, Du Bois RM, Hansell DM. Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography. *Am J Respir Crit Care Med*. 2003;**167**:962-9.

22. Corte TJ, Wort SJ, Gatzoulis MA, Macdonald P, Hansell DM, Wells AU. Pulmonary vascular resistance predicts early mortality in patients with diffuse fibrotic lung disease and suspected pulmonary hypertension. *Thorax*. 2009;**64**(10):883-8.

23. Flaherty KR, Toews GB, Travis WD, Colby TV, Kazerooni EA, Gross BH, Jain A, Strawderman RL, III, Paine R, Flint A, Lynch JP, III, Martinez FJ. Clinical significance of histological classification of idiopathic interstitial pneumonia. *EurRespir J*. 2002;**19**:275-83.

24. Hanumegowda C, Farkas L, Kolb M. Angiogenesis in pulmonary fibrosis: Too much or not enough? *Chest*. 2012;**142**:200-7.

25. Farkas L, Farkas D, Ask K, Möller A, Gauldie J, Margetts P, Inman M, Kolb M. VEGF ameliorates pulmonary hypertension through inhibition of endothelial apoptosis in experimental lung fibrosis in rats. *The Journal of Clinical Investigation*. 2009;**119**(5):1298-311.

26. Renzoni EA, Walsh DA, Salmon M, Wells AU, Sestini P, Nicholson AG, Veeraraghavan S, Bishop AE, Romanska HM, Pantelidis P, Black CM, du Bois RM. Interstitial Vascularity in Fibrosing Alveolitis. *Am J Respir Crit Care Med*. 2003 2003/02/01;**167**(3):438-43.

27. Ebina M, Shimizukawa M, Shibata N, Kimura Y, Suzuki T, Endo M, Sasano H, Kondo T, Nukiwa T. Heterogeneous Increase in CD34-positive Alveolar Capillaries in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2004;**169**:1203-8.

28. Cosgrove GP, Brown KK, Schiemann WP, Serls AE, Parr JE, Geraci MW, Schwarz MI, Cool CD, Worthen GS. Pigment Epithelium–derived Factor in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 2004;**170**(3):242-51.

29. Turner-Warwick M. Precapillary systemic-pulmonary anastomoses. *Thorax*. 1963;**18**:225-37.

30. Miller WCH, J.G.; Unger, K.M.; Suich, D.M;. Anatomical lung shunting in pulmonary fibrosis. *Thorax*. 1986;**41**:208-9.

Variable (n = 283 unless stated) Units are percentage unless stated	Value
Median Age (years)	67
Male/female	219/64
Never smokers	97 (34%)
Ex-smokers	178 (63%)
Current smokers	8 (3%)
Lung biopsy	60 (21%)
Mean follow up time (months)	30·0 ± 21·5
FEV1 % predicted (n=257)	70·8 ± 19·1
FVC % predicted (n=257)	68·8 ± 20·5
DLco % predicted (n=254)	36·1 ± 12·9
Kco % predicted (n=254)	69·0 ± 19·2
TLC % predicted (n=241)	63·7 ± 15·9
CPI (n=249)	55·1 ± 11·7
CALIPER total ILD extent	26·5 ± 18·1
CALIPER ground glass opacity	17·0 ± 14·7
CALIPER reticular pattern	8·5 ± 6·0
CALIPER honeycombing	1·0 ± 1·7
CALIPER PVV	5·1 ± 1·7
CALIPER Grade 1 decreased attenuation	20·8 ± 20·7
CALIPER Grade 2 decreased attenuation	0·8 ± 2·6
CALIPER Grade 3 decreased attenuation	0·5 ± 2·8
Visual ILD extent	43·1 ± 17·8
Visual ground glass opacity	10·4 ± 11·4
Visual reticular pattern	21·7 ± 10·9
Visual honeycombing	9·8 ± 12·6
Visual consolidation	1·1 ± 3·3
Visual emphysema	4·7 ± 10·9
Visual mosaic attenuation	0·8 ± 2·1
Visual TxBx severity (maximum score 18)	7·0 ± 3·3
Main pulmonary artery diameter (mm)	30·3 ± 4·8
Ascending aorta diameter (mm)	34·8 ± 4·2
Echocardiography RVSP (mmHg) (n=150)	45·1 ± 16·8

Table 1. Patient demographics and mean and standard deviations of pulmonary function tests, CALIPER and visually scored CT parameters and echocardiography data. CPI=composite physiologic index, CT=computed tomography, FEV1=forced expiratory volume in one second, FVC=forced vital capacity, DLco=diffusing capacity for carbon monoxide, Kco=Carbon monoxide transfer coefficient, TLC=total lung capacity, RV/TLC=residual volume expressed as a percent of total lung capacity, ILD=interstitial lung disease, PVV=pulmonary vessel volume, TxBx=traction bronchiectasis, RVSP=right ventricular systolic pressure.

	Patient Number	Hazard ratio	P Value	95.0% Confidence Interval	
				Lower	Upper
CALIPER score					
Total ILD extent	283	1·03	<0·0001	1·03	1·04
Ground glass opacity	283	1·03	<0·0001	1·02	1·04
Reticular pattern	283	1·10	<0·0001	1·07	1·12
Honeycombing	283	1·11	0·006	1·03	1·20
Grade 1 DA	283	0·97	<0·0001	0·97	0·98
Grade 2 DA	283		NS		
Grade 3 DA	283		NS		
Normal lung	283		NS		
PVV	283	1·52	<0·0001	1·40	1·65
VISUAL score					
ILD extent	283	1·03	<0·0001	1·02	1·04
Ground glass opacity	283	1·01	0·03	1·00	1·03
Reticular pattern	283	1·02	0·002	1·01	1·03
Honeycombing	283	1·03	<0·0001	1·03	1·07
Consolidation	283	1·08	<0·0001	1·04	1·12
Total emphysema	283		NS		
Mosaicism	283		NS		
TxBx severity	283	1·11	<0·0001	1·07	1·16
Main PA	283		NS		
PA:Ao ratio	283		NS		
Pulmonary Function Tests					
FEV1	257	0·97	<0·0001	0·97	0·98
FVC	257	0·97	<0·0001	0·96	0·98
TLC	241	0·96	<0·0001	0·95	0·97
DLco	254	0·94	<0·0001	0·93	0·95
Kco	254	0·99	0·001	0·98	1·00
CPI	249	1·07	<0·0001	1·05	1·09
Echocardiography RVSP	150	1·02	<0·0001	1·02	1·03
GAP score	249	1.45	<0·0001	1.30	1.62

Table 2. Univariate Cox regression analysis demonstrating mortality according to CALIPER indices (top white), visually derived HRCT indices (light grey), pulmonary function tests (dark grey) and echocardiography and the GAP score (bottom white). ILD=Interstitial lung disease, DA=decreased attenuation, TxBx=traction bronchiectasis, PA=pulmonary artery, Ao=Aorta, PVV=pulmonary vessel volume, FEV1=forced expiratory volume in one second, FVC=forced vital capacity, TLC=total lung capacity, DLco=diffusing capacity for carbon monoxide,

Kco=Carbon monoxide transfer coefficient, CPI=composite physiologic index, RVSP=right ventricular systolic pressure, NS=not significant.

	Hazard ratio	P Value	95.0% Confidence Interval	
			Lower	Upper
CALIPER analysis				
PVV	1·53	<0·0001	1·41	1·66
Honeycombing	1·12	0·004	1·04	1·21
VISUAL score analysis				
Ground glass opacity	1·03	<0·0001	1·01	1·04
Reticular pattern	1·03	<0·0001	1·01	1·04
Honeycombing	1·04	<0·0001	1·03	1·05
Consolidation	1·10	<0·0001	1·06	1·15
CALIPER PVV	1·23	0·001	1·08	1·40
CALIPER Honeycombing	1·18	0·002	1·06	1·32
CPI	1·05	<0·0001	1·02	1·07
CALIPER PVV	1·39	<0·0001	1·25	1·54
CALIPER Honeycombing	1·21	0·0004	1·09	1·34
GAP score	1.20	0.004	1.06	1.35

Table 3. Multivariate Cox regression analysis demonstrating mortality separately for CALIPER indices (top white), visually derived HRCT indices (light grey) and combined CALIPER, visually derived HRCT indices and pulmonary function tests (dark grey). The GAP score was substituted for CPI in the final combined multivariate model but was a weaker predictor of mortality than the CPI (lower white). PVV=pulmonary vessel volume, CPI=composite physiologic index.

	Hazard ratio	P Value	95% Confidence Interval		Harrell C Index
			Lower	Upper	
GAP index staging system	2.00	<0·0001	1·61	2.48	0.62
CALIPER-CPI	2·23	<0·0001	1·85	2·69	0.67
CALIPER-only	2·14	<0·0001	1·78	2·58	0.67
GAP index staging system	1·25	0·09	0.97	1.62	
CALIPER-CPI	1.99	<0·0001	1·58	2·50	
GAP index staging system[b]		0·09	-0·03	0·48	
CALIPER_CPI[b]		0·001	0·46	0.95	
GAP index staging system	1·41	0·0045	1·11	1·80	
CALIPER-only	1·87	<0·0001	1·51	2·30	
GAP index staging system[b]		0·007	0·10	0·59	
CALIPER-only[b]		0·001	0·41	0·86	

Table 4. Univariate Cox regression analyses (white) comparing the Gender, Age, Physiology (GAP) index staging system with scores derived from two separate hazard ratio formulae. The first formula derived from the hazard ratios of the four independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The second formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). Bivariate analyses (light and dark grey) compared the hazard ratio derived scores to the GAP index staging system with the results confirmed on bootstrapping of 1000 samples. CPI=Composite Physiologic Index, [b]=bootstrapped results.

Figure 1. CONSORT diagram illustrating the selection of patients for the final study population. ILD = interstitial lung disease, CTD = connective tissue disease, IPF = idiopathic pulmonary fibrosis, LCH = Langerhans cell histiocytosis, LAM = Lymphangioleiomyomatosis, CT = computed tomography.

Figure 2. Two examples of delineation of pulmonary arteries and veins (coloured red), quantified by CALIPER as a percentage of the total lung volume and expressed as the pulmonary vessel volume (PVV). Note that segmentation is incomplete – for example a left upper lobe vessel in Figure 2c (arrow). The CALIPER software also misclassifies small volumes of reticulation as vessels (arrowhead).

Case 1 - A 62-year-old female ex-smoker, with a 15-pack year smoking history. Mean visual scores characterized the lungs as consisting of 9% total ILD extent (ground glass opacity=4%; reticular pattern=5%) (A). CALIPER identified a total ILD extent of 2% of the lung (ground glass opacity=1%; reticular pattern=1%). PVV represented 2% of the lung volume (B).

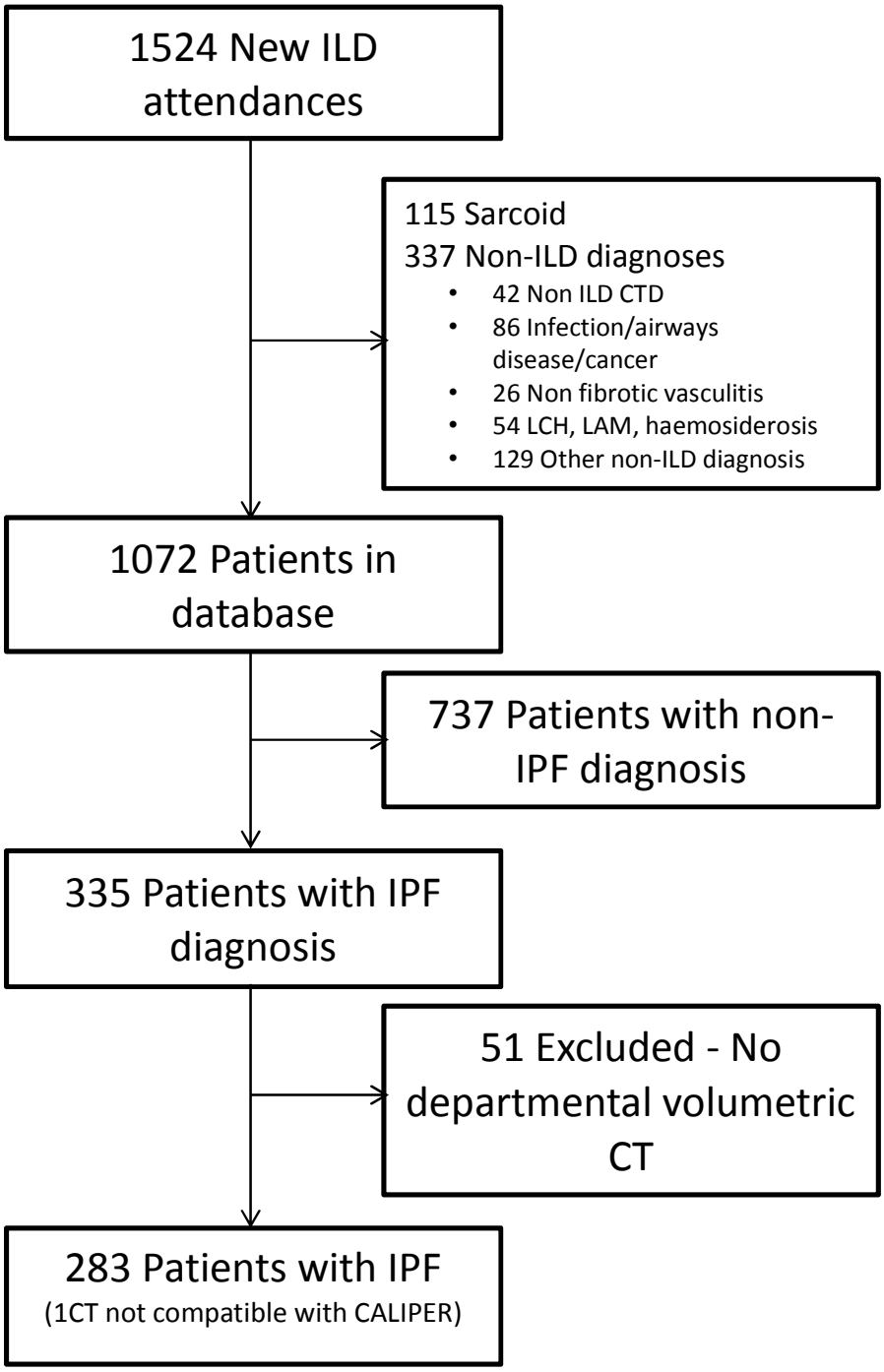
Case 2 - A 61-year-old female ex-smoker, with a 25-pack year smoking history. Visual scores characterized the lungs as consisting of 29% total ILD extent (ground glass opacity=3%; reticular pattern=8%; honeycombing=17%)(C). CALIPER identified a total ILD extent of 27% (ground glass opacity=15%; reticular pattern=11%). PVV represented 6% of the lung volume (D).

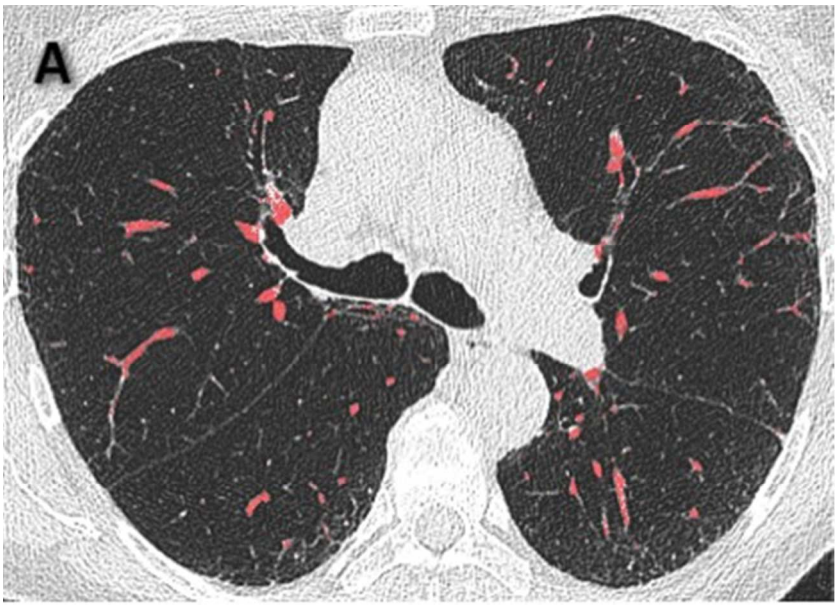
Figure 3. Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender,

Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.

In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.

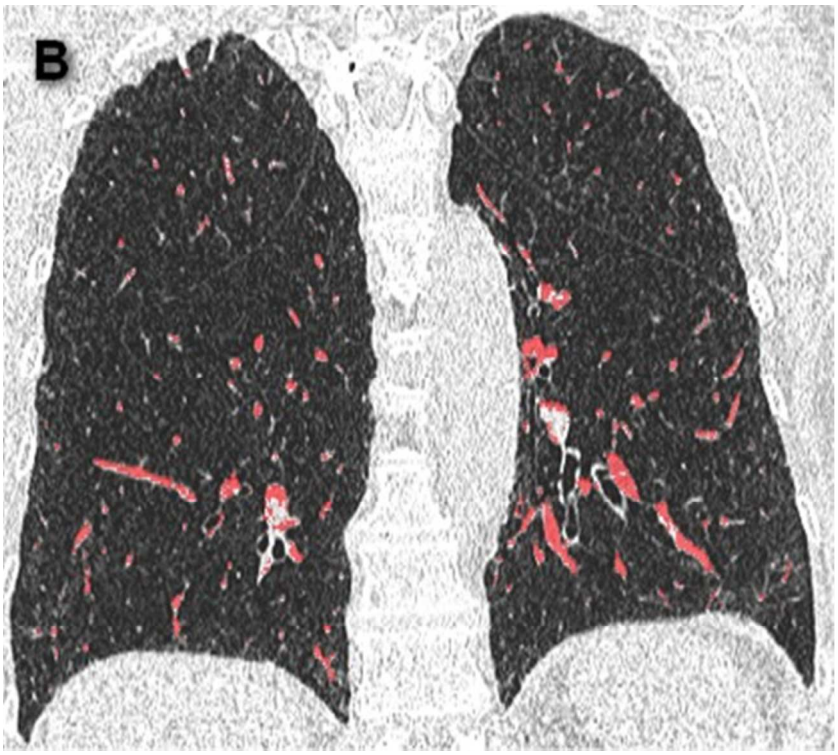
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.


1524 New ILD attendances
115 Sarcoid
337 Non-ILD diagnoses
• 42 Non ILD CTD
• 86 Infection/airways disease/cancer
• 26 Non fibrotic vasculitis
• 54 LCH, LAM, haemosiderosis
• 129 Other non-ILD diagnosis
1072 Patients in database
737 Patients with non-IPF diagnosis
335 Patients with IPF diagnosis
51 Excluded - No departmental volumetric CT
283 Patients with IPF
(1CT not compatible with CALIPER)



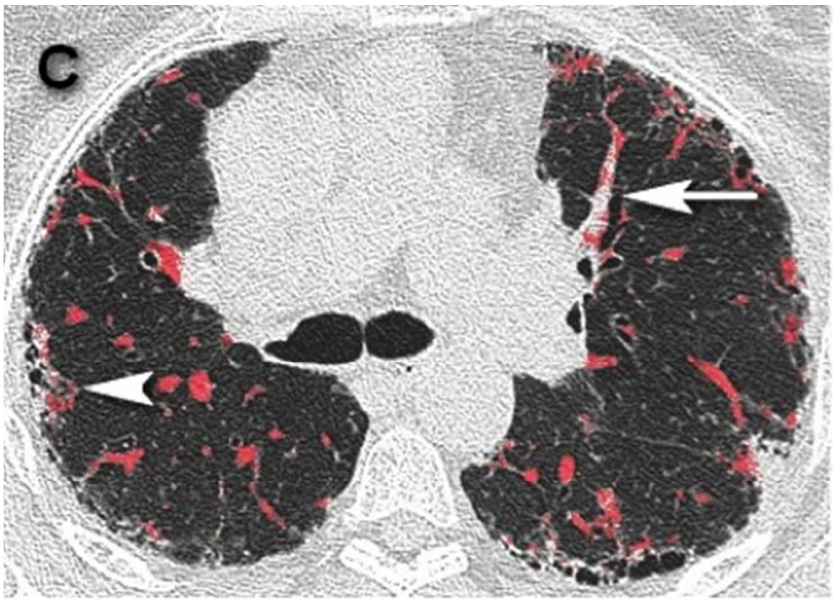
Two examples of delineation of pulmonary arteries and veins (coloured red), quantified by CALIPER as a percentage of the total lung volume and expressed as the pulmonary vessel volume (PVV). Note that segmentation is incomplete – for example a left upper lobe vessel in Figure 2c (arrow). The CALIPER software also misclassifies small volumes of reticulation as vessels (arrowhead).
Case 1 - A 62-year-old female ex-smoker, with a 15-pack year smoking history. Mean visual scores characterized the lungs as consisting of 9% total ILD extent (ground glass opacity=4%; reticular pattern=5%) (A). CALIPER identified a total ILD extent of 2% of the lung (ground glass opacity=1%; reticular pattern=1%). PVV represented 2% of the lung volume (B).
Case 2 - A 61-year-old female ex-smoker, with a 25-pack year smoking history. Visual scores characterized the lungs as consisting of 29% total ILD extent (ground glass opacity=3%; reticular pattern=8%; honeycombing=17%)(C). CALIPER identified a total ILD extent of 27% (ground glass opacity=15%; reticular pattern=11%). PVV represented 6% of the lung volume (D).

146x105mm (72 x 72 DPI)



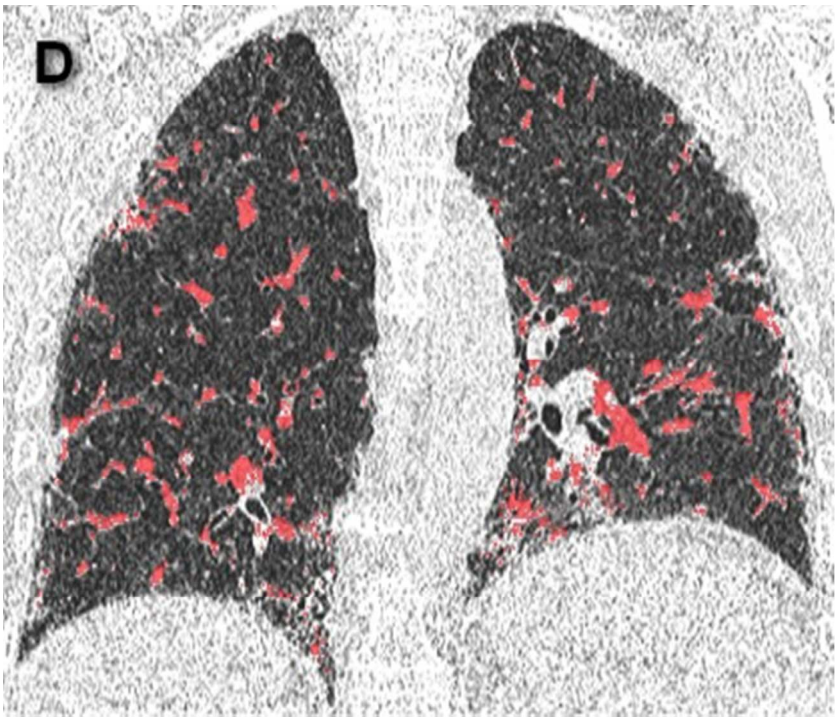
Two examples of delineation of pulmonary arteries and veins (coloured red), quantified by CALIPER as a percentage of the total lung volume and expressed as the pulmonary vessel volume (PVV). Note that segmentation is incomplete – for example a left upper lobe vessel in Figure 2c (arrow). The CALIPER software also misclassifies small volumes of reticulation as vessels (arrowhead).
Case 1 - A 62-year-old female ex-smoker, with a 15-pack year smoking history. Mean visual scores characterized the lungs as consisting of 9% total ILD extent (ground glass opacity=4%; reticular pattern=5%) (A). CALIPER identified a total ILD extent of 2% of the lung (ground glass opacity=1%; reticular pattern=1%). PVV represented 2% of the lung volume (B).
Case 2 - A 61-year-old female ex-smoker, with a 25-pack year smoking history. Visual scores characterized the lungs as consisting of 29% total ILD extent (ground glass opacity=3%; reticular pattern=8%; honeycombing=17%)(C). CALIPER identified a total ILD extent of 27% (ground glass opacity=15%; reticular pattern=11%). PVV represented 6% of the lung volume (D).

146x131mm (72 x 72 DPI)



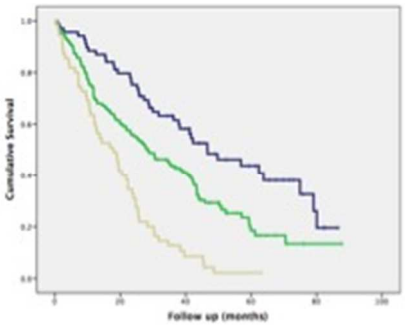
Two examples of delineation of pulmonary arteries and veins (coloured red), quantified by CALIPER as a percentage of the total lung volume and expressed as the pulmonary vessel volume (PVV). Note that segmentation is incomplete – for example a left upper lobe vessel in Figure 2c (arrow). The CALIPER software also misclassifies small volumes of reticulation as vessels (arrowhead).
Case 1 - A 62-year-old female ex-smoker, with a 15-pack year smoking history. Mean visual scores characterized the lungs as consisting of 9% total ILD extent (ground glass opacity=4%; reticular pattern=5%) (A). CALIPER identified a total ILD extent of 2% of the lung (ground glass opacity=1%; reticular pattern=1%). PVV represented 2% of the lung volume (B).
Case 2 - A 61-year-old female ex-smoker, with a 25-pack year smoking history. Visual scores characterized the lungs as consisting of 29% total ILD extent (ground glass opacity=3%; reticular pattern=8%; honeycombing=17%)(C). CALIPER identified a total ILD extent of 27% (ground glass opacity=15%; reticular pattern=11%). PVV represented 6% of the lung volume (D).

146x104mm (72 x 72 DPI)



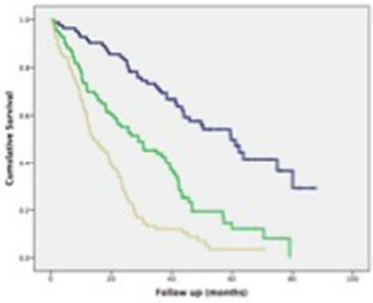
Two examples of delineation of pulmonary arteries and veins (coloured red), quantified by CALIPER as a percentage of the total lung volume and expressed as the pulmonary vessel volume (PVV). Note that segmentation is incomplete – for example a left upper lobe vessel in Figure 2c (arrow). The CALIPER software also misclassifies small volumes of reticulation as vessels (arrowhead).
Case 1 - A 62-year-old female ex-smoker, with a 15-pack year smoking history. Mean visual scores characterized the lungs as consisting of 9% total ILD extent (ground glass opacity=4%; reticular pattern=5%) (A). CALIPER identified a total ILD extent of 2% of the lung (ground glass opacity=1%; reticular pattern=1%). PVV represented 2% of the lung volume (B).
Case 2 - A 61-year-old female ex-smoker, with a 25-pack year smoking history. Visual scores characterized the lungs as consisting of 29% total ILD extent (ground glass opacity=3%; reticular pattern=8%; honeycombing=17%)(C). CALIPER identified a total ILD extent of 27% (ground glass opacity=15%; reticular pattern=11%). PVV represented 6% of the lung volume (D).

146x125mm (72 x 72 DPI)



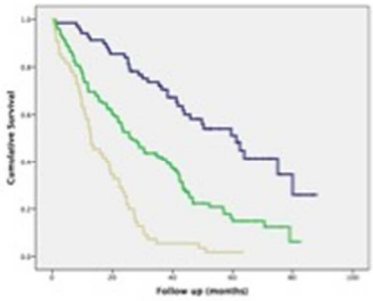
Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender, Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.
In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.

71x60mm (72 x 72 DPI)



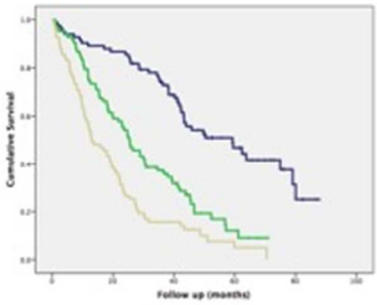
Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender, Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.
In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.

66x55mm (72 x 72 DPI)



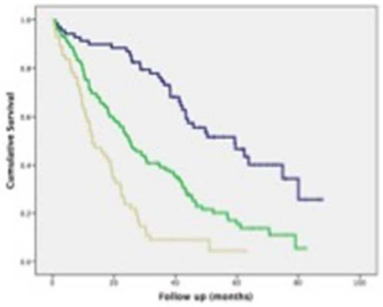
Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender, Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.
In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.

66x55mm (72 x 72 DPI)



Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender, Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.
In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.

67x56mm (72 x 72 DPI)



Survival curves for the IPF population in the current study. In the topmost graph the cases have been separated into three groups according to the Gender, Age, Physiology (GAP) index staging system. Mean survival and standard errors: Stage 1 (n=69, blue) 50·2±3·7; Stage 2 (n=125, green) 35·5±2·6; Stage 3 (n=55, yellow) 19·0±1·9.
In the middle row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of the three independent predictors of mortality in a combined visual, CALIPER and pulmonary function index multivariate model (CALIPER-CPI). The middle-left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (n=83, blue) 56·0±3·4; Group 2 (n=83, green) 31·9±2·7; Group 3 (n=83, yellow) 19·2±1·8. The middle-right graph consists of three groups each equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 56.0±3·6; Group 2 (n=125, green) 32·8±2·4; Group 3 (n=55, yellow) 16·7±1·8.
On the bottom row, stratification of patients into three groups was achieved using a formula derived from the hazard ratios of CALIPER variables that were independent predictors of mortality (CALIPER-only). The lower left graph comprises three groups of equal size. Mean survival and standard errors: Group 1 (blue) 55·4±3·3; Group 2 (green) 30·3±2·3; Group 3 (yellow) 20·1±2·1. The lower right graph contains three groups equal in size to the GAP groups. Mean survival and standard errors: Group 1 (n=69, blue) 55·5±3·6; Group 2 (n=125, green) 32·1±2·3; Group 3 (n=55, yellow) 18·9±2·1.

67x56mm (72 x 72 DPI)

Online Appendices:

CT Protocols

The CT scans were obtained using a 64-slice multiple detector CT scanner (Somatom Sensation 64, Siemens, Erlangen, Germany) or a 4-slice multiple detector CT scanner (Siemens Volume Zoom, Siemens, Erlangen, Germany. To satisfy requirements for processing by the CALIPER algorithm, all scans were reconstructed using a high spatial frequency, B70 kernel (Siemens, Munich, Germany). All patients were scanned from lung apices to bases, supine, at full inspiration, with 1·0mm section thicknesses using a peak voltage of 120kVp with tube current modulation (range 30-140 mA). Images were viewed at window settings optimized for the assessment of the lung parenchyma (width 1500 H.U.; level -500 H.U.).

Pulmonary function tests:

Pulmonary function tests were analyzed if performed within 3 months of the corresponding CT scan according to established protocols [31]. Spirometry (Jaeger Master screen PFT, Carefusion Ltd., Warwick, UK), plethysmographic lung volumes (Jaeger Master screen Body, Carefusion Ltd., Warwick, UK), and diffusion capacity for carbon monoxide (DLco) (Jaeger Master screen PFT, Carefusion Ltd., Warwick. UK) Parameters assessed: forced expiratory volume in one second (FEV1), forced vital capacity (FVC), total lung capacity (TLC), transfer coefficient of the lung for carbon monoxide (Kco) and single breath carbon monoxide diffusing capacity corrected for hemoglobin concentration (DLco). The composite physiologic index (CPI) was

calculated using the formula: 91·0 - (0·65 x % predicted DLco) - (0·53 x % predicted FVC) + (0·34 x % predicted FEV1).

Echocardiography:

A measure of pulmonary hypertension was obtained from right ventricular systolic pressure (RVSP) evaluated on transthoracic echocardiography, and only included if performed within 3 months of the CT scan.

Consensus formulation for visual scores:

The identification of systematic biases in visual scores was achieved by plotting the spread of differences in parenchymal pattern scores between observers. The most disparate 5% (two standard deviations) of values were arbitrated by a third scorer for all parameters except traction bronchiectasis, thereby minimizing bias within the original scorers. The original scorers derived a consensus for the traction bronchiectasis score. If a single parenchymal subtype extent was changed at consensus, the other parameters were modified, following CT review, to retain an overall sum of 100% for the four parenchymal subtypes. Similarly, if the lobar percentages of total interstitial disease, emphysema or mosaicism varied, the other two parameter extents were rescored.

Visual CT Variable (n = 283)	Single determination standard deviation
CT Interstitial lung disease extent	7·24
CT Ground glass opacity	6·15
CT Reticular pattern	5·24
CT Honeycombing	7·88
CT Consolidation	2·69
CT Total emphysema	4·99
CT Mosaic attenuation	3·83
CT Traction bronchiectasis severity	1·43

Table 1. Single determination standard deviation values of visual CT scores for idiopathic pulmonary fibrosis cases. CT = computed tomography.

BLINDED TITLE PAGE

Automated quantitative CT versus visual CT scoring in idiopathic pulmonary fibrosis: validation against pulmonary function

PURPOSE: To determine whether a novel CT post-processing software technique (CALIPER) is superior to visual CT scoring as judged by functional correlations in idiopathic pulmonary fibrosis.

MATERIALS and METHODS: 283 consecutive patients with idiopathic pulmonary fibrosis had CT parenchymal patterns evaluated quantitatively with CALIPER and by visual scoring. These two techniques were evaluated against: FEV1, FVC, DLco, Kco and a composite physiological index (CPI), with regard to extent of interstitial lung disease, extent of emphysema and pulmonary vascular abnormalities.

RESULTS: CALIPER-derived estimates of interstitial lung disease extent demonstrated stronger univariate correlations than visual scores for most pulmonary function tests: (FEV1: CALIPER R^2=0·29, visual R^2=0·18, FVC: CALIPER R^2=0·41, visual R^2=0·27; DLco: CALIPER R^2=0·31, visual R^2=0·35; CPI: CALIPER R^2=0·48, visual R^2=0·44). Correlations between CT measures of emphysema extent and pulmonary function tests were weak and did not differ significantly between CALIPER and visual scoring. Intriguingly, the pulmonary vessel volume provided similar correlations to total interstitial lung disease extent scored by CALIPER for FVC, DLco and CPI (FVC: R^2=0·45; DLco: R^2=0·34; CPI: R^2=0·53).

CONCLUSIONS: CALIPER was superior to visual scoring as validated by functional correlations with pulmonary function tests. The pulmonary vessel volume, a novel CALIPER CT parameter with no visual scoring equivalent, has the potential to be a CT feature in the assessment of patients with idiopathic pulmonary fibrosis and requires further exploration.

KEY WORDS

QUANTITATIVE COMPUTER ANALYSIS

IDIOPATHIC PULMONARY FIBROSIS

VISUAL CT ANALYSIS

PULMONARY VESSEL VOLUME

INTRODUCTION

Computer-based quantitative CT evaluation has the potential for greater precision than visual scoring in the estimation of the extent of diffuse parenchymal diseases. A new generation of computer-based CT software tools have demonstrated similar results between computer quantitation and visual quantitative scoring in small-scale studies in patients with idiopathic pulmonary fibrosis (IPF), [1, 2] with distinct improvement in performance on older, less sophisticated software programs.[3, 4]

Pulmonary damage in IPF is the consequence of pathological involvement of three components of the lung: the parenchyma, the vasculature (largely due to pulmonary hypertension)[5] and, in a large proportion of IPF patients who are cigarette smokers, co-existent emphysema.[6, 7] Our study assessed baseline involvement of these three compartments using traditional visual CT evaluation and a sophisticated quantitative CT software tool, CALIPER (Computer-Aided Lung Informatics for Pathology Evaluation and Rating), developed at the (_____). In our study, we set out to validate computer-based CT scoring by examining correlations between CT patterns against pulmonary function tests.

MATERIALS and METHODS

Study population and clinical information:

A retrospective analysis of an interstitial lung disease database identified all new consecutive patients, (over a four and a half year period (January 2007 to July 2011)), with a multidisciplinary team diagnosis of IPF according to published guidelines.[8] Patients with a non-contrast, supine, volumetric thin section CT were collected, and subsequent exclusions are shown (as per CONSORT in Figure 1). Pulmonary function and echocardiography protocols are explained in the online appendix. The DICOM images for the CT scans were transferred to the (_____) for blinded CALIPER processing. Approval for this study of clinically indicated CT and pulmonary function data was obtained from the Institutional Ethics Committee of the (_____) and the Institutional Review Board of (_____).

CT protocol:

The CT scans were obtained using a 64-slice multiple detector CT scanner (Somatom Sensation 64, Siemens, Erlangen, Germany) or a 4-slice multiple detector CT scanner (Siemens Volume Zoom, Siemens, Erlangen, Germany. To satisfy requirements for processing by the CALIPER algorithm, all scans were reconstructed using a high spatial frequency, B70 kernel (Siemens, Munich, Germany). All patients were scanned from lung apices to bases, at full inspiration, using a peak voltage of 120kVp with tube current modulation (range 30-140 mA). Images of 1mm thickness were viewed at window settings optimized for the assessment of the lung parenchyma (width 1500 H.U.; level -500 H.U.).

CT visual evaluation:

Each CT scan was evaluated independently by two radiologists (___,___) with 5 and 7 years thoracic imaging experience respectively, blinded to all clinical information. An initial training dataset of 15 non-study cases was used to help to identify pre-existing biases. The scores of the test cases were reviewed and the most widely discrepant results discussed with a third radiologist (__).

CTs were scored on a lobar basis using a continuous scale. The total interstitial lung disease (ILD) extent was estimated to the nearest 5%, and sub-classified into four patterns: reticular pattern, ground glass opacification, honeycombing and consolidation, using definitions from the Fleischner Society glossary of terms for thoracic imaging.[9] To derive a lobar percentage for each parenchymal pattern, the total lobar ILD extent was multiplied by individual lobar parenchymal pattern extents and divided by 100. The percentage (to the nearest 5%) of each lobe that contained mosaicism (decreased attenuation component) or emphysema was recorded. The individual lobar percentages of each parenchymal pattern were summed for each radiologist and divided by six to create an averaged lobar score per pattern, per scorer per case.

Traction bronchiectasis, as defined in the Fleischner society glossary of terms, [9] was assigned with a categorical "severity" score that took into account the average degree of airway dilatation within areas of fibrosis as well as the extent of dilatation throughout the lobe and was given an overall score of: none=0, mild=1, moderate=2, severe=3. An index of pulmonary hypertension (main pulmonary artery:ascending aorta ratio) was assessed by a single scorer using electronic caliper diameter measurements of the ascending aorta and pulmonary artery diameters at the level of the pulmonary artery bifurcation.[10]

The identification of systematic biases in visual scores was achieved by plotting the spread of differences in parenchymal pattern scores between observers. The most disparate 5% (two standard deviations) of values were arbitrated by a third scorer for all parameters except traction bronchiectasis, thereby minimizing bias within the original scorers. The original scorers derived a consensus for the traction bronchiectasis score. If a single parenchymal subtype extent was changed at consensus, the other parameters were modified, following CT review, to retain an overall sum of 100% for the four parenchymal subtypes. Similarly, if the lobar percentages of total interstitial disease, emphysema or mosaicism varied, the other two parameter extents were rescored.

CALIPER CT evaluation:

Data processing: Initial data processing steps involved extraction of the lung from the surrounding thoracic structures and segmentation into upper, middle and lower zones. Lung segmentation was performed with an adaptive density-based morphological approach,[11] whilst airway segmentation involved iterative three-dimensional region growing, density thresholding (thresholds including -950HU and -960HU) and connected components analysis.

Parenchymal tissue type classification was applied to 15x15x15 voxel volume units using texture analysis, computer vision-based image understanding of volumetric histogram signature mapping features and 3D morphology.[12] The CALIPER tool was trained by sub-specialty thoracic radiologist consensus assessment of pathologically confirmed datasets.[2, 12]

Pattern evaluation: CALIPER and visual CT were used to quantify pulmonary variables in three domains:

i) total extent of interstitial disease

ii) total extent of emphysema

iii) pulmonary vessels

CALIPER evaluation of CT data involved algorithmic identification and volumetric quantification of every voxel volume unit into one of eight radiological parenchymal features: normal lung, three grades of decreased lung attenuation (grade 1=mild, 2=moderate 3=marked), ground glass opacification, reticular pattern, honeycombing and the pulmonary vessels (Figure 2). Volumes for all eight parenchymal features were converted into a percentage using the total lung volume also measured by CALIPER.

The final CALIPER emphysema index was defined by pre-processing the CALIPER variables against PFTs and visual comparison of CT scans. Using univariate linear correlations, Grade 1 decreased attenuation (DA) demonstrated no fit against the visual emphysema extent score or against Kco and was therefore not included within the CALIPER emphysema score. On inspection of colour maps, Grade 1 DA was shown to encompass areas of patchy centrilobular emphysema and considerable

amounts of normal lung (Figure 2a-c), which accounted for its large extent and lack of functional impact (Table 1). The sum of Grade 2 and 3 DA was taken to represent CALIPER emphysema and the suitability of the new variable was confirmed on analysis of individual colour maps (Figure 2a-c) that demonstrated that Grade 2 and 3-DA corresponded to discrete and conglomerate foci of emphysema.

Segmentation of pulmonary vessels, prior to their extraction, was achieved using an optimized multi-scale tubular structure enhancement filter based on the eigenvalues of the Hessian matrix. The filters calculated the 2nd-order derivatives that occurred in the regions that surrounded each pulmonary voxel. The eigenvalues of the Hessian matrix that were constructed from the derivatives were then analyzed, and from these values, it was possible to determine the likelihood that an underlying voxel was connected to a dense tubular structure and therefore represented a vessel.[12, 13] The pulmonary vessel volume (PVV) score quantified the volumes of pulmonary arteries and veins excluding vessels at the lung hilum as a percentage of lung volume (Figure 3). Total ILD extent represented the sum of ground glass, reticular and honeycomb percentages.

Statistical analysis:

Data are given as means with standard deviations, or numbers of patients with percentages where appropriate. Interobserver variation for visual scores was assessed using the single determination standard deviation.

Correlations between the extents of parenchymal patterns and individual PFTs were examined using Pearson's product moment correlation. Univariate and multivariate analyses were undertaken to investigate relationships between CALIPER or visual CT evaluation and pulmonary function tests. In multivariate analyses, robustness of relationships were tested by bootstrapping the dataset with 1000 samples. In all study analyses, a p-value of <0.01 was considered significant. Models were formally tested for heteroscedasticity to confirm that the assumptions of parametric analysis had been satisfied. Statistical analyses were performed with STATA (version 12, StatCorp, College Station, TX, USA).

RESULTS:

Baseline data

The final study group comprised 283 consecutive patients with a multidisciplinary diagnosis of IPF. Age, gender, mean visual scores, CALIPER scores and pulmonary function tests are shown in Table 1. Discordances between CALIPER and visual estimations of total ILD extent, emphysema and the various individual parenchymal patterns are shown in Table 1 and an illustrative example is shown in Figure 2. Visual scores identified on average 1.6 times more ILD than CALIPER and 10 times more honeycombing than CALIPER. Interobserver variation values for the visual scores are provided in Table 2.

Univariate relationships

The relationships between pulmonary function parameters (FEV1, FVC, DLco, Kco, CPI) and CT variables are shown in Table 3. Taken across the PFTs, CALIPER ILD extent was either superior to (FEV1, FVC) or comparable with (DLco, CPI) visual scoring (Table 3). CALIPER ILD extent and PVV had very similar correlations with pulmonary function indices. The PVV correlations were either superior to (FEV1, FVC, CPI) or comparable with (DLco) visual ILD extent scores. PVV increased with ILD extent (Quadratic R^2=0.76), but less so with more advanced disease. The visual pulmonary artery:aorta diameter ratio demonstrated no relationship to any functional index.

Multiple regression

Multivariate regression analyses of visual and CALIPER scores of pulmonary vasculature, total interstitial disease extent and emphysema extent were analysed against pulmonary function tests (Table 4). PVV and ILD extent could not be included in the same model due to major co-linearity and so were examined in separate models (Model 1 containing CALIPER ILD and Model 2 containing PVV).

In Model 1, CALIPER ILD extent was clearly superior for two PFTs (FEV1, FVC), (confirmed on bootstrapping the dataset with 1000 samples), with visual ILD extent discarded. CALIPER and visual ILD extents were complementary for two PFTs (DLco, CPI) with both variables retained (again

confirmed on bootstrapping the dataset with 1000 samples). In the second model, CALIPER PVV was the strongest determinant of all examined PFTs. Although visual ILD extent was retained for DLco and CPI, there was minimal effect on model fit when it was discarded (DLco change: R^2 0.04; CPI change R^2 0.03). Similarly, the inclusion of individual CT patterns (e.g. ground glass opacity, reticular pattern and honeycombing), whether quantified by CALIPER or visually, resulted in only a minimal improvement in correlations with PFTs (R^2 values increasing by <0.02).

Given the overall strong correlations between PVV and the various pulmonary function tests, relationships between PVV and RVSP were explored. Univariate analysis of a subgroup of 150 patients with concurrent right ventricular systolic pressure (RVSP) measured on echocardiography was performed. RVSP was found to explain 20% of the variability of the PVV (R^2=0·20, P<0·0001). In the same 150 patients, the extent of ILD measured by both CALIPER (R^2=0·73, P<0·0001) and visually (R^2=0·47, P<0·0001), better explained PVV variability. Furthermore, with these two CT variables included in separate models, RVSP had no independent linkage with PVV.

Post-hoc evaluation: CALIPER-based validation of CPI

On the basis of the strong linkages between CALIPER ILD extent and PFTs, the robustness of the CPI was validated using CALIPER-scored ILD extent in the current study group. This showed that the same three best-fit PFTs, as used in the original CPI model, predicted CALIPER ILD extent:

$$\text{CALIPER-derived CPI} = 66.0 - (0.47 \times \text{DLco}) - (0.67 \times \text{FVC}) + (0.32 \times \text{FEV1})$$

Correlation of the new CPI against the old CPI score demonstrated a strong linkage as shown in the graph in Figure 4 (R^2=0.95).

DISCUSSION

Our study has shown that in an IPF population, CALIPER derived interstitial and vascular parameters correlated more strongly with pulmonary function indices (FEV1, FVC) or were at least comparable to visual scores (DLco, CPI). Consequently, it was felt logical to explore the robustness of the CPI using an objective scoring methodology (CALIPER) and, in so doing, the CPI was vindicated. Importantly, we have also shown that the pulmonary vessel volume, a novel CALIPER-derived percentage of the lung composed of pulmonary arteries and veins, is surprisingly strongly linked to the extent of interstitial disease.

The increase in the size and number of drug trials in IPF has necessitated the development of new automated computer-based algorithms capable of analysing hundreds of CTs per study. The accuracy of the new computer-based techniques requires validation to ensure that they are at least comparable to visual CT scoring. One of the steps in validating the accuracy of CT in assessing disease extent is by examining the relationship between CT estimates, however obtained, and pulmonary function measures of disease severity. There are numerous structure-function studies in interstitial lung disease, but these have almost exclusively relied on visual scoring of total and individual pattern extents, with all the inherent problems of interobserver agreement.[14-16]

In our study, CALIPER-derived interstitial and vascular markers clearly demonstrated stronger or comparable correlations with all cardinal pulmonary function tests than visual scores. The strong univariate correlations between both visual and CALIPER-scored total ILD extent and pulmonary function parameters are in line with previous IPF studies. [16, 17] Whilst the primary purpose of this study was a validation exercise of CALIPER against visual scores, the strong correlations between CALIPER scores and PFTs made it possible to examine historical derivation of the CPI. The CPI, originally derived using a subjective visual scoring system, was validated as a robust variable following replication of the CPI by using scores from the objective CALIPER system.

CALIPER was less sensitive to the extent of emphysema as compared to visual scoring and less accurate as judged by correlations with pulmonary function indices. Nevertheless, the positive correlations between visually and CALIPER scored emphysema and FVC, identified in the current study, are consistent with previous studies in IPF patients with co-existent emphysema. [18, 19]

On multivariate analysis of CALIPER and visual variables, it was striking that the strongest independent parameter predictive of pulmonary function tests was the PVV. We had assumed a priori that PVV was a measure of pulmonary vessel involvement, but intriguingly our investigations revealed that PVV was not only a ILD marker per se, but that it was co-linear with both CALIPER and visual ILD extents. PVV was at least as strong as CALIPER ILD extent in its linkage with key PFTs (FVC, CPI). Furthermore, after correcting for the extent of ILD, no linkage between PVV and RVSP remained, thereby establishing that PVV was not an index of pulmonary hypertension.

Discordances between CALIPER and visual scores of total ILD extent and honeycombing were assessed by analyzing outlying cases and were found to relate largely to differences in scoring methodologies. On visual evaluation, each lobe represented a sixth of the total lung volume, regardless of the extent of lobar disease. CALIPER however, assessed parenchymal patterns, as a proportion of the total lung volume. Lower lobes, contracted by the retractile fibrosis of UIP, contribute a smaller percentage of disease than, for example, non-fibrotic upper lobes and were thus under-represented by CALIPER when compared to visual scores. The disparity in visual and CALIPER estimates of reticular pattern and ground glass opacity reflect differences in categorization of a pattern of intermixed fine reticulation and ground glass opacity (Figure 2d-f). Furthermore, review of individual cases showed that quite frequently, CALIPER characterized visually scored honeycombing as reticular pattern and ground glass opacity (Figure 2g-i).

Our findings suggest that a computer-based quantitative CT tool such as CALIPER, has several valuable roles for the evaluation of patients with IPF. The improved sensitivity of CALIPER in evaluating ILD extent, when compared to visual scoring, has the potential to enhance understanding of the natural history of IPF by improving the accuracy of identifying serial change. In the sphere of drug trials,

computer-based CT evaluation has several possible applications. CALIPER could be used to correct for baseline CT disease extent in patients at the start of a trial and it could also be applied as a monitoring tool in the context of end-points. In the current study, recalibration of CALIPER software was required to analyse the Siemens B70f algorithms performed in our department, with the result that post-recalibration, similar results for parenchymal pattern extents were achieved when CALIPER analysed both edge-enhancing algorithms such as a Siemens B70f and "less edge-enhancing" algorithms such as a Siemens B46f, (the algorithm constituting most of CALIPER's original training dataset). A consequence of CALIPERs recalibration was an improvement in its versatility, which has relevance for multicentre drug trials, where CTs in different centres can be reconstructed with a range of different algorithms.

CALIPER could also make a useful contribution to the investigation of combined fibrosis and emphysema (CPFE). The strong correlations with functional indices we have demonstrated suggest that CALIPER would be a more suitable tool than visual scores to quantify ILD extent in CPFE, although conversely visual scoring may be best placed to quantify emphysema extent. Both methodologies used together may better delineate the contribution of each component of CPFE.

With regard to PVV, individual colour overlay maps demonstrated some contamination of the variable by areas of reticulation (Figure 3), particularly in cases with extensive pulmonary fibrosis. Nonetheless, the variable primarily reflects the quantitation, by CALIPER, of large and small vessels in the lung, in a way that has not, to date, been possible by human scorers. Evaluating this new parameter has resulted in a credible additional CT measure to visual ILD extent scores when quantifying interstitial involvement in IPF. A possible explanation for the relationship between PVV and ILD extent relates to the increased negative intra-thoracic pressure that non-compliant fibrotic lungs generate during inspiration. The transmission of high negative pressures into the lung parenchyma could in turn affect compliant vessels, resulting in vascular dilatation throughout the lung and an increase in capacitance. However deciphering the exact pathophysiological mechanisms that link interstitial damage to vascular volume requires further investigation. Furthermore, the potential prognostic role of PVV in patients with fibrosing lung disease is worthy of exploration.

Some limitations to the CALIPER technique are evident. One lies in the poor correlations we have identified between CALIPER emphysema extent and functional indices, (in particular Kco). Improving the detection of non-conglomerate emphysema by CALIPER would be a preliminary yet feasible objective. The complexities in scoring CT parenchymal pattern extents, be that visual or computer-based, in a disease that is inevitably associated with volume loss are considerable. Whilst discrepancies between CALIPER and visual scores for parenchymal patterns such as honeycombing have been partly explained, further studies directed towards clarifying the reasons behind the differences in disease extent scores are needed. Lastly, the minor contamination of the PVV signal by reticulation in cases with severe fibrosis, might be considered a limitation, as it could be thought to dilute the relationships between PVV and functional indices. However improvement and greater sophistication of the algorithm to detect vessels (versus reticulation) may result in simply strengthening the correlations we have already shown.

In conclusion, we have shown that CALIPER measures of lung disease are more strongly related to pulmonary function tests than visual scores. Strong links between CALIPER estimation of pulmonary vessel volume and pulmonary function tests suggests that evaluation of pulmonary vessel volume may be an important new index when assessing disease severity in patients with IPF.

REFERENCES

1. 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 7 Months. *Acad Radiol*. 2015;**22**(1):70-80.
2. Maldonado F, Moua T, Rajagopalan S, et al. Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. *Eur Respir J*. 2014;**43**(1):204-12.
3. Best AC, Meng J, Lynch AM, et al. Idiopathic Pulmonary Fibrosis: Physiologic Tests, Quantitative CT Indexes, and CT Visual Scores as Predictors of Mortality. *Radiology*. 2008;**246**(3):935-40.
4. Iwasawa T, Asakura A, Sakai F, et al. Assessment of prognosis of patients with idiopathic pulmonary fibrosis by computer-aided analysis of CT images. *J Thorac Imaging*. 2009;**24**(3):216-22.
5. Nadrous HF, Pellikka PA, Krowka MJ, et al. Pulmonary hypertension in patients with idiopathic pulmonary fibrosis. *Chest*. 2005;**128**(4):2393-9.
6. Cottin V, Nunes H, Brillet PY, et al. Combined pulmonary fibrosis and emphysema: a distinct underrecognised entity. *European Respiratory Journal*. 2005;**26**(4):586-93.
7. Sugino K, Ishida F, Kikuchi N, et al. Comparison of clinical characteristics and prognostic factors of combined pulmonary fibrosis and emphysema versus idiopathic pulmonary fibrosis alone. *Respirology*. 2014;**19**(2):239-45.
8. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med*. 2011;**183**(6):788-824.
9. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: glossary of terms for thoracic imaging. *Radiology*. 2008;**246**(3):697-722.
10. Ng CS, Wells AU, Padley SP. A CT sign of chronic pulmonary arterial hypertension: the ratio of main pulmonary artery to aortic diameter. *J Thorac Imaging*. 1999;**14**(4):270-8.
11. Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE Trans Med Imaging*. 2001;**20**(6):490-8.
12. Bartholmai BJ, Raghunath S, Karwoski RA, et al. Quantitative CT imaging of interstitial lung diseases. *J Thorac Imaging*. 2013;**28**(5):298-307.
13. Shikata H, McLennan G, Hoffman EA, Sonka M. Segmentation of pulmonary vascular trees from thoracic 3D CT images. *Int J Biomed Imaging*. 2009:11.
14. Mogulkoc N, Brutsche MH, Bishop PW, Greaves SM, Horrocks AW, Egan JJ. Pulmonary function in idiopathic pulmonary fibrosis and referral for lung transplantation. *Am J Respir Crit Care Med*. 2001;**164**:103-8.
15. Lynch DA, David GJ, Safrin S, et al. High-resolution computed tomography in idiopathic pulmonary fibrosis: diagnosis and prognosis. *Am J Respir Crit Care Med*. 2005;**172**:488-93.
16. Xaubet A, AgustÍ C, Luburich P, et al. Pulmonary Function Tests and CT Scan in the Management of Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med*. 1998;**158**(2):431-6.

17. Wells AU, Desai SR, Rubens MB, et al. Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography. *Am J Respir Crit Care Med*. 2003;**167**:962-9.
18. Akagi T, Matsumoto T, Harada T, et al. Coexistent emphysema delays the decrease of vital capacity in idiopathic pulmonary fibrosis. *Respir Med*. 2009;**103**:1209-15.
19. Kim YJ, Shin SH, Park J-W, et al. Annual Change in Pulmonary Function and Clinical Characteristics of Combined Pulmonary Fibrosis and Emphysema and Idiopathic Pulmonary Fibrosis: Over a 3-Year Follow-up. *Tuberc Respir Dis*. 2014;**77**:18-23.

FIGURE CAPTIONS

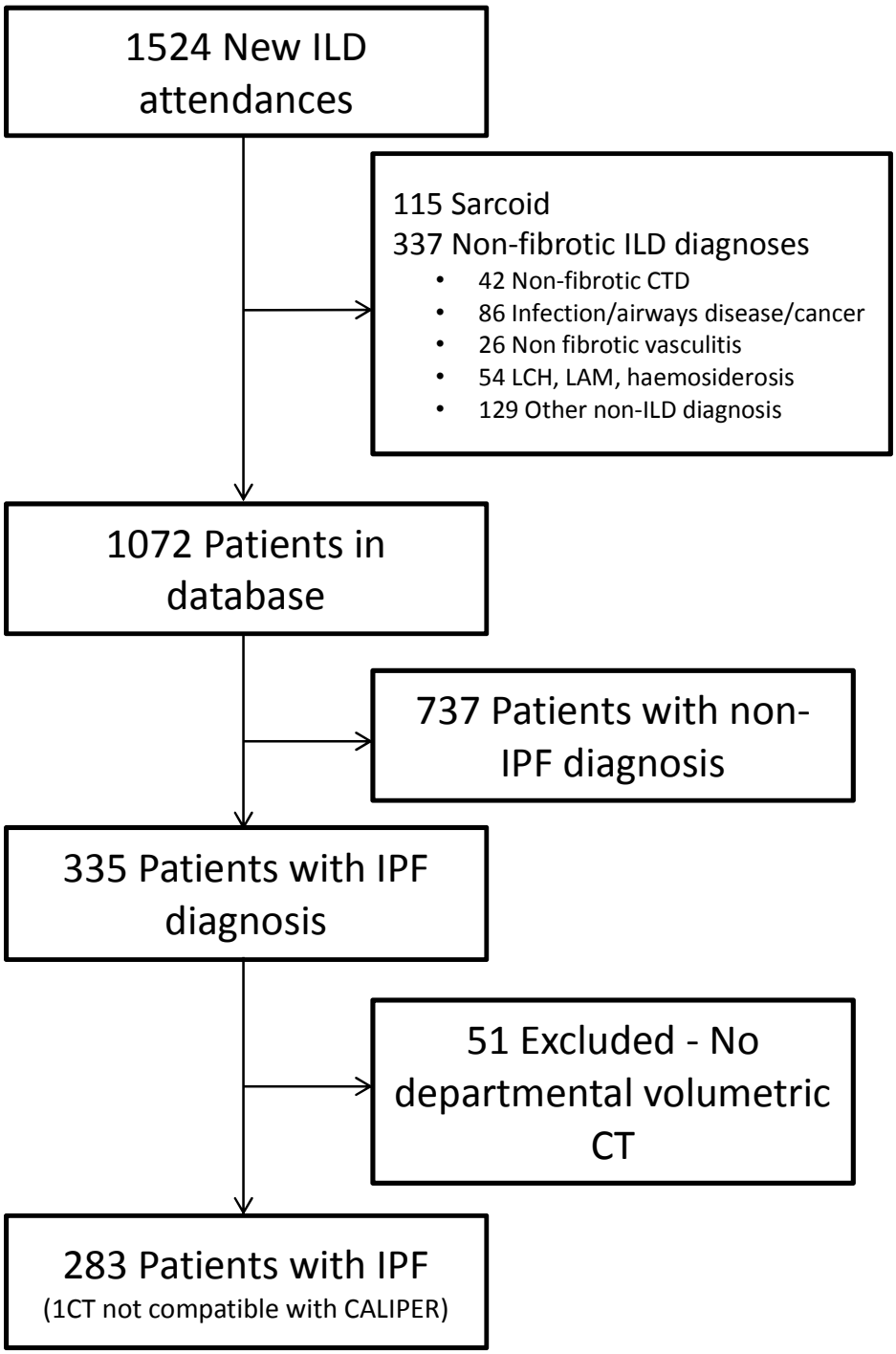
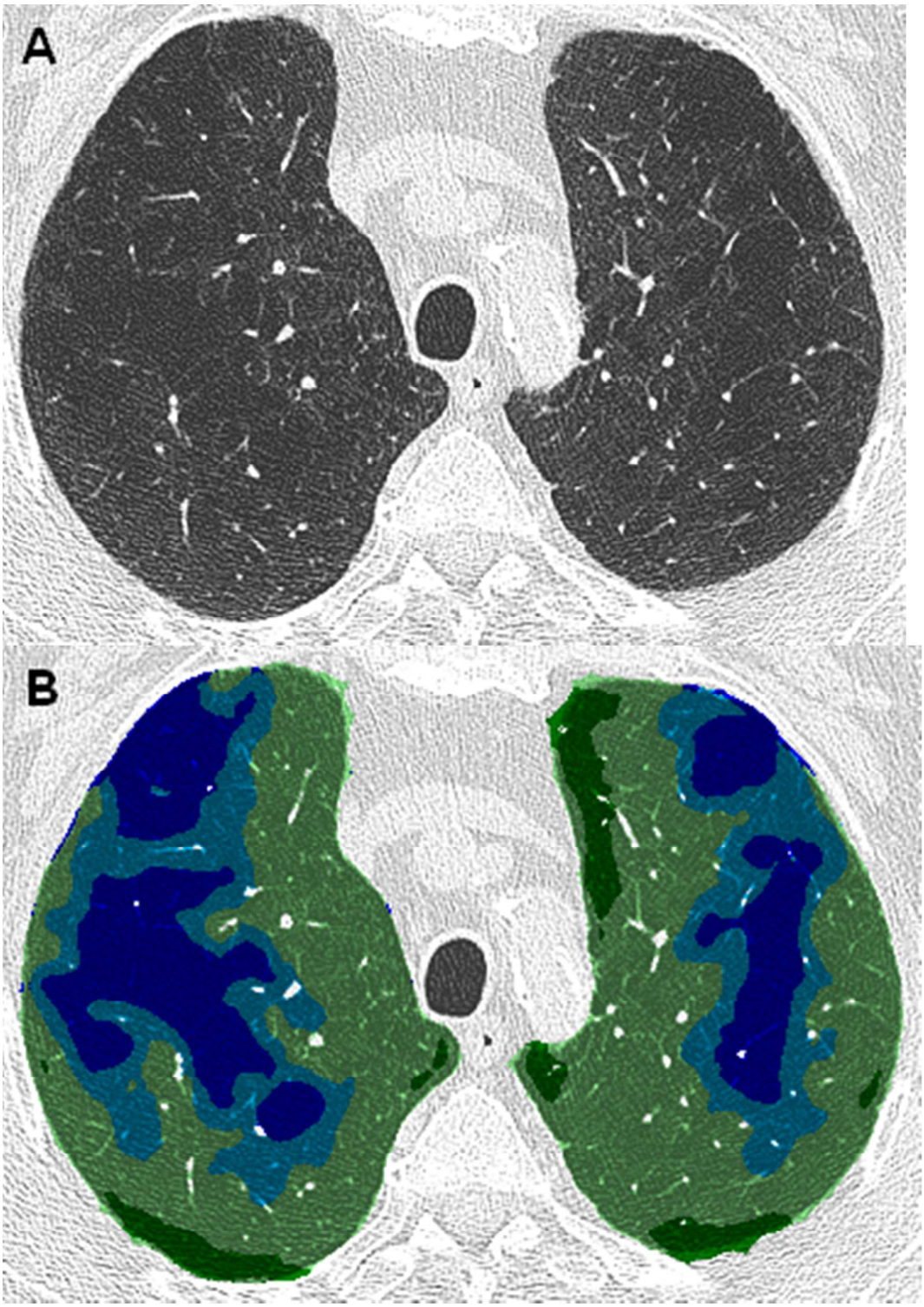
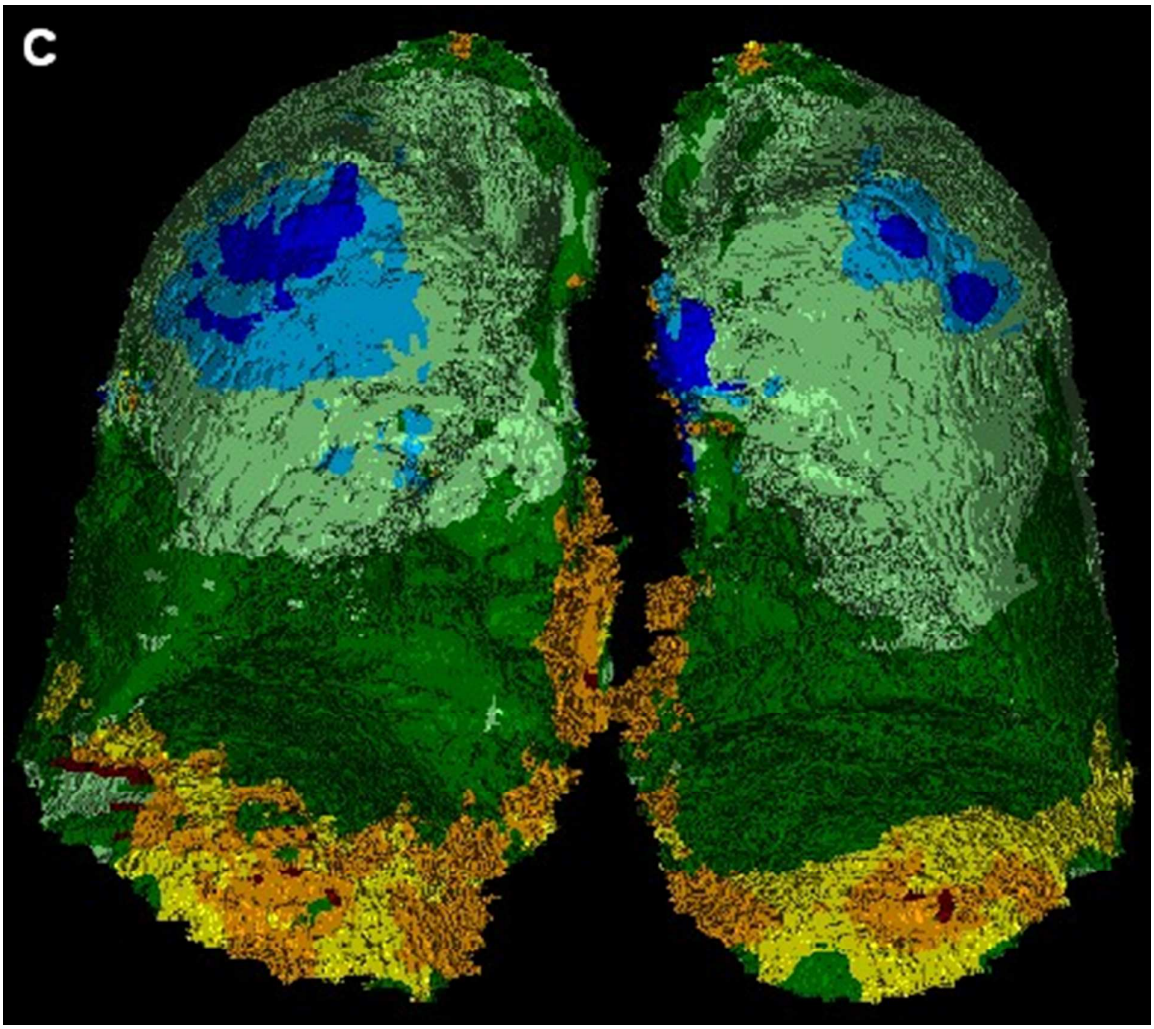
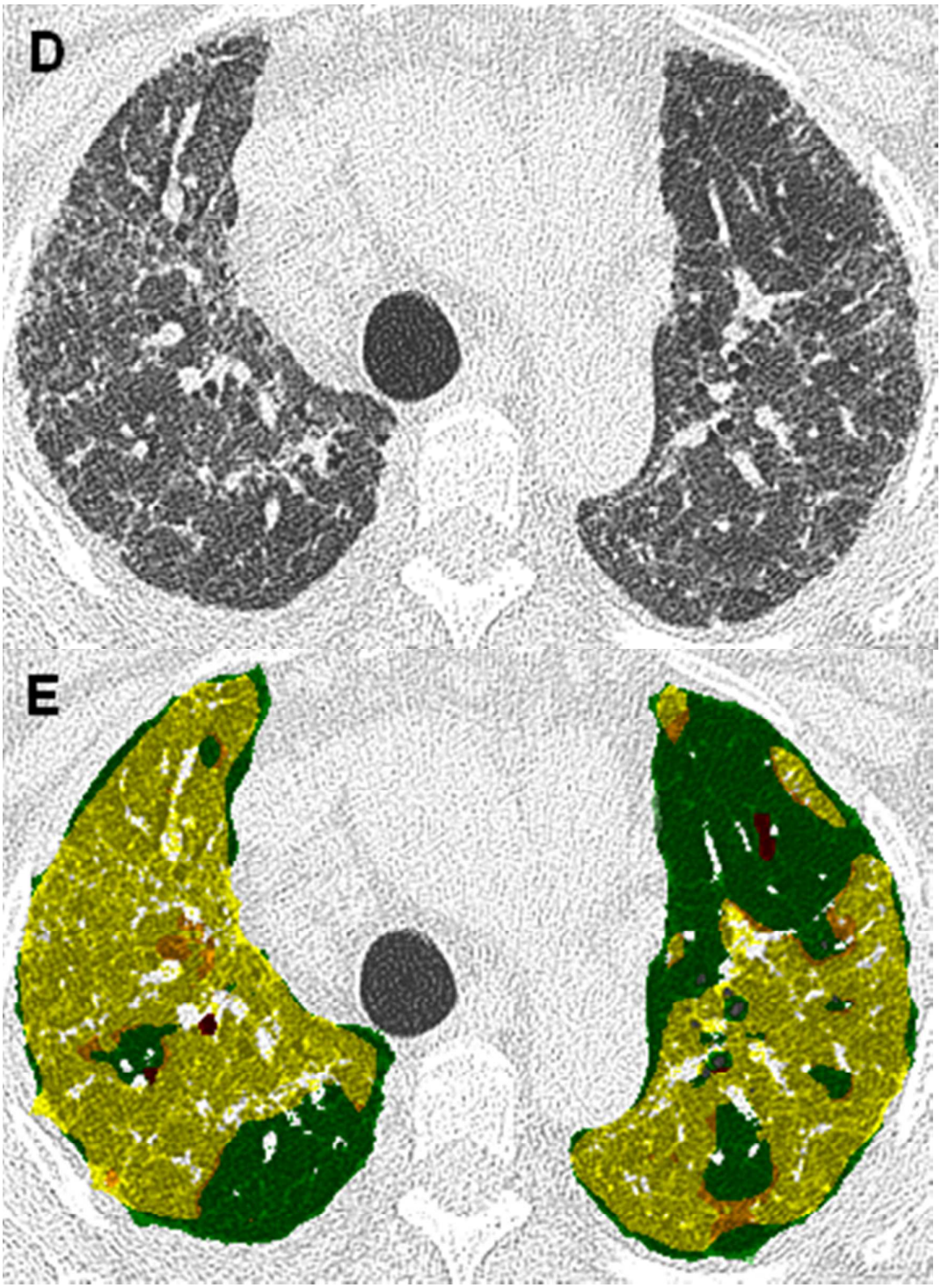
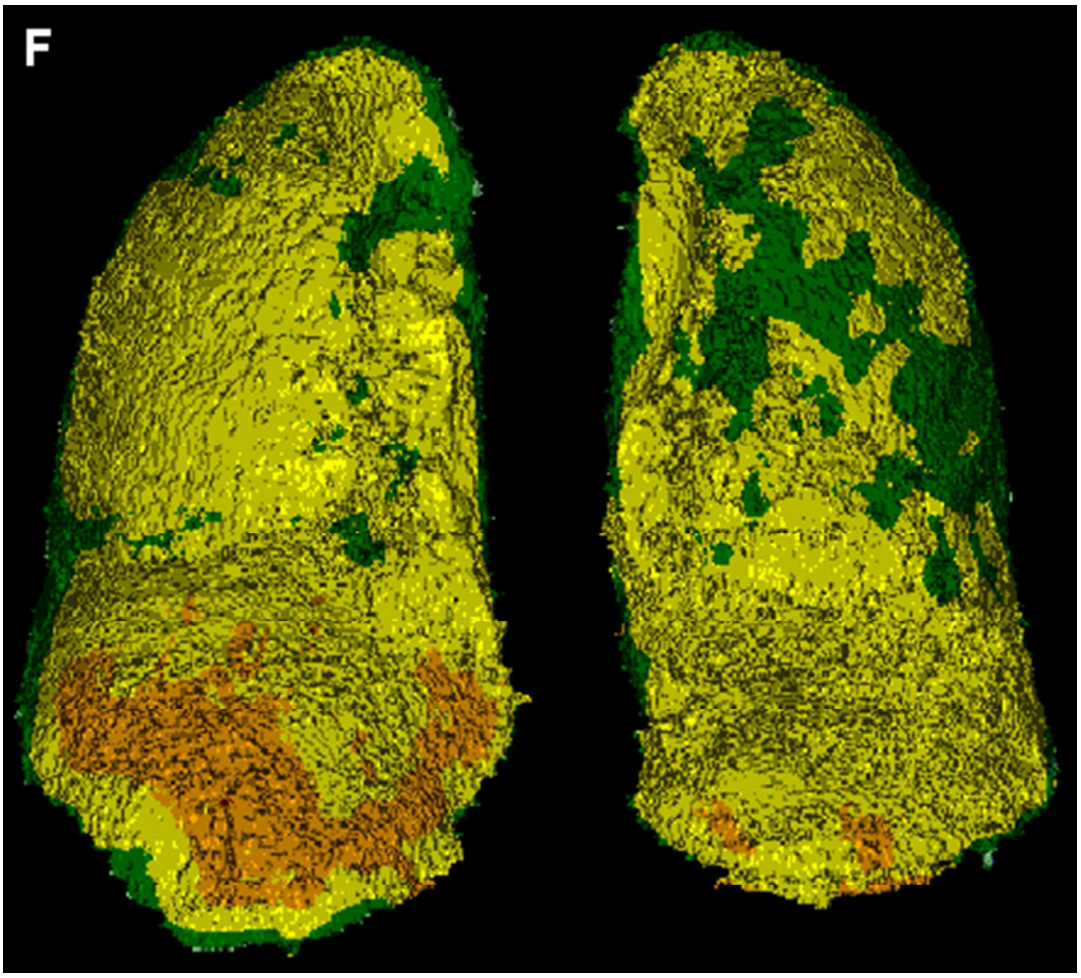


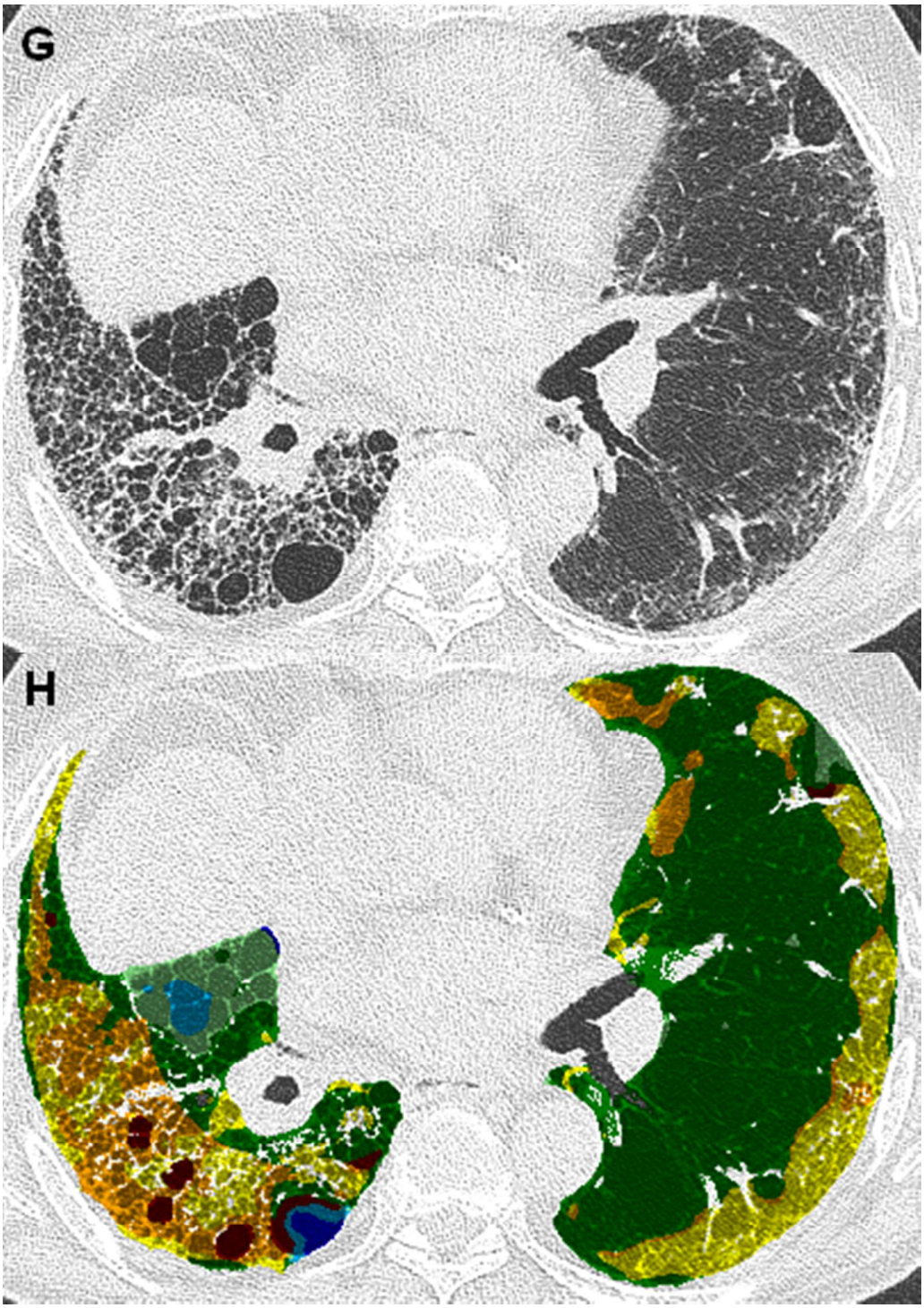
Figure 1. CONSORT diagram illustrating the selection of patients for the final study population. ILD = interstitial lung disease, CTD = connective tissue disease, IPF = idiopathic pulmonary fibrosis, LCH = Langerhans cell histiocytosis, LAM = Lymphangioleiomyomatosis, CT = computed tomography.


A
B


C


D
E


F


G
H

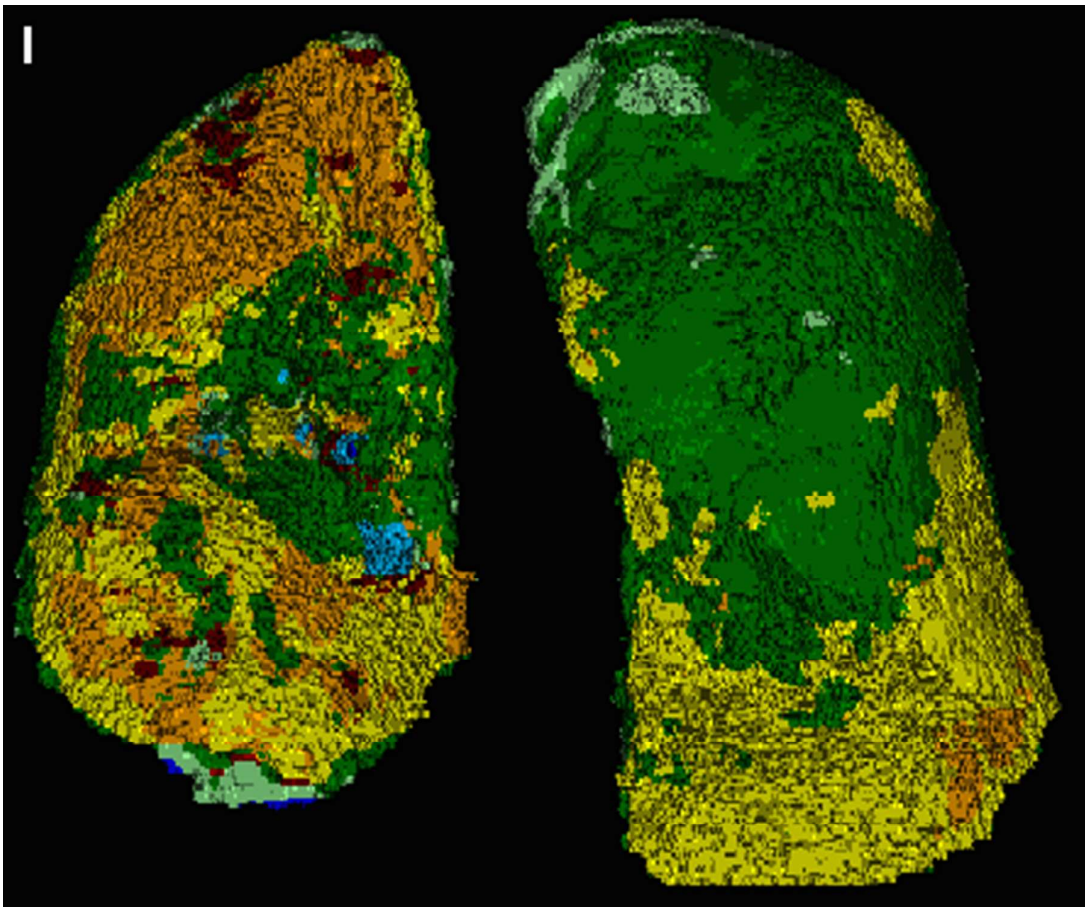


Figure 2A-C. Axial CT slices, axial CALIPER-derived colour image overlays and three-dimensional CALIPER coronal rendering of the lungs displaying parenchymal patterns as various colours (Dark green=normal lung, light green=grade 2 decreased attenuation area, light blue=grade 2 decreased attenuation area, dark blue=grade 3 decreased attenuation area, yellow=ground glass opacity, orange=reticular pattern, brown=honeycombing).

2A(i-iii) 75-year-old male ex-smoker with a 40-pack year smoking history. Mean visual scores of the CT: 7% reticular pattern, 3% honeycombing, no ground glass opacity and 24% emphysema (i). CALIPER characterized 5% reticular pattern, 1% honeycombing, 2% ground glass opacity, 43% Grade 1, 10% Grade 2 and 4% Grade 3 decreased attenuation and 3% pulmonary vessel volume (PVV). A large proportion of the areas with interspaced patches of centrilobular emphysema were characterized as Grade I decreased attenuation by CALIPER as demonstrated on overlaid axial ii) and 3D rendered

images (iii) but characterized as predominantly normal lung with only minor emphysema on visual scores.

2B(i-iii) 48-year-old female never-smoker. Mean visual scores of the CT: 34% reticular pattern, 0.5% honeycombing, 45% ground glass opacity with no emphysema (i). CALIPER characterized 6% reticular pattern, 1% honeycombing, 62% ground glass opacity, no Grade 2 and 3 decreased attenuation and 8% PVV. A large proportion of the areas visually labeled reticular pattern were characterized as ground glass opacity by CALIPER as demonstrated on overlaid axial (ii) and 3D rendered images (iii) reflecting a pattern of textured ground glass opacity that is often difficult to classify.

2C(i-iii) 76-year-old male ex-smoker with a 20-pack year smoking history. Mean visual scores of the CT: 14% reticular pattern, 49% honeycombing, 5% ground glass opacity with no emphysema (i). CALIPER characterized 16% reticular pattern, 2% honeycombing, 24% ground glass opacity, 0·5% Grade 2 and 3 decreased attenuation and 9% PVV. A substantial proportion of the areas visually labeled honeycombing were characterized as reticular pattern and/or ground glass opacity by CALIPER as shown on overlaid axial (ii) and 3D rendered images (iii).

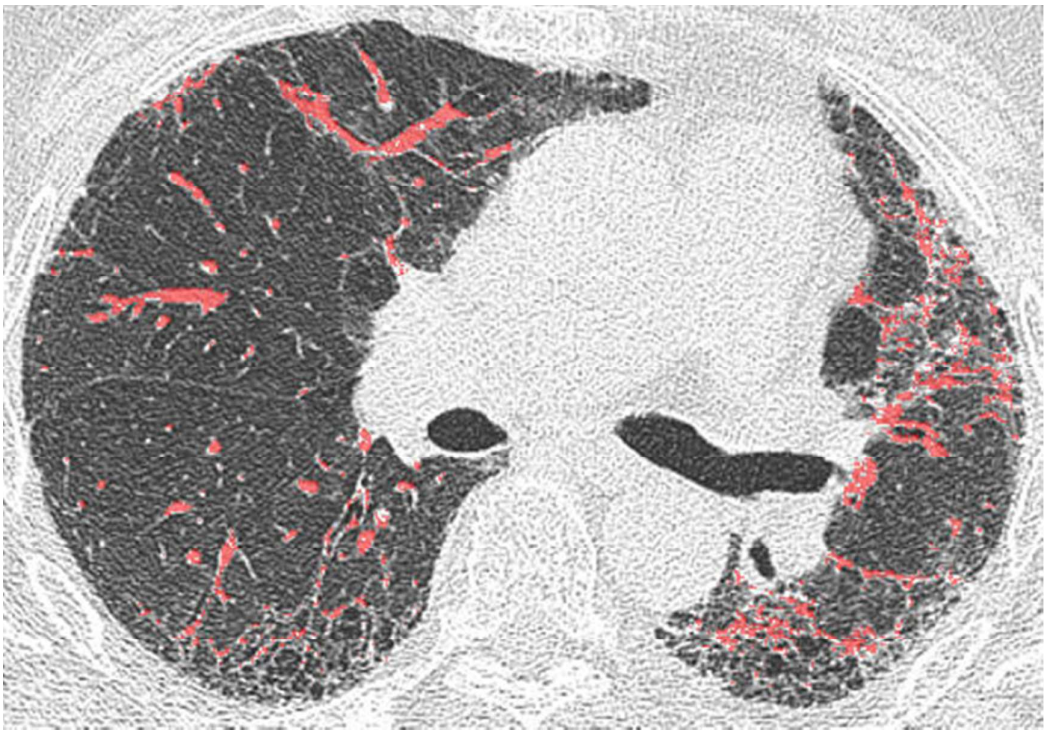


Figure 3. A 75-year-old ex-smoker diagnosed with idiopathic pulmonary fibrosis. Axial CT slice with CALIPER-derived colour overlay demonstrating intraparenchymal pulmonary arteries and veins (red). During the initial extraction process separating the lungs from the mediastinum and chest wall, the hilar structures including the central pulmonary arteries and veins were removed. Pulmonary vessels are classified by CALIPER using structure and textural analysis and computer vision-based image understanding of volumetric histogram signature mapping features for 9x9x9 voxel volume units. The pulmonary vessel volume is calculated by dividing the total lung vessel volume by the total lung volume and multiplying by 100. The caliber of vessels in the spared right lung are increased when compared to vessels within areas of fibrosis.

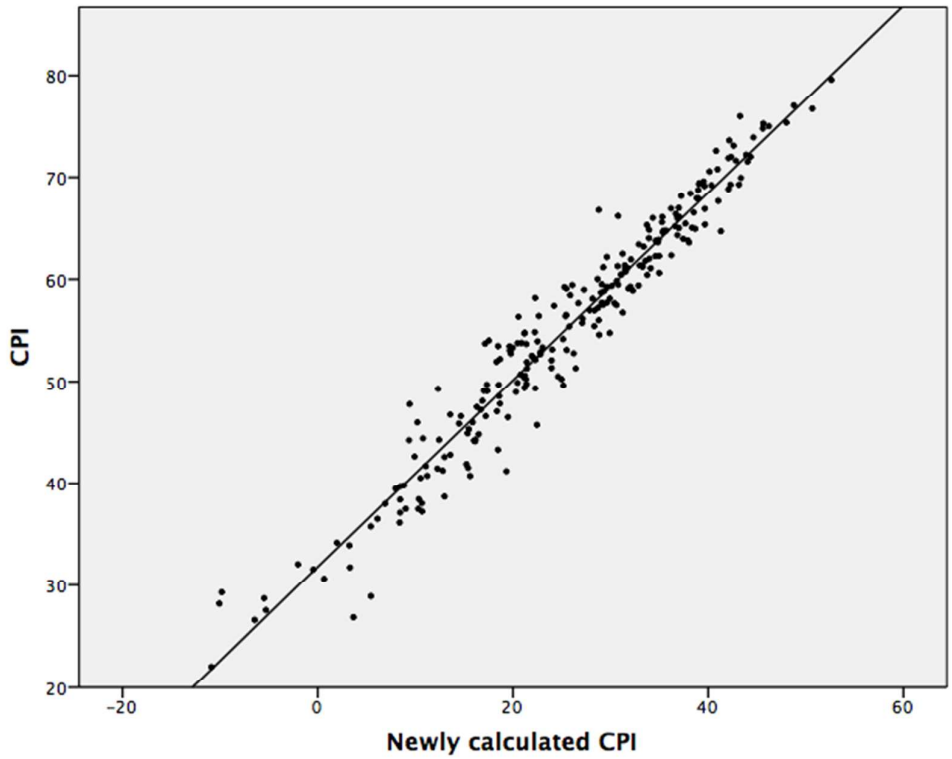


Figure 4. Graph demonstrating the relationship between CPI derived from correlations between pulmonary function tests and ILD extent scored visually in a separate, historic cohort of IPF patients, and the CPI calculated using a new formula derived from correlations between pulmonary function tests and ILD extent scored by CALIPER in the current cohort (R^2=0·95).

Variable (n = 283 unless stated) Units are percentage unless stated	Value
Median Age (years)	67
Male/female (ratio)	219/64
FEV1 % predicted (n=257)	70·8 ± 19·1
FVC % predicted (n=257)	68·8 ± 20·5
DLco % predicted (n=254)	36·1 ± 12·9
Kco % predicted (n=254)	69·0 ± 19·2
CPI (n=249)	55·1 ± 11·7
Echocardiography RVSP (mmHg) (n=150)	45·1 ± 16·8
CALIPER ILD extent	26·5 ± 18·1
CALIPER Ground glass opacity	17·0 ± 14·7
CALIPER Reticular pattern	8·5 ± 6·0
CALIPER Honeycombing	1·0 ± 1·7
CALIPER Grade 1 decreased attenuation	20·8 ± 20·7
CALIPER Grade 2 decreased attenuation	0·8 ± 2·6
CALIPER Grade 3 decreased attenuation	0·5 ± 2·8
CALIPER pulmonary vessel volume	5·1 ± 1·7
Visual ILD extent	43·1 ± 17·8
Visual Ground glass opacity	10·4 ± 11·4
Visual Reticular pattern	21·7 ± 10·9
Visual Honeycombing	9·8 ± 12·6
Visual Consolidation	1·1 ± 3·3
Visual Emphysema	4·7 ± 10·9
Visual TxBx severity (max score 18)	7·0 ± 3·3
Main pulmonary artery diameter (mm)	30·3 ± 4·8
Ascending aorta diameter (mm)	34·8 ± 4·2

Table 1. Patient age, gender and mean and standard deviations of pulmonary function tests, CALIPER and visually scored CT parameters and echocardiography data. Data represent mean values with standard deviations. CT = computed tomography, FEV1 = forced expiratory volume in one second, FVC = forced vital capacity, DLco = diffusing capacity for carbon monoxide, Kco = carbon monoxide transfer coefficient, TLC = total lung capacity, CPI = composite physiological index, ILD = interstitial lung disease, RVSP = right ventricular systolic pressure, TxBx = traction bronchiectasis.

Visual CT Variable (n = 283)	Single determination standard deviation
CT Interstitial lung disease extent	7·24
CT Ground glass opacity	6·15
CT Reticular pattern	5·24
CT Honeycombing	7·88
CT Consolidation	2·69
CT Total emphysema	4·99
CT Mosaic attenuation	3·83
CT Traction bronchiectasis severity	1·43

Table 2. Single determination standard deviation values of visual CT scores for idiopathic pulmonary fibrosis cases. CT = computed tomography.

	CT Variable	FEV1	FVC	DLco	Kco	CPI
ILD	Visual	0.19, <0.0001	0.27, <0.0001	0.35, <0.0001	0.07, <0.0001	0.44, <0.0001
	CALIPER	0.29, <0.0001	0.41, <0.0001	0.31, <0.0001	NS	0.48, <0.0001
Emphysema	Visual	NS	0.13, <0.0001	0.05, <0.0001	0.26, <0.0001	NS
	CALIPER	NS	0.06, <0.0001	NS	0.08, <0.0001	0.03, 0.004
Vessels	Visual	NS	NS	0.03, 0.008	0.04, 0.001	NS
	CALIPER	0.31, <0.0001	0.45, <0.0001	0.34, <0.0001	NS	0.53, <0.0001

Table 3. Univariate linear regression demonstrating relationships between disease in three compartments characterized by visual and CALIPER-derived scores and pulmonary function tests. CT=Computed tomography, ILD=interstitial lung disease, FEV1=forced expiratory volume in one second, FVC=forced vital capacity, DLco=diffusing capacity for carbon monoxide, Kco = carbon monoxide transfer coefficient, CPI=composite physiological index, NS=not significant.

	Pulmonary function test	CT Pattern	Beta Coefficient	95% Confidence Interval	P value	Model R^2
Model 1	FEV1	CALIPER Emphysema	-0·97	-1·49, -0·46	0·0002	0·33
		CALIPER ILD extent	-0·59	-0·70, -0·48	<0·0001	
		Visual Emphysema	0.33	0.09, 0.57	0.008	
	FVC	CALIPER ILD extent	-0·67	-0·78, -0·57	<0·0001	0·46
		Visual Emphysema	0·40	0·23, 0·57	<0·0001	
	DLco	CALIPER ILD extent	-0·28	-0·37, -0·18	<0·0001	0·51
		Visual Emphysema	-0·40	-0·51, -0·30	<0·0001	
		Visual ILD extent	-0·28	-0·37, -0·19	<0·0001	
	CPI	CALIPER ILD extent	0·30	0·22, 0·39	<0·0001	0·54
		Visual ILD extent	0·23	0·15, 0·31	<0·0001	
Model 2	FEV1	CALIPER Emphysema	-0·63	-1·01, -0·25	0·001	0·34
		CALIPER PVV	-6·85	-8·03, -5·66	<0·0001	
	FVC	CALIPER PVV	-7·59	-8·72, -6·47	<0·0001	0.48
		Visual Emphysema	0·33	0·16, 0·50	0·0001	
	DLco	CALIPER PVV	-3·72	-4·77, -2.67	<0·0001	0·53
		Visual Emphysema	-0·44	-0·54, -0·34	<0·0001	
		Visual ILD extent	-0·22	-0·31, -0·12	<0·0001	
	CPI	CALIPER PVV	3·80	2.91, 4·70	<0·0001	0·56
		Visual ILD extent	0·18	0·10, 0·26	<0·0001	

Table 4. Multivariate linear regression demonstrating relationships significant to a level of 0·01 between parenchymal patterns characterized by both visual and CALIPER scores with pulmonary function tests. Model 1 contained CALIPER ILD extent but excluded PVV. Model 2 contained PVV but excluded CALIPER ILD extent. CT= Computed tomography, FEV1=forced expiratory volume in one second, FVC=forced vital capacity, DLco=diffusing capacity for carbon monoxide, CPI=composite physiological index, ILD=interstitial lung disease, PVV= pulmonary vessel volume.