European Radiology
https://doi.org/10.1007/s00330-019-06619-5

CHEST

Stratification of long-term outcome in stable idiopathic pulmonary fibrosis by combining longitudinal computed tomography and forced vital capacity

Nicola Sverzellati[1] · Mario Silva[1] · Valeria Seletti[1] · Carlotta Galeone[2] · Stefano Palmucci[3] · Sara Piciucchi[4] · Carlo Vancheri[5] · Venerino Poletti[6] · Sara Tomassetti[7] · Ronald Karwoski[8] · Brian J. Bartholmai[9]

Received: 29 May 2019 / Revised: 28 October 2019 / Accepted: 11 December 2019

Abstract

Objectives To test HRCT with either visual or quantitative analysis in both short-term and long-term follow-up of stable IPF against long-term (transplant-free) survival, beyond 2 years of disease stability.

Methods Fifty-eight IPF patients had FVC measurements and HRCTs at baseline (HRCT0), 10–14 months (HRCT1) and 22–26 months (HRCT2). Visual scoring, CALIPER quantitative analysis of HRCT measures, and their deltas were evaluated against combined all-cause mortality and lung transplantation by adjusted Cox proportional hazard models at each time interval.

Results At HRCT1, a ≥ 20% relative increase in CALIPER-total lung fibrosis yielded the highest radiological association with outcome (C-statistic 0.62). Moreover, the model combining FVC% drop ≥ 10% and ≥ 20% relative increase of CALIPER-total lung fibrosis improved the stratification of outcome (C-statistic 0.69, high-risk category HR 12.1; landmark analysis at HRCT1 C-statistic 0.66, HR 14.9 and at HRCT2 C-statistic 0.61, HR 21.8). Likewise, at HRCT2, the model combining FVC% decrease trend and ≥ 20% relative increase of CALIPER-pulmonary vessel–related volume (VRS) improved the stratification of outcome (C-statistic 0.65, HR 11.0; landmark analysis at HRCT1 C-statistic 0.62, HR 13.8 and at HRCT2 C-statistic 0.58, HR 12.6). A less robust stratification of outcome distinction was also demonstrated with the categorical visual scoring of disease change.

Conclusions Annual combined CALIPER -FVC changes showed the greatest stratification of long-term outcome in stable IPF patients, beyond 2 years.

Key Points

- *Longitudinal high-resolution computed tomography (HRCT) data is more helpful than baseline HRCT alone for stratification of long-term outcome in IPF.*
- *HRCT changes by visual or quantitative analysis can be added with benefit to the current spirometric reference standard to improve stratification of long-term outcome in IPF.*
- *HRCT follow-up at 12–14 months is more helpful than HRCT follow-up at 23–26 months in clinically stable subjects with IPF.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06619-5) contains supplementary material, which is available to authorized users.

✉ Nicola Sverzellati
nicola.sverzellati@unipr.it

[1] Scienze Radiologiche, Department of Medicine and Surgery, University of Parma, Padiglione Barbieri, V. Gramsci 14, Parma, Italy

[2] Department of Clinical Sciences and Community, University of Milan, Milan, Italy

[3] Department of Medical and Surgical Sciences and Advanced Technologies-Radiodiagnostic and Radiotherapy Unit, University Hospital "Policlinico-Vittorio Emanuele", Catania, Italy

[4] Department of Radiology, Azienda USL Romagna, Ospedale GB Morgagni, Forlì, Italy

[5] Department of Clinical and Experimental Medicine, University of Catania, Catania, Italy

[6] Department of Respiratory Diseases and Allergy, Aarhus University Hospital, Aarhus, Denmark

[7] Department of Diseases of the Thorax, Ospedale G.B. Morgagni, Forli, Italy

[8] Biomedical Imaging Resource, Mayo Clinic, Rochester, MN, USA

[9] Division of Radiology, Mayo Clinic, Rochester, MN, USA

Published online: 31 January 2020

Keywords Idiopathic pulmonary fibrosis · Multidetector computed tomography · Usual interstitial pneumonia · Outcome · Machine learning

Abbreviations

CALIPER	Computer-Aided Lung Informatics for Pathology Evaluation and Rating
CI	Confidence interval
CT	Computed tomography
DLco	Diffusing capacity for carbon monoxide
FEV1	Forced expiratory volume in one second
FVC	Forced vital capacity
GAP	Gender-age-physiology
HR	Hazard ratio
HRCT	High-resolution computed tomography
ILD	Interstitial lung disease
IPF	Idiopathic pulmonary fibrosis
PFT	Pulmonary function test
VRS	Pulmonary vessel–related volume

Introduction

Idiopathic pulmonary fibrosis (IPF) is a primary interstitial lung disease with exceptionally poor prognosis due to rapidly progressive evolution towards pulmonary failure and comorbidities [1]. However, marked variations in disease behavior can occur in IPF. Median survival is generally quoted at 3 years [2, 3], but some patients may survive up to 5–10 years [4]. Disease stage at the time of diagnosis and longitudinal behavior both account for the prognostic variability, and these features drive clinical treatment strategy [5–7].

Several demographic-functional characteristics of subjects with IPF are associated with increased mortality, particularly older age. Such characteristics have been incorporated into multidimensional indexes for personalized prognostication [8–15]. However, the repeatability and sensitivity limitations of pulmonary function tests (PFT) illuminate the need for more detailed measures of disease severity such as high-resolution computed tomography (HRCT), which allows to capture the heterogeneity of disease extent and character throughout different lung regions [16].

Several studies have shown that disease extent at baseline HRCT either by computer-based quantification or visual scoring [6, 17–19] may increase the accuracy of multivariate prediction models. The utility of HRCT metrics on longitudinal scans at multiple time points is also proven [20–24]. However, there is scarce evidence upon the additional value of serial HRCT over forced vital capacity (FVC), which still represents the reference endpoint in both clinical practice and scientific setting. In particular, there is neither consensus recommendation for the most robust HRCT scoring technique nor validation of standardized interval for HRCT scan in patients without acute exacerbation [17, 18, 20, 25]. Detailed investigation of these issues is foremost because spirometric follow-up by FVC may be biased by several factors [26] and might be not sufficiently sensitive in capturing subtle changes in relatively stable or insidious conditions [16, 27].

The purpose of this study was to test HRCT with either visual or quantitative analysis in both short-term and long-term follow-up of IPF against long-term (transplant-free) survival, beyond 2 years of disease stability.

Methods

Patients with diagnosis of IPF between December 2011 and October 2014 according to published guidelines were retrospectively retrieved from the electronic registers of two Italian referral centers (Ospedale Morgagni di Forlì and Policlinico di Catania) [1]. Institutional review board approval was obtained from both centers.

Selection criteria were as follows: (a) three serial volumetric HRCTs at baseline, 10–14 month and 22–26 months (hereafter HRCT0, HRCT1, and HRCT2, respectively); (b) three serial full PFTs within 30 days of the selected HRCT time point; (c) systematic collection of full clinical history; (d) absence of acute complications (e.g., acute exacerbation, pulmonary edema, embolism, or infection) or lung cancer within HRCT0-HRCT2 (Fig. 1).

Data collection was closed in March 2017; data censoring became as a consequence of the study time frame.

Pulmonary function testing

Pulmonary function testing was done locally at study sites (supplementary material). For the study objectives, the Gender-Age-Physiology (GAP) stage at baseline [14] was recorded and serial %-predicted FVC and %-predicted DLCO were calculated at each time point.

Visual scoring of HRCT scans

Each HRCT scan was evaluated independently by two radiologists with 14 and 5 years of experience in thoracic imaging. The HRCT time points were read in a random order. The radiologists were blinded to clinical information.

HRCTs were scored on a lobar basis using a continuous scale [6, 28]. The total extent of interstitial lung disease (ILD) was estimated to the nearest 5% and then was subclassified into relative lobar extent of each from the following four

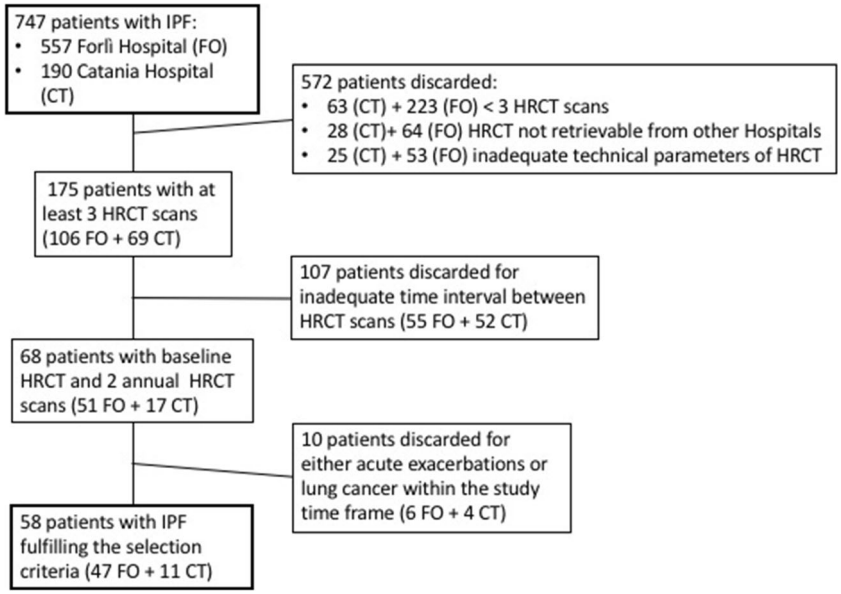

Fig. 1 STROBE diagram illustrating the selection of idiopathic pulmonary fibrosis (IPF) patients for the study. HRCT, high-resolution computed tomography; FO, Forlì cases; CT, Catania cases

parenchymal patterns: reticular, ground glass opacification, honeycombing, and consolidation (supplementary material).

Traction bronchiectasis [29] was assigned with a categorical "severity" score accounting for the average degree of airway dilatation within the areas of fibrosis as well as the extent of dilatation throughout the lobe. An ordinal Gestalt score was given as follows: none (0), mild (1), moderate (2), and severe (3) [6, 30].

Longitudinal differences in overall extent were calculated from the baseline HRCT (e.g., HRCT1-HRCT0 and HRCT2-HRCT0) and represented the absolute change. The relative change was calculated as the ratio between the absolute change divided by the baseline extent.

Following the above-described visual scoring in a random fashion, serial HRCTs were also assessed in side-by-side comparison to grant fine description of overall and relative extent of each parenchymal finding. This side-by-side comparison was summarized into nominal categories of disease behavior, as follows: (a) progression (e.g., increase of ILD extent in at least one lobe); (b) stability; (c) improvement (e.g., decrease in ILD extent in at least one lobe, with no disease progression in the remainders). Consensus formulation for visual scores is outlined in the supplementary material.

CALIPER evaluation

HRCTs were evaluated by CALIPER as previously described [28]. Patterns scored volumetrically by CALIPER included ground glass opacities, reticular pattern, honeycombing, emphysema, normal lung, and vessel-related structures (VRS) (Fig. 2). The total extent of lung fibrosis was represented by the sum of ground glass opacities, reticular pattern, and honeycombing.

Longitudinal differences were calculated as for the visual scores. Further details of HRCT technique and quantitative analysis are given in the supplementary material.

Statistical analysis

Interobserver agreement was evaluated with either the intraclass correlation coefficient or the kappa statistics, as appropriate [31].

Details on the change in both FVC and CALIPER quantification are given in the supplementary material. Relationship between visual score, CALIPER score, and FVC% of predicted or DLco at each HRCT time point was tested by Pearson correlation coefficients.

The study outcome was defined by any of all-cause death or lung transplantation beyond 2 years since HRCT0. The outcome was tested against all variables, independently at each time point, as reported thereafter.

HRCT0 Cox proportional hazard models (both univariate and multivariate) assessed associations between baseline variables (e.g., visual, CALIPER scores, age, gender, FVC %-predicted, DLCO %-predicted, GAP score) and the study outcome. In multivariate models, stepwise proportional hazards analyses were performed, if not otherwise specified.

HRCT1 and HRCT2 Cox proportional hazard models at both HRCT1 and HRCT2 were adjusted for both GAP score and pack-years of smoking. We tested the proportional hazards assumption by including time-dependent effects in the model (i.e., a covariate for interaction of the predictor and the logarithm of survival time), and no violation was found. The index of concordance (C-statistic) was used to compare model discrimination between various models.

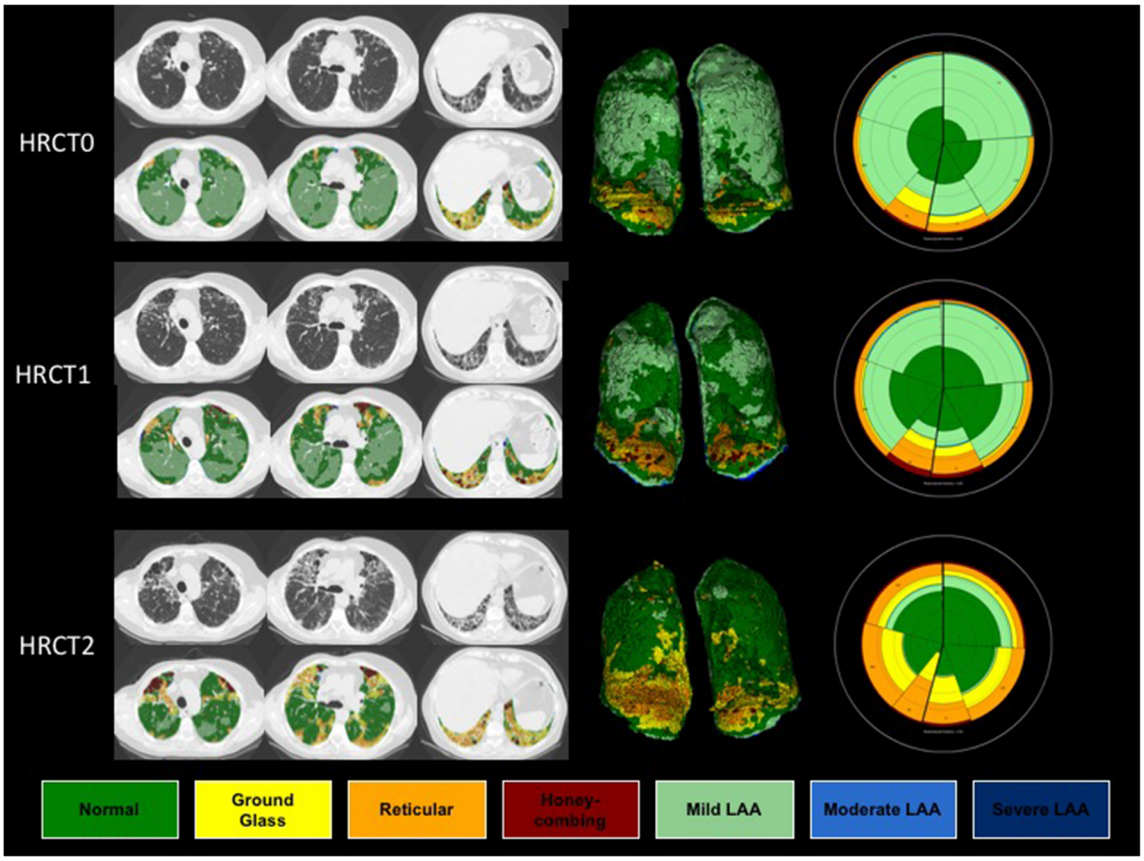

Fig. 2 Axial HRCTs at the study time points of a 67-year-old male former smoker. CALIPER-HRCT images demonstrate increase in reticulation, progression of volume loss in the lung bases, and increase in the extent of disease in the mid to upper lungs across HRCT1-HRCT2. Honeycombing slightly progressed as shown in a tiny rim on the glyphs

Event-free survival (EFS) was analyzed using the Kaplan-Meier product-limit survival curve estimates [32]. EFS was defined as the time from HRCT0 until outcome and by two further analyses with landmark at HRCT1 or HRCT2. In the figures, follow-up was cut at 5 years. Kaplan-Meier plots showed EFS for models of FVC decline with or without combined visual or CALIPER changes below and above predefined thresholds considered significant for disease progression at both HRCT1 and HRCT2. Thresholds were selected using the receiver operating characteristic (ROC) curve analysis [33].

Statistical analyses were performed using SAS 9.4 (SAS Institute) and the figures were obtained using STATA 15.0 (StataCorp LP) statistical software.

Results

Baseline data

The study population consisted of 58 subjects (Fig. 1). Diagnosis was corroborated by either surgical biopsy or cryobiopsy in 30/58 (51.7%) subjects. A total of 38/58 (65.5%) subjects were treated with either pirfenidone or nintedanib, at some point. The mean age at presentation was 66.4 years ± 8.2, with 69% ($n = 40$) male and 67.2% ($n = 39$) former or current smokers (Table 1).

The median follow-up of the study was 44 months (quartile range, 36–57). Median HRCT1-HRCT0 time interval was 12 months (quartile range, 12–14), whereas median HRCT2-HRCT0 was 25 months (quartile range, 23–26). After 2 years since HRCT0, 2 (3.4%) patients underwent lung transplant and 15 (25.9%) died.

Interobserver variation for the visual scores and HRCT-functional correlations are given in the supplementary material (Supplementary Tables 1–2).

Baseline indexes

Most of visual and CALIPER scores at HRCT0 were associated with the study outcome (Table 2 and Supplementary Table 3). The model with GAP score (HR 1.47, 95%CI 1.04–2.08) and pack-years of smoking (HR 1.01, 95%CI 0.97–1.04) showed the best discrimination at baseline (C-statistic 0.712). The inclusion of either HRCT0 visual score (C-statistic 0.596) or HRCT0 CALIPER metrics (C-statistic

Table 1 Demographics, PFT, visual score CALIPER metrics, and selected follow-up information of the 58 IPF patients

	Mean	SD	Median	Quartile range*
Demographics and pulmonary function test				
Male (*N*, %)	40	69.0%	–	–
Age at diagnosis (years)	66.4	8.2	67.0	61.0–72.0
Pack-years of smoking	15.5	14.2	20.0	0–22.0
FVC %pred	80.1	18.0	77.5	67.0–92.0
DLco %pred	57.8	16.5	55.0	45.0–69.0
GAP score (max score 8)	3.12	1.52	3.0	2.0–4.0
Stage I (*N*, %)	36	63.2%	–	–
Stage II (*N*, %)	19	33.3%	–	–
Stage III (*N*, %)	2	3.5%	–	–
Visual score (%)				
Total ILD extent	48.7	16.5	48.8	39.6–60.0
Ground glass opacity	6.2	11.3	0.48	0–7.63
Reticular pattern	34.5	17.7	36.6	17.6–49.1
Honeycombing	7.80	12.7	1.81	0–9.04
Consolidation	0.17	0.81	0	0–0
Total emphysema	2.43	5.50	0	0–2.92
Mosaicism	1.49	4.19	0	0–1.25
TxBx severity	6.15	2.89	6.0	4.5–8.0
CALIPER metrics (%)				
Total lung fibrosis extent	17.86	13.46	13.82	8.50–27.91
Ground glass opacity	8.39	10.30	5.20	1.89–11.86
Reticular pattern	8.02	6.21	5.59	4.18–10.58
Honeycombing	1.44	2.77	0.50	0.23–1.63
Emphysema	2.00	3.68	0.71	0.25–1.96
Non-specific air-trapping	32.03	21.87	29.39	9.59–50.75
VRS	4.18	1.69	3.89	3.00–5.13
Follow-up information				
Follow-up (months)	47.0	14.7	43.8	35.8–57.0
Event-free survival (months)	62.5	29.2	68.4	45.7–73.7

FVC%pred, forced vital capacity percentage of predicted; *DLco%pred*, diffusing capacity of the lung for carbon monoxide percentage of predicted; *GAP*, gender age physiology; *ILD*, interstitial lung disease; *TxBx*, traction bronchiectasis; *VRS*, pulmonary vessel–related volume

*Quartile range was presented as the 25th percentile and 75th percentile

0.641) did not refine the stratification by baseline GAP score and pack-years.

HRCT1-HRCT0 changes

At HRCT1, visual categories of disease behavior (progression vs. stability/improvement; HR 7.46; 95%CI, 2.39–23.19) and continuous VRS change (HR 1.55; 95%CI, 0.91–2.6) were the HRCT metrics that best associated with the long-term outcome (adjusted for both baseline GAP score and pack-years of smoking, Table 3).

Table 2 Multivariate Cox regression analyses of the study outcome (death or lung transplant) according to baseline indexes

	HR*	95% CI	HR**	95% CI
Visual scoring				
Reticular pattern	1.054	1.020–1.090	1.05	1.01–1.09
GAP score	–	–	1.28	0.90–1.82
Pack-years of smoking	–	–	1.00	0.95–1.04
C-statistic	0.661		0.596	
CALIPER metrics				
Reticular pattern	1.046	0.956–1.145	1.06	0.96–1.16
VRS	1.469	1.026–2.102	1.40	0.91–2.15
GAP score	–	–	1.00	0.65–1.55
Pack-years of smoking	–	–	1.01	0.97–1.05
C-statistic	0.673		0.641	
GAP score	–	–	1.47	1.04–2.08
Pack-years of smoking	–	–	1.01	0.97–1.04
C-statistic			0.712	

HR, hazard ratio; *CI*, confidence interval; *GAP*, gender age physiology; *ILD*, interstitial lung disease; *VRS*, pulmonary vessel–related volume

*HR for continuous increase of one unit mutually adjusted for all variables (selected according to stepwise proportional hazards analyses) reported in each subgroup of analyses types

**Cox regression analyses in each subgroup of analysis types included also adjustment for GAP score and pack-years of smoking

The relative increase in CALIPER-total lung fibrosis volume ≥ 20% was the best threshold for the stratification of outcome, according to the ROC analysis. The C-statistic of relative increase in CALIPER-total lung fibrosis volume ≥ 20% showed the best association (C-statistic 0.62; HR 2.89; 95%CI, 1.13–8.37) compared with either visual scoring of disease progression (C-statistic 0.52; HR 7.46; 95%CI, 2.39–23.19) or FVC% decline ≥ 10% (C-statistic 0.58; HR 9.45; 95%CI, 3.14–30.87).

The model combining FVC% drop ≥ 10% and relative increase of CALIPER-total lung fibrosis ≥ 20% improved the stratification of outcome (C-statistic 0.69) by definition of 3 risk categories, notably with the "high-risk" category showing HR 12.1 (95%CI, 3.1–46.7). The landmark analysis at HRCT1 substantially confirmed the stratification of outcome (Fig. 3), and the HR for the high-risk category was 14.9 (95%CI 3.9–56.8; C-statistic 0.66).

Likewise, the combination HRCT1-HRCT0 visual categories of disease behavior to FVC% drop ≥ 10% showed improved discrimination of long-term outcome compared with either metric alone (C-statistic 0.64). The landmark analysis at HRCT1 substantially confirmed the stratification of outcome, as detailed in Fig. 4b (C-statistic 0.60).

Similar findings were obtained by including acute exacerbation into the study outcome ("Post hoc analysis" in supplementary material).

Table 3 Multivariate Cox regression analyses* of the study outcome (death or lung transplant) according to HRCT1-HRCT0 and HRCT2-HRCT0 changes

	HRCT1-HRCT0		HRCT2-HRCT0	
	HR	95% CI	HR	95% CI
Pulmonary function test				
FVC%pred drop > 10%	9.85	3.14–30.88		
DLCO%pred drop > 15%	4.37	1.32–14.46		
FVC%pred trend decrease during all study period		3.87 1.16–12.93		
Visual score (%)				
Total ILD extent	1.375	0.943–2.006	1.294	1.070–1.566
Ground glass opacity	1.068	0.978–1.166	0.981	0.895–1.075
Reticular pattern	1.137	1.017–1.270	1.115	1.042–1.194
Honeycombing	0.996	0.863–1.149	0.966	0.885–1.055
Consolidation	0.044	0.001–1.667	0.317	0.001–174.99
Total emphysema	1.221	0.865–1.723	1.184	0.984–1.422
Mosaicism	0.597	0.324–1.098	0.830	0.453–1.519
TxBx severity	1.281	0.932–2.721	1.199	0.972–2.331
Disease behavior change				
Stability/improvement	1	Reference	1	Reference
Progression	7.456	2.396–23.198	2.264	0.671–7.639
CALIPER metrics (%)				
Total lung fibrosis extent	1.057	1.006–1.111	1.075	1.031–1.121
Ground glass opacity	1.040	0.980–1.105	1.068	1.028–1.110
Reticular pattern	1.006	0.971–1.042	0.985	0.906–1.071
Honeycombing	1.637	0.719–3.728	0.960	0.734–1.257
Emphysema	1.266	0.818–1.959	0.956	0.797–1.148
Non-specific air-trapping	0.977	0.932–1.025	0.985	0.944–1.027
VRS	1.549	0.908–2.644	0.987	0.946–1.028

HRCT, high-resolution chest computed tomography; *HR*, hazard ratio; *CI*, confidence interval; *FVC%pred*, forced vital capacity percentage of predicted; *DLco%pred*, diffusing capacity of the lung for carbon monoxide percentage of predicted; *GAP*, gender age physiology; *ILD*, interstitial lung disease; *TxBx*, traction bronchiectasis; *VRS*, pulmonary vessel–related volume

*HR for continuous increase of one unit, if not otherwise specified, adjusted for GAP score at baseline and pack-years of smoking

HRCT2-HRCT0 changes

The FVC% decrease trend until HRCT2 showed lower association with outcome (HR 3.87; 95%CI, 1.16–12.93) than that of FVC% drop ≥ 10% at HRCT1 (HR 9.85; 95%CI, 3.14–30.87), after adjusting for both baseline GAP score and pack-years of smoking (Table 3).

The relative increase in CALIPER-VRS ≥ 20% from baseline to HRCT2 was the best threshold for stratification of outcome, according to the ROC analysis. The relative increase in CALIPER-VRS ≥ 20% showed the best association (C-statistic 0.59; HR 5.65; 95%CI, 1.69–18.87) compared with either visual scoring of disease progression (C-statistic 0.52; HR 2.26; 95%CI, 0.67–7.64) or FVC% decrease trend during the study period (C-statistic 0.45; HR 3.87; 95%CI, 1.16–12.93). The model combining FVC% decrease trend until HRCT2 and relative increase of CALIPER-VRS ≥ 20% improved the stratification of outcome (C-statistic 0.65). The landmark analysis at HRCT2 substantially confirmed the stratification of outcome, as detailed in Fig. 4a (C-statistic 0.58).

Likewise, combining HRCT2-HRCT0 visual categories of disease behavior with FVC% trend decline showed improved discrimination of long-term outcome compared with either metric alone (C-statistic 0.60). The landmark analysis at HRCT1 and HRCT2 substantially confirmed the stratification of outcome, as detailed in Fig. 4b (C-statistic 0.59).

Similar findings were obtained by including acute exacerbation into the study outcome ("Post hoc analysis" in supplementary material).

Discussion

The role of HRCT in the follow-up of subjects of IPF without acute exacerbation is still under debate. The study results suggest that serial HRCT data yield potentially better stratification of

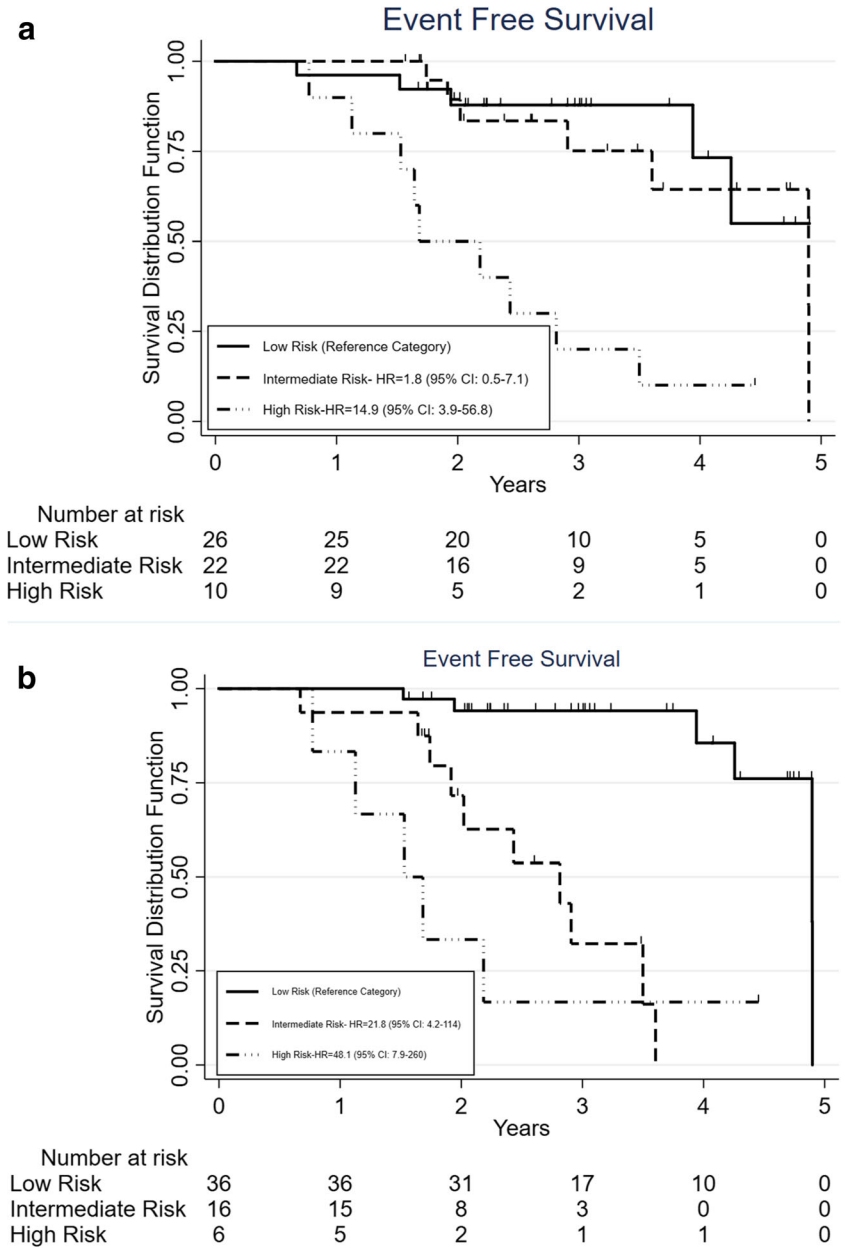

Fig. 3 **a** Event-free survival calculated in landmark analysis beginning at HRCT1 after the enrolment for patients in strata of FVC%pred drop ≥ 10% and CALIPER-total lung fibrosis relative increase ≥ 20% at HRCT1. Study outcome: death or lung transplant. (C-statistic 0.66) FVC%pred, forced vital capacity percentage of predicted; ILD, interstitial lung disease. Risk categories were defined as follows: Low risk, no FVC% drop ≥ 10% and no total fibrosis extent relative increase ≥ 20% at HRCT1. Intermediate risk, FVC% drop ≥ 10% or total fibrosis extent relative increase ≥ 20% at HRCT1. High risk, FVC%pred drop ≥ 10% and total fibrosis extent relative increase ≥ 20% at HRCT1. Hazard ratios (HR) and corresponding 95% confidence intervals (95% CI) were derived using Cox regression model adjusted for GAP score and pack-years of smoking. **b** Event-free survival calculated in landmark analysis beginning at HRCT1 after the enrolment for patients in strata of FVC%pred drop ≥ 10% and visually assessed lLD behavior at HRCT1. Study outcome: death or lung transplant. (C-statistic 0.60) FVC%pred, forced vital capacity percentage of predicted; ILD, interstitial lung disease. Risk categories were defined as follows: Low risk, no FVC%pred ≥ 10% and no visual disease progression at HRCT1. Intermediate risk, FVC%pred drop ≥ 10% or visual disease progression at HRCT1. High risk, FVC%pred drop and visual disease progression at HRCT1. Hazard ratios (HR) and corresponding 95% confidence intervals (95% CI) were derived using Cox regression model adjusted for GAP score and pack-years of smoking

outcome beyond 2 years of disease stability, in addition to baseline GAP. In particular, longitudinal HRCT changes in volume of lung fibrosis at 12–14 months may augment the stratification of outcome beyond the lone decline in FVC.

The narrowed study inclusion criteria are of particular interest because subjects with IPF in early-/mid-term relatively stable clinical conditions represent a sizeable group of IPF patients in the era of anti-fibrotic treatment. Indeed, the study cohort was

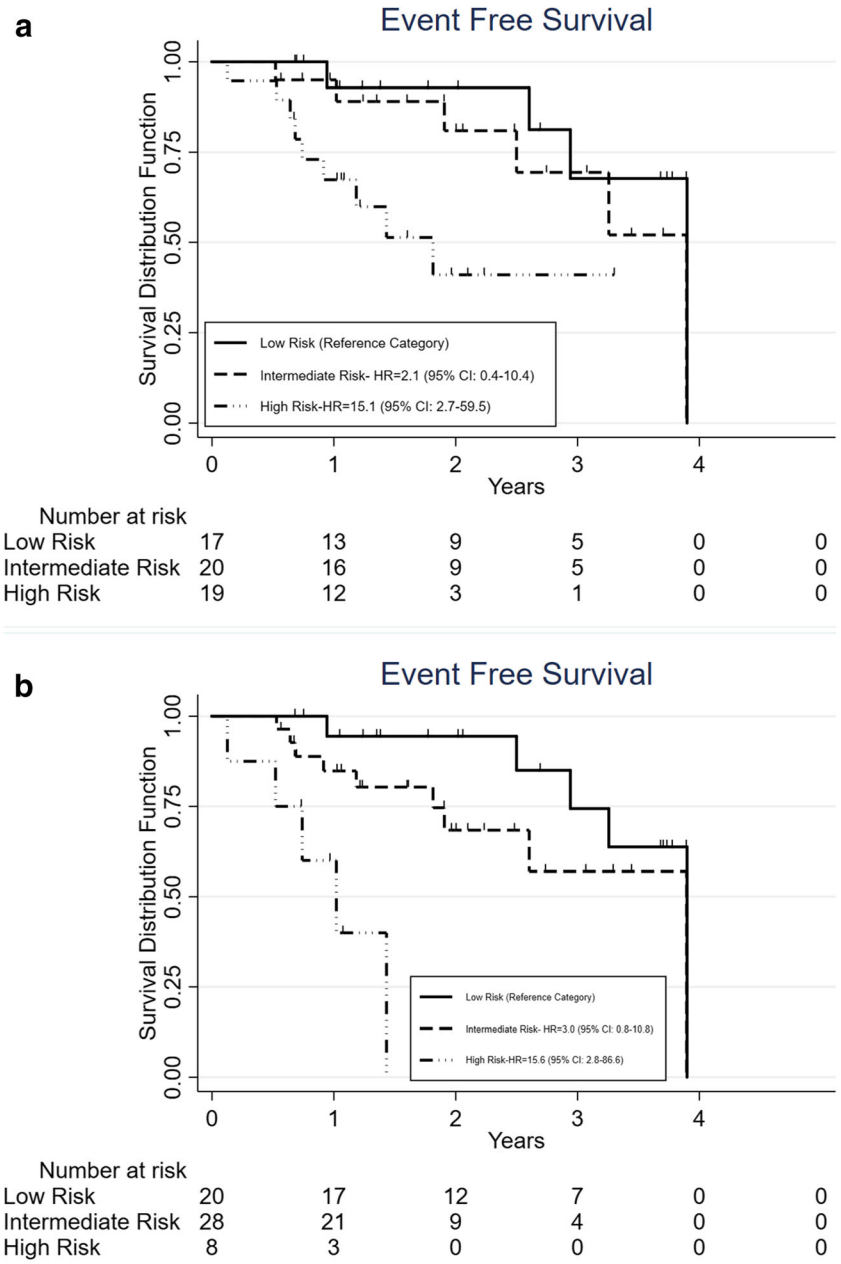

Fig. 4 **a** Event-free survival calculated in landmark analysis beginning at HRCT2 after the enrolment for patients in strata of FVC%pred trend and VRS relative increase ≥ 20% at HRCT2. Study outcome, death or lung transplant. (C-statistic 0.58). FVC%pred, forced vital capacity percentage of predicted; VRS, pulmonary vessel related volume. Risk categories were defined as follows: Low risk, no FVC%pred trend increase and no VRS relative increase ≥ 20% at HRCT2. Intermediate risk, FVC% trend increase or VRS relative increase ≥ 20% at HRCT2. High risk, patients with FVC% trend increase and VRS relative increase ≥ 20% at HRCT2. Hazard ratios (HR) and corresponding 95% confidence intervals (95% CI) were derived using Cox regression model adjusted for GAP score and pack-years of smoking. **b** Event-free survival calculated in landmark analysis beginning at HRCT2 after the enrolment for patients in strata of FVC%pred trend and visually assessed disease behavior at HRCT2. Study outcome, death or lung transplant. (C-statistic 0.59) FVC%pred, forced vital capacity percentage of predicted. Risk categories were defined as follows: Low risk, no FVC%pred trend increase and no visual disease progression at HRCT2. Intermediate risk, FVC%pred trend increase or visual disease progression at HRCT2. High risk, FVC% trend increase and visual disease progression at HRCT2. Hazard ratios (HR) and corresponding 95% confidence intervals (95% CI) were derived using Cox regression model adjusted for GAP score and pack-years of smoking

mostly constituted by subjects with mild to moderate disease (as defined by GAP staging), including anti-fibrotic treatment, without major complications through 2 years. In this context, clinico-functional changes are often subtle and prediction of prognosis may be extremely difficult. While GAP is clinically engraved and tightly associated with IPF outcome, these data show that longitudinal HRCT data can add potentially relevant information for both clinical practice and drug trials.

At baseline, no HRCT index substantially improved the stratification of outcome by GAP score and pack-years of smoking. This observation is in contrast with those reported by Jacob who showed that the GAP staging system was marginally weaker in predicting mortality than combined CALIPER-VRS and honeycombing, at baseline [6]. Several factors may explain these differences, as discussed in the supplementary material. Of note, the result of the present study should be deemed relevant in the group of subjects with prolonged stability that are mostly represented in the non-advanced GAP stages.

Discrepancy between physiology and morphology is well known in follow-up of ILD [34, 35]. In keeping with prior studies, the change in the total extent of ILD as assessed by quantitative analysis or visual scoring poorly or moderately correlated with changes in FVC [18, 22, 36]. This complex structure/function relationship suggests that complementary information is expected by imaging.

In keeping with the study by Flaherty [5], we observed that the association of outcome and FVC decline at 12–14 months outstanded the above changes in any HRCT score. Nevertheless, our study models show that either imaging domain—visual or CALIPER score—can be gathered to physiology for refined stratification of IPF outcome. The study models entail both discordance and concordance between critical functional and HRCT scores that can be frequently encountered in clinical practice. Notably, in the absence of FVC and HRCT changes, the percentage of subjects who either died or underwent lung transplantation was lower (low risk) than that recorded when either one was present (intermediate risk). Furthermore, the likelihood of meeting the study outcome was substantially higher when both FVC and CALIPER inputs were present (high risk). Such a stratification was clear-cut using either a side-by-side visual comparison between HRCT time points or relative change in the CALIPER score. The moderate to good levels of interobserver agreement for the visual comparison of the HRCT scans are in contrast with the worse agreement reported by prior studies [37]. Indeed, the utility of the software in the follow-up of IPF seems even more obvious than at baseline.

In keeping with Jacob's studies, the longitudinal change in VRS was an optimal biomarker of "silent" disease progression [38, 39]. The VRS measure not only predominantly quantifies pulmonary arteries and veins but also captures connected tubular structures mainly representing adjoining regions of fibrosis. In fact, major collinearity of VRS with the CALIPER-total volume of lung fibrosis was recorded in the multivariate models at both baseline and longitudinal analyses. However, the selected variables and thresholds were designated by ROC analysis for stratification of outcome: we found that the relative change of total volume of lung fibrosis best fits the model at HRCT1, whereas relative change of VRS did it at HRCT2, maybe reflecting a relatively minor progressive asymptomatic accumulation hardly detected by visual assessment. Of note, at HRCT2, VRS in isolation was superior to the FVC trend in predicting the study outcome.

In a population of early IPF patients, the evaluation of the longer follow-up was considered appropriate since it was not fully clear when HRCT changes could become appreciable either visually or quantitatively. The study findings suggest that the shorter time interval at 12–14 months is more associated with the study outcome. Importantly, this was confirmed by landmark analysis further corroborated by adding subsequent acute exacerbations—an important clinical trial endpoint—to the study outcome (see supplementary material).

Most prior studies evaluating longitudinal HRCT data included only one follow-up HRCT scan and composite outcomes such as functional decline, hospitalization, and death [17, 18, 38, 40, 41]. Conversely, the present study applies composite stratification approach by radiology and functional decline because these are indeed the preclinical inputs before actual disease progression with hospitalization, transplant, or death. In keeping with the present study, Maldonado first suggested that short-term changes in either total volume of interstitial abnormalities or volume of reticular opacities by the CALIPER were predictive of survival in 55 subjects with IPF free of acute complications [20]. The radiologist interpretation of overall global change in terms of progression of ILD was also predictive. However, compared with the present study, the authors investigated a single follow-up time point with a substantially wider time interval starting at 3 months (range 3–15 months compared with 12–14 months in our study) and a fairly lower median follow-up time (28 months compared with 43 months in our study). Furthermore, Maldonado did not evaluate the interaction between FVC% and HRCT changes in stratification of outcome. Our findings are in keeping with those reported by Lee in a recent study of 144 subjects with IPF [21]. The change in the extent of fibrosis at 12-month follow-up CT scan (quantified by an in-house texture-based automated system) was significantly associated with survival along with age, desaturation, and the quantitative extent of lung fibrosis at baseline. Despite the technical differences between the computer-based methods, such consistency emphasizes the utility of the software as a surrogate endpoint for the longitudinal evaluation of IPF, especially in long-term stable subjects such as those under anti-fibrotic treatment. A major difference between our study and Lee's is represented by the clinical applicability. While Lee et al reported the association by single-unit increase (continuous fashion), we report a discrete threshold for the definition of nominal risk classes based on HRCT change $\geq 20\%$ relative extent of fibrosis by CALIPER at 12 months.

Although a change $\geq 20\%$ relative extent of fibrosis may seem a big increase at first sight, it is worth clarifying that such a relative change in a subject with, for example, 17.9% of lung

fibrosis at baseline (i.e., the mean extent in our study population as quantitated by the CALIPER) would correspond to an absolute increase of 3.6%. This figure is in keeping with that of Humpries who reported the absolute increase of 3.4% of lung fibrosis by a machine learning–based automated tool which represents a clinically important change [22].

This study has limitations, and the first was related to the observational study in terms of collection criteria and results that should be carefully confirmed before integration in clinical practice. Noteworthily, the retrospective fashion allows stratification of risk in terms of association rather than prediction. Second, the narrowed inclusion criteria greatly reduced the size of the study population: this selection was specifically intended in the foreseeable perspective of patients with early-stage IPF, both in clinical practice and in clinical trials. Also, it is worth mentioning that a sizeable proportion of subjects from one center could not be included because of non-volumetric HRCT that hampered the CALIPER analysis. The relatively small study population was an important limitation also for the multivariate models (as shown by the wide 95%CIs). However, we restricted the multivariate models to a few adjustment variables, according to the rule proposed by Harrell [42]. Nonetheless, the strict selection criteria were intended for the simulation of the prototypical early-diagnosed IPF patient who is expected in the next future of clinical practice and trial recruitment [39].

In conclusion, the simple combination of FVC% and HRCT changes shows the synergistic role that physiology and morphology play in IPF outcome. Noteworthily, both FVC% and HRCT changes are independently associated with IPF outcome, while their combination at 12–14 months refines stratification compared with either individual one. Prospective studies are fostered to test the reproducibility of the proposed radiological contribution to longitudinal IPF characterization.

Funding information This study has received funding by the European Society of Thoracic Imaging.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Nicola Sverzellati.

Conflict of interest The authors of this manuscript declare relationships with the following companies:

Dr. Sverzellati reports grants and personal fees from Roche, personal fees from Boehringer-Ingelheim, outside the submitted work.

Dr. Tomassetti reports grants and personal fees from Roche, personal fees from Boehringer-Ingelheim, outside the submitted work.

Dr. Palmucci reports personal fees as speaker (Roche and Boehringer Ingelheim); as writer (Boehringer Ingelheim).

Dr. Silva reports grants from European Society of Thoracic Imaging, during the conduct of the study.

Dr. Bartholmai reports personal fees from Promedior, LLC, and from Imbio, LLC, outside the submitted work. Mayo Clinic has received grants from NIH/NHLBI, fees from Imbio, LLC, and Boehringer Ingelheim outside the submitted work. In addition, Dr. Bartholmai has a patent SYSTEMS AND METHODS FOR ANALYZING IN VIVO TISSUE VOLUMES USING MEDICAL IMAGING pending to Mayo Clinic.

Mr. Karwoski reports other from Imbio Inc., outside the submitted work.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- prognostic study/observational
- multicenter study

References

1. Raghu G, Collard HR, Egan JJ et al (2011) An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. Am J Respir Crit Care Med 183:788–824
2. Daniil ZD, Gilchrist FC, Nicholson AG et al (1999) A histologic pattern of nonspecific interstitial pneumonia is associated with a better prognosis than usual interstitial pneumonia in patients with cryptogenic fibrosing alveolitis. Am J Respir Crit Care Med 160:899–905
3. Raghu G, Chen SY, Yeh WS et al (2014) Idiopathic pulmonary fibrosis in US Medicare beneficiaries aged 65 years and older: incidence, prevalence, and survival, 2001-11. Lancet Respir Med 2:566–572
4. Nathan SD, Shlobin OA, Weir N et al (2011) Long-term course and prognosis of idiopathic pulmonary fibrosis in the new millennium. Chest 140:221–229
5. Flaherty KR, Mumford JA, Murray S et al (2003) Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. Am J Respir Crit Care Med 168:543–548
6. Jacob J, Bartholmai BJ, Rajagopalan S et al (2017) Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based CT analysis with conventional severity measures. Eur Respir J:49
7. Lee SH, Kim SY, Kim DS et al (2016) Predicting survival of patients with idiopathic pulmonary fibrosis using GAP score: a nationwide cohort study. Respir Res 17:131
8. King TE Jr, Albera C, Bradford WZ et al (2014) All-cause mortality rate in patients with idiopathic pulmonary fibrosis. Implications for the design and execution of clinical trials. Am J Respir Crit Care Med 189:825–831
9. Martinez FJ, Safrin S, Weycker D et al (2005) The clinical course of patients with idiopathic pulmonary fibrosis. Ann Intern Med 142:963–967
10. Kishaba T, Tamaki H, Shimaoka Y, Fukuyama H, Yamashiro S (2014) Staging of acute exacerbation in patients with idiopathic pulmonary fibrosis. Lung 192:141–149
11. Kolb M, Collard HR (2014) Staging of idiopathic pulmonary fibrosis: past, present and future. Eur Respir Rev 23:220–224
12. Egan JJ, Martinez FJ, Wells AU, Williams T (2005) Lung function estimates in idiopathic pulmonary fibrosis: the potential for a simple classification. Thorax 60:270–273

13. du Bois RM, Weycker D, Albera C et al (2011) Ascertainment of individual risk of mortality for patients with idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 184:459–466
14. Ley B, Ryerson CJ, Vittinghoff E et al (2012) A multidimensional index and staging system for idiopathic pulmonary fibrosis. Ann Intern Med 156:684–691
15. Wells AU, Desai SR, Rubens MB et al (2003) Idiopathic pulmonary fibrosis: a composite physiologic index derived from disease extent observed by computed tomography. Am J Respir Crit Care Med 167:962–969
16. Hansell DM, Goldin JG, King TE Jr, Lynch DA, Richeldi L, Wells AU (2015) CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the Fleischner Society. Lancet Respir Med 3:483–496
17. Humphries SM, Yagihashi K, Huckleberry J et al (2017) Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. Radiology. https://doi.org/10.1148/radiol.2017161177:161177
18. Salisbury ML, Lynch DA, van Beek EJ et al (2017) Idiopathic pulmonary fibrosis: the association between the adaptive multiple features method and fibrosis outcomes. Am J Respir Crit Care Med 195:921–929
19. Park HJ, Lee SM, Song JW et al (2016) Texture-based automated quantitative assessment of regional patterns on initial CT in patients with idiopathic pulmonary fibrosis: relationship to decline in forced vital capacity. AJR Am J Roentgenol 207:976–983
20. Maldonado F, Moua T, Rajagopalan S et al (2014) Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. Eur Respir J 43:204–212
21. Lee SM, Seo JB, Oh SY et al (2018) Prediction of survival by texture-based automated quantitative assessment of regional disease patterns on CT in idiopathic pulmonary fibrosis. Eur Radiol 28:1293–1300
22. Humphries SM, Swigris JJ, Brown KK et al (2018) Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis. Eur Respir J 52(3)
23. Yoon RG, Seo JB, Kim N et al (2013) Quantitative assessment of change in regional disease patterns on serial HRCT of fibrotic interstitial pneumonia with texture-based automated quantification system. Eur Radiol 23:692–701
24. Clukers J, Lanclus M, Mignot B et al (2018) Quantitative CT analysis using functional imaging is superior in describing disease progression in idiopathic pulmonary fibrosis compared to forced vital capacity. Respir Res 19:213
25. Schmidt SL, Tayob N, Han MK et al (2014) Predicting pulmonary fibrosis disease course from past trends in pulmonary function. Chest 145:579–585
26. Cottin V, Hansell DM, Sverzellati N et al (2017) Effect of emphysema extent on serial lung function in patients with idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. https://doi.org/10.1164/rccm.201612-2492OC
27. Collard HR (2017) Improving survival in idiopathic pulmonary fibrosis: the race has just begun. Chest 151:527–528
28. Jacob J, Bartholmai BJ, Rajagopalan S et al (2016) Automated quantitative computed tomography versus visual computed tomography scoring in idiopathic pulmonary fibrosis: validation against pulmonary function. J Thorac Imaging 31:304–311
29. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J (2008) Fleischner society: glossary of terms for thoracic imaging. Radiology 246:697–722
30. Edey AJ, Devaraj AA, Barker RP, Nicholson AG, Wells AU, Hansell DM (2011) Fibrotic idiopathic interstitial pneumonias: HRCT findings that predict mortality. Eur Radiol 21:1586–1593
31. Coblentz CL, Babcook CJ, Alton D, Riley BJ, Norman G (1991) Observer variation in detecting the radiologic features associated with bronchiolitis. Invest Radiol 26:115–118
32. Kaplan ELM, P (1958) Non-parametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
33. Linden A (2006) Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. J Eval Clin Pract 12:132–139
34. Gafa G, Sverzellati N, Bonati E et al (2012) Follow-up in pulmonary sarcoidosis: comparison between HRCT and pulmonary function tests. Radiol Med 117:968–978
35. Tashkin DP, Roth MD, Clements PJ et al (2016) Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): a randomised controlled, double-blind, parallel group trial. Lancet Respir Med 4:708–719
36. Lee HY, Lee KS, Jeong YJ et al (2012) High-resolution CT findings in fibrotic idiopathic interstitial pneumonias with little honeycombing: serial changes and prognostic implications. AJR Am J Roentgenol 199:982–989
37. Kim HG, Tashkin DP, Clements PJ et al (2010) A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol 28:S26–S35
38. Jacob J, Bartholmai BJ, Rajagopalan S et al (2018) Serial automated quantitative CT analysis in idiopathic pulmonary fibrosis: functional correlations and comparison with changes in visual CT scores. Eur Radiol 28:1318–1327
39. Jacob J, Bartholmai BJ, Rajagopalan S et al (2018) Predicting outcomes in idiopathic pulmonary fibrosis using automated computed tomographic analysis. Am J Respir Crit Care Med 198:767–776
40. De Giacomi F, Raghunath S, Karwoski R, Bartholmai BJ, Moua T (2018) Short-term automated quantification of radiologic changes in the characterization of idiopathic pulmonary fibrosis versus non-specific interstitial pneumonia and prediction of long-term survival. J Thorac Imaging 33:124–131
41. Hwang JH, Misumi S, Curran-Everett D, Brown KK, Sahin H, Lynch DA (2011) Longitudinal follow-up of fibrosing interstitial pneumonia: relationship between physiologic testing, computed tomography changes, and survival rate. J Thorac Imaging 26:209–217
42. Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15:361–387