Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis

Stephen M. Humphries[1], Jeffrey J. Swigris[2], Kevin K. Brown[2], Matthew Strand[3], Qi Gong[4], John S. Sundy[4], Ganesh Raghu[5], Marvin I. Schwarz[6], Kevin R. Flaherty[7], Rohit Sood[8], Thomas G. O'Riordan[4] and David A. Lynch[1]

Affiliations: [1]Dept of Radiology, National Jewish Health, Denver, CO, USA. [2]Division of Pulmonary and Critical Care Medicine, National Jewish Health, Denver, CO, USA. [3]Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, CO, USA. [4]Gilead Sciences Inc., Foster City, CA, USA. [5]Center for Interstitial Lung Diseases, Dept of Medicine, University of Washington, Seattle, WA, USA. [6]Division of Pulmonary Sciences and Critical Care Medicine, University of Colorado, Aurora, CO, USA. [7]Division of Pulmonary and Critical Care Medicine, University of Michigan, Ann Arbor, MI, USA. [8]PAREXEL International, Billerica, MA, USA.

Correspondence: Stephen M. Humphries, Quantitative Imaging Laboratory, Dept of Radiology, National Jewish Health, 1400 Jackson Street, Denver, CO 80206-2761, USA. E-mail: humphriess@njhealth.org

@ERSpublications
In subjects with IPF, quantification of lung fibrosis extent on HRCT using data-driven texture analysis shows acceptable performance characteristics and minimal clinically important difference in the range of 3.4–6.4% http://ow.ly/fFNc30lfAGh

Cite this article as: Humphries SM, Swigris JJ, Brown KK, *et al*. Quantitative high-resolution computed tomography fibrosis score: performance characteristics in idiopathic pulmonary fibrosis. *Eur Respir J* 2018; 52: 1801384 [https://doi.org/10.1183/13993003.01384-2018].

ABSTRACT We evaluated performance characteristics and estimated the minimal clinically important difference (MCID) of data-driven texture analysis (DTA), a high-resolution computed tomography (HRCT)-derived measurement of lung fibrosis, in subjects with idiopathic pulmonary fibrosis (IPF).

The study population included 141 subjects with IPF from two interventional clinical trials who had both baseline and nominal 54- or 60-week follow-up HRCT. DTA scores were computed and compared with forced vital capacity (FVC), diffusing capacity of the lung for carbon monoxide, distance covered during a 6-min walk test and St George's Respiratory Questionnaire scores to assess the method's reliability, validity and responsiveness. Anchor- and distribution-based methods were used to estimate its MCID.

DTA had acceptable reliability in subjects appearing stable according to anchor variables at follow-up. Correlations between the DTA score and other clinical measurements at baseline were moderate to weak and in the hypothesised directions. Acceptable responsiveness was demonstrated by moderate to weak correlations (in the directions hypothesised) between changes in the DTA score and changes in other parameters. Using FVC as an anchor, MCID was estimated to be 3.4%.

Quantification of lung fibrosis extent on HRCT using DTA is reliable, valid and responsive, and an increase of ~3.4% represents a clinically important change.

This article has supplementary material available from erj.ersjournals.com

Received: July 23 2018 | Accepted after revision: July 26 2018

Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic fibrosing interstitial lung disease whose aetiology remains unknown [1]. It is characterised by progressive scarring of the lung parenchyma that leaves patients with increasing dyspnoea, decreasing quality of life and shortened survival. Its median survival is estimated at 3–5 years, although individual prognosis can vary significantly [2–5]. Some patients suffer rapid disease progression and early death, while others decline more slowly with variable periods of clinical stability [6]. Accurate assessment of disease severity and of change over time is critical for both clinical care and therapeutic trials.

Pulmonary physiology, particularly forced vital capacity (FVC), is the standard method for longitudinal monitoring of disease progression. However, it is an indirect measure of disease activity with drawbacks including dependency on technique and patient effort, variability in rate of change [7], and lack of sensitivity to subtle changes in disease status [8]. While it has been the subject of debate [9], decline in FVC is generally accepted as a surrogate end-point for death in IPF [10]. Still, it is widely recognised that additional valid and reliable outcome measures are needed [6, 11].

High-resolution computed tomography (HRCT) plays an essential role in the evaluation of patients with IPF. It provides noninvasive visualisation of lung parenchyma, can diagnose IPF without lung biopsy in the majority of patients and has been used for entry into clinical trials. However, visual assessment of HRCT images is limited by interobserver variation [12] and is insufficiently precise for longitudinal evaluation [13].

Computational methods for quantitative evaluation of HRCT have emerged as promising objective markers of disease severity in pulmonary fibrosis [14]. HRCT-derived scores for fibrosis extent correlate with degree of physiological impairment at baseline and may be more sensitive to subtle changes in disease status than physiological metrics [15]. In previous work, we showed that extent of lung fibrosis, quantified on HRCT using a method called data-driven texture analysis (DTA), provides an IPF severity index that correlates with expert visual assessment and lung function, and could be used to predict longitudinal disease behaviour better than semiquantitative visual scores or HRCT lung histogram-based metrics [16].

To be accepted as outcome measures, the performance characteristics of HRCT lung fibrosis scores require further study. In this work, we analysed pooled subject-level data from two IPF treatment trials (PANTHER-IPF [17] and RAINIER [18]) to assess the reliability, validity and responsiveness, and to estimate the minimal clinically important difference (MCID) (the smallest difference in an outcome measure that would be meaningful to the patient [19]) of DTA.

The sequential HRCT analysis of the PANTHER-IPF data has been reported previously [16], but test characteristics were not evaluated. Some of the data reported here have been presented in poster format [20].

Methods

Study design and population

As all analyses were conducted retrospectively on previously collected, de-identified data, this study was exempt from additional institutional review board approval. Methods for the PANTHER-IPF and RAINIER trials have been published previously [17, 18]. We included subjects who had both baseline and follow-up (15 and 12.5 months for PANTHER-IPF and RAINIER, respectively) data. Briefly, for inclusion in PANTHER-IPF, a three-armed trial of placebo *versus* *N*-acetylcysteine *versus* a three-drug combination (prednisone, azathioprine and *N*-acetylcysteine), patients required a diagnosis of IPF made using criteria similar to subsequently published international consensus guidelines [1]. A subset of 72 subjects underwent both baseline and nominal 15-month follow-up volumetric (axial slice thickness and spacing $\leqslant$1.25 mm) HRCT. RAINIER was a placebo-controlled trial of simtuzumab, a monoclonal antibody against the lysyl-oxidase like-2 enzyme, conducted from March 2011 to January 2016 and terminated prematurely for lack of efficacy. In this trial the diagnosis of IPF was made in accordance with accepted criteria [1]. A subset of subjects from sites in the USA underwent both baseline and nominal 54-week follow-up HRCT. In RAINIER, HRCT protocols were more varied, but only series with axial slice thickness $\leqslant$2.5 mm and limited gaps (slice spacing $\leqslant$10 mm) were included in this analysis (n=69). HRCT scans showing excessive motion artefacts, inadequate inspiration or an incomplete depiction of the lung parenchyma, identified by visual assessment, were omitted from this analysis. Baseline and follow-up HRCT with similar protocols were matched, to the extent possible. Case selection and summary characteristics of HRCT are included in supplementary figure E1 and supplementary table E1. In each trial, standard demographic, physiological and patient-reported outcome variables were collected, including FVC, diffusing capacity of the lung for carbon monoxide (D_{LCO}), distance covered during a 6-min walk test (6MWD) and response data from the St George's Respiratory Questionnaire (SGRQ). The

SGRQ is a respiratory disease-specific health-related quality of life questionnaire with 50 items separated into three domains (Symptoms, Activity and Impacts). Each domain score and the SGRQ total score has a range from 0 to 100, with higher scores corresponding to greater impairments [21].

HRCT analysis

DTA is a machine learning method capable of automatic detection and quantification of lung fibrosis on HRCT [16]. It is trained to discriminate fibrosis using radiologist-identified image regions demonstrating normal lung parenchyma and usual interstitial pneumonia patterns. Exemplar regions labelled as reticulation, honeycombing or traction bronchiectasis were used to define the fibrosis category. The algorithm classifies local regions in axial sections as either normal lung or fibrosis in a sliding window fashion over lung fields, which are identified in a separate segmentation process. The DTA fibrosis score is computed as the percentage of the total number of window regions classified as fibrosis (figure 1). Performing classification on axial images enables analysis of studies with noncontiguous sections.

Statistical analyses

Summary statistics were generated for baseline characteristics. In the performance characteristics analyses, we used several disease severity variables as anchors against which DTA scores were compared. Anchors included FVC, *D*LCO, 6MWD and SGRQ scores.

Several analyses were conducted to support the validity of the DTA score as a measure capable of capturing baseline and change in IPF severity. For concurrent validity analyses, we examined associations between baseline values for the DTA score and each anchor (FVC, *D*LCO, 6MWD and SGRQ scores) by using Spearman correlation coefficients. Known-groups validity was assessed by comparing mean DTA fibrosis scores across anchor-defined, discrete subgroups of IPF severity determined by stratifying the cohort on baseline values for FVC, *D*LCO, 6MWD and SGRQ. One-way ANOVA was used for statistical comparisons, as well as p-value-adjusted pairwise comparisons using the Tukey method.

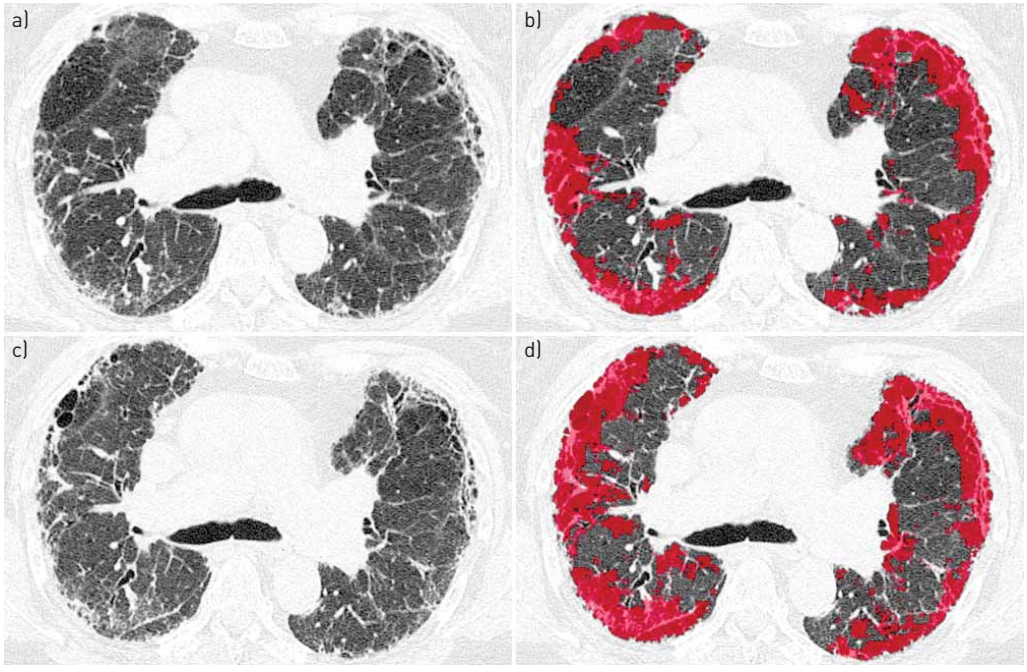

FIGURE 1 High-resolution computed tomography (HRCT) images and regions classified as fibrosis by data-driven texture analysis (DTA). FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire. a) Axial baseline unenhanced HRCT image in a 66-year-old female. b) Regions classified as fibrosis by DTA shown in red. The baseline DTA fibrosis score was 39.5%. Baseline FVC % pred was 88.11%, *D*LCO % pred was 43.0%, 6MWD was 366 m and SGRQ total score was 42.5 points. c) Similar axial unenhanced HRCT image in the same subject at nominal 60-week follow-up. d) Regions classified as fibrosis by DTA shown in red. The DTA fibrosis score increased 12.6 percentage points at follow-up. FVC declined 20.6% (relative to baseline), *D*LCO declined 16.0% (relative to baseline), 6MWD declined 52 m and SGRQ total score increased 23.5 points. Considering the cubic relationship between volume and radius, this level of change in fibrosis extent represents an increase of only a few millimetres in the effective radius of fibrotic regions. This may be difficult to detect by eye as a change of a few pixels on cross-sectional images.

Responsiveness was assessed with Spearman correlation coefficients between change from baseline in the DTA score and change from baseline in each anchor variable. These values were also used to gauge the appropriateness of anchors for estimation of MCID. Following Cohen's rule of thumb, anchors with correlation coefficients ⩾0.30 when compared with DTA change were considered appropriate [22, 23].

DTA scores were compared across groups of subjects stratified into discrete, anchor-defined categories of change in IPF severity. Stratification levels for each anchor variable were selected to represent groups that were "much worse", "slightly worse", "same", "slightly better" and "much better". Cut-off values for each anchor were selected based on published data. For example, in IPF the MCID for FVC is estimated to be ~5% [24], so the "same" group for this anchor was defined as subjects whose change in FVC was within ±5% relative to baseline and the cut-off between "slightly worse" and "much worse" was set at twice this value. Other published MCIDs in IPF are ~30 m for 6MWD [25] and 7 points for SGRQ total score [26]. In chronic obstructive pulmonary disease, MCID for *D*LCO has been estimated to be ~10% [27]. Test–retest reliability was evaluated by computing the intraclass correlation coefficient (ICC) of baseline and follow-up DTA scores for subjects in the "same" groups defined by stable values in each external anchor. ICC values <0.5, 0.5–0.75, 0.75–0.9 and >0.90 were interpreted as poor, moderate, good and excellent reliability, respectively [28]. Mean changes in DTA scores for anchor-defined groups were compared using ANOVA and p-value-adjusted pairwise comparisons using t-tests.

MCID is defined as the smallest difference in an outcome measure that can be considered important and that would lead a clinician to consider a change in therapy [25]. While there is no consensus agreement on the ideal method to determine MCID, current best practice is to use several approaches to estimate a practical range [23, 26]. Attempts to triangulate the MCID for DTA were made using both anchor- and distribution-based methods. Anchor-based estimation of MCID is a special case of responsiveness where mean change in the DTA score for subjects who change minimally according to a given anchor provides an estimate for MCID. We considered the mean values of DTA change scores in the "slightly worse" groups, described earlier, as estimates of MCID for worsening. Effect size, the difference in mean DTA scores at baseline and follow-up divided by standard deviation of the DTA score at baseline, was used to estimate the magnitude of MCID. Effect size values of 0.2, 0.5 and 0.8 are considered small, medium and large, respectively [24].

Distribution-based methods use sample data only, relying on statistical properties of the distribution of outcome scores to estimate the MCID [19]. We used the standard deviation of the DTA score at baseline and the standard error of measurement for these estimates. The amount of change in the outcome variable that corresponds to a moderate effect size, *i.e.* one half of the standard deviation at baseline, can be used as an estimate of MCID [25, 29]. Finally, the standard error of measurement (another estimate of MCID [24]) was calculated as SEM=(SD of DTA at baseline)×sqrt(1–ICC).

Statistical analyses were performed using R version 3.4.2 [30] and p-values <0.05 were considered statistically significant.

Results

The final cohort was comprised of 141 subjects who had baseline and follow-up data available for analysis. Demographics, baseline values and changes at follow-up are presented in table 1. The mean±SD age of the

TABLE 1 Study population[#]

	Baseline		Change at follow-up	
	Subjects n	Mean±SD	Subjects n	Mean±SD
Age years	141	68.0±8.2	141	
FVC % pred	141	68.9±15.2	141	−5.68±8.79%[¶]
FVC L	141	2.74±0.73	141	−6.14±8.79%[¶]
***D*LCO % pred**	141	43.6±11.9	141	−6.89±18.65%[¶]
***D*LCO mL·min^{-1}·mmHg^{-1}**	141	12.91±4.03	141	−7.16±18.50%[¶]
6MWD m	139	393.0±93.5	138	−21.5±80.9
SGRQ total score	138	39.4±17.2	135	3.4±12.6
DTA score	141	28.0±12.9	141	4.4±7.5
CT TLC L	141	3.83±1.08	141	−0.17±0.58

FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire; DTA: data-driven texture analysis; CT: computed tomography; TLC: total lung capacity. [#]: PANTHER-IPF n=72, RAINIER n=69, male n=108 (76.6%); [¶]: relative to baseline (follow-up value minus baseline divided by baseline).

TABLE 2 Validity testing using Spearman's correlation coefficient (ρ) between the data-driven texture analysis (DTA) score and anchors at baseline

	Subjects n	Baseline DTA score	
		ρ	p-value
FVC % pred	141	−0.46	<0.001
FVC	141	−0.32	<0.001
***D*LCO % pred**	141	−0.61	<0.001
***D*LCO**	141	−0.49	<0.001
6MWD	139	−0.15	0.079
SGRQ total score	138	0.28	<0.001

FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire.

pooled cohort was 68.0±8.2 years and 108 (76.6%) were male. Mean±SD baseline FVC % pred was 68.9±15.2%, *D*LCO % pred was 43.6±11.9%, 6MWD was 393.0±93.5 m and SGRQ total score was 39.4±17.2. The mean±SD DTA score at baseline was 28.0±12.9%. On average, subjects showed slight progression (mean FVC decline 6.14% relative to baseline) over the follow-up period.

Table 2 shows baseline correlations between the DTA score and the clinical variables. There were weak to moderate correlations in the expected directions between DTA score and each anchor at baseline. Table 3 presents the results for the known-groups validity analyses of baseline data. Mean values for the DTA score were generally higher for subjects with poorer lung function, 6MWD and health-related quality of life. ANOVA showed the mean baseline DTA score differed across FVC, *D*LCO and SGRQ groups defined by severity, but not across the spectrum of 6MWD values. Tukey multiple comparison of means showed the mean DTA score was significantly different between any two FVC tertiles and any two *D*LCO tertiles ($p<0.05$, with Bonferroni adjustment). The subgroup with the lowest SGRQ score had a mean DTA score significantly different from the other two SGRQ-defined subgroups, whose mean DTA scores were not significantly different from each other.

Table 4 demonstrates responsiveness using Spearman correlations between the DTA change score, calculated as follow-up score minus baseline score, and changes in FVC, *D*LCO, 6MWD and SGRQ values. Correlations were weakly to moderately strong and in the expected directions. The absolute value of all correlation coefficients was ⩾0.30, supporting the appropriateness of each anchor for estimation of MCID.

TABLE 3 Mean of baseline data-driven texture analysis (DTA) scores by tertiles of anchor variables

Anchor variable	p-value[#]	Subjects n	Mean DTA score (95% CI)
FVC % pred			
⩽55		31	36.1 (32.1–40.1)
>55 to ⩽70	<0.0001	49	29.4 (25.8–33.0)
>70		61	22.8 (19.7–25.8)
***D*LCO % pred**			
⩽35		33	37.2 (32.1–42.3)
>35 to ⩽55	<0.0001	84	28.0 (25.7–30.2)
>55		24	15.4 (12.3–18.5)
6MWD m			
⩽350		35	31.4 (26.4–36.3)
>350 to ⩽450	0.173	68	27.8 (24.9–30.7)
>450		37	25.8 (21.6–29.9)
SGRQ total score			
>48		43	30.3 (26.1–34.4)
>33 to ⩽48	0.003	46	31.4 (27.7–35.1)
⩽33		50	23.3 (20.0–26.5)

FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire. [#]: one-way ANOVA.

TABLE 4 Responsiveness shown as Spearman's correlation coefficient (ρ) between change in the data-driven texture analysis score and changes in anchors at follow-up

	Subjects n	Change in DTA score	
		ρ	p-value
FVC[#]	141	−0.46	<0.001
***D*LCO**[#]	141	−0.38	<0.001
6MWD m	138	−0.30	<0.001
SGRQ total score	135	0.37	<0.001

FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire. [#]: percentage change relative to baseline (follow-up value minus baseline divided by baseline).

Data for anchor-based estimation of MCID are presented in table 5. Mean change in DTA scores was stratified into groups according to "much worse", "slightly worse", "same", "slightly better" and "much better" changes in each external anchor variable. ICC for baseline and follow-up DTA scores in each subgroup of stable subjects ranged from 0.78 to 0.91 (mean 0.83), showing good to excellent reliability. ANOVA showed significant differences in the means of DTA change scores across each set. Increase in the DTA score was consistently greater for subjects with larger declines in pulmonary function, 6MWD and health-related quality of life. The mean DTA change score in the "slightly worse" group is an estimate of MCID. Effect sizes in this group are small to medium for each anchor.

Table 6 summarises results for estimation of MCID using anchor- and distribution-based methods. Estimates of MCID were fairly consistent, ranging from 3.4% to 6.4%.

TABLE 5 Anchor-based estimation of minimal clinically important difference (MCID)

Anchor	p-value[#]	Subjects n	Mean change in DTA score (95% CI)	Effect size	ICC
FVC[¶]					
Much worse (⩽−10%)		45	9.1 (7.0–11.3)	0.75	
Slightly worse (>−10% to ⩽−5%)		33	*3.4 (0.5–6.3)*	0.28	
Same (>−5% to ⩽5%)	<0.0001	47	1.8 (0.1–3.6)	0.13	0.91
Slightly better (>5% to ⩽10%)		13	0.7 (−2.0–3.5)	0.06	
Much better (>10%)		3	1.1 (−12.2–14.4)	0.15	
***D*LCO**[¶]					
Much worse (⩽−15%)		47	7.6 (5.6–9.6)	0.53	
Slightly worse (>−15% to ⩽−10%)		10	*5.1 (−1.6–11.8)*	0.48	
Same (>−10% to ⩽10%)	0.0028	66	3.1 (1.5–4.7)	0.28	0.84
Slightly better (>10% to ⩽15%)		6	3.4 (−10.8–17.6)	0.21	
Much better (>15%)		12	−1.1 (−5.0–2.8)	−0.09	
6MWD m					
Much worse (⩽−60)		33	7.7 (5.1–10.1)	0.54	
Slightly worse (>−60 to ⩽−30)		25	*5.3 (3.0–7.5)*	0.40	
Same (>−30 to ⩽30)	0.0002	47	2.7 (0.3–5.1)	0.23	0.78
Slightly better (>30 to ⩽60)		18	5.1 (1.4–8.8)	0.53	
Much better (>60)		15	0.0 (−2.7–2.7)	0.0	
SGRQ total score					
Much worse (>12)		31	7.4 (4.4–10.3)	0.51	
Slightly worse (>7 to ⩽12)		24	*5.4 (2.5–8.2)*	0.45	
Same (>−7 to ⩽7)	0.0090	57	3.7 (1.6–5.9)	0.31	0.78
Slightly better (>−12 to ⩽−7)		11	2.2 (−1.1–5.5)	0.12	
Much better (⩽−12)		12	−0.3 (−3.0–2.5)	−0.24	

DTA: data-driven texture analysis; ICC: intraclass correlation coefficient; FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire. Mean change in DTA score (95% CI) and effect sizes stratified by changes in other measures. Stratification levels were selected to represent groups that were "much worse", "slightly worse", "same", "slightly better" and "much better". Cut-off values for each anchor were selected based on published MCIDs. Mean change in DTA in each "slightly worse" group (shown in italics) is an anchor-based estimate of MCID. ICC of baseline and follow-up DTA score in each "same" group is an estimate of reliability. [#]: one-way ANOVA; [¶]: percentage change relative to baseline (follow-up value minus baseline divided by baseline).

TABLE 6 Estimates of the minimal clinically important difference (MCID) for worsening of the data-driven texture analysis (DTA) score in patients with idiopathic pulmonary fibrosis

Method		MCID %
Anchor	FVC	3.40
Anchor	*D*LCO	5.09
Anchor	6MWD	5.28
Anchor	SGRQ total score	5.35
Distribution	0.5×SD at baseline	6.44
Distribution	SEM=(SD of DTA at baseline)×sqrt(1–ICC[#])	3.86
Distribution	SEM=(SD of DTA at baseline)×sqrt(1–ICC[¶])	6.04

FVC: forced vital capacity; *D*LCO: diffusing capacity of the lung for carbon monoxide; 6MWD: 6-min walk distance; SGRQ: St George's Respiratory Questionnaire; SEM: standard error of measurement; ICC: intraclass correlation coefficient. [#]: using ICC for FVC anchor (0.91); [¶]: using ICC for 6MWD anchor (0.78).

Discussion

HRCT of the chest is relied upon for the diagnosis and management of patients with IPF. Computational methods that produce quantitative HRCT scores for disease extent show promise as precise and objective outcome measures in IPF, but require further systematic performance testing. In this study, we used baseline and longitudinal data from two well-characterised populations to evaluate the performance characteristics of DTA. Confirming our prior work, correlations between the DTA score and pulmonary function tests at baseline were moderate. Extending our previous findings, additional anchor variables (6MWD and SGRQ) and known-groups validity testing show that DTA can distinguish subjects with differing levels of disease severity. It also showed good to excellent test–retest reliability in subjects determined to be stable based on anchor variables and it was responsive to changes in measured disease severity. Finally, we estimated, using both anchor- and distribution-based methods, MCID for worsening to be in the range of 3.4–6.4%.

Other researchers have used computational methods to evaluate lung fibrosis on HRCT. Like DTA, scores from multiple methods correlate with measures of pulmonary physiology. For example, JACOB *et al.* [31] showed that CALIPER, a quantification method based on local histograms of pixel intensity within volumes of interest, provided image-derived metrics of lung fibrosis that correlated more strongly with FVC at baseline than did visual scoring. PARK *et al.*'s [32] texture-based quantification system showed fibrosis score correlations with baseline FVC and their measure of reticulation was predictive of decline in FVC at 1-year follow-up. KIM *et al.* [33] observed that a quantitative lung fibrosis score, computed by a machine learning algorithm trained with image textural features and expert labelled image regions, correlated with baseline values for FVC and *D*LCO. At 7-month follow-up, change in quantitative lung fibrosis score was also associated with changes in FVC and *D*LCO. SALISBURY *et al.* [15] have also analysed HRCT scans from the PANTHER-IPF cohort. Using the AMFM (Adaptive Multi-Feature Method) algorithm they showed that baseline score for a ground-glass reticular pattern was independently associated with risk of a composite outcome of death, hospitalisation or 10% decline in FVC over 60 weeks. The change in this score was only weakly correlated ($r=-0.25$; $p=0.01$) with change in FVC at follow-up.

DTA is implemented as a simple convolutional neural network. It is based on unsupervised feature learning; image features used for classification were discovered in an initial clustering process, in contrast to engineered features that are chosen by the algorithm designer. In image texture analysis, engineered features are often based on first- and second-order pixel statistics within local regions. A weakness of feature engineering is the bias introduced in the design and feature selection process. Learned features rely on fewer design choices and tend to capture important details better than manually designed features [34]. Future work will evaluate the benefits of more complex convolutional neural network architectures in detection and quantification of diffuse lung diseases.

In October 2014, the US Food and Drug Administration approved two antifibrotic drugs for IPF (pirfenidone and nintedanib) based on changes in FVC [35]. In the confirmatory trials, the modelled average decline in FVC was ~100 mL per year in subjects on either of the approved treatments [36, 37] compared with ~200 mL per year in subjects on placebo [38]. These approvals have reshaped the landscape for future drug trials in IPF, as most upcoming trial subjects will be on one of these drugs [8, 39]. As measuring differences in FVC below 100 mL will be difficult, additional reliable, responsive and validated outcome measures of disease activity are needed.

While repeat HRCT within a short time interval was not available in this study, we observed acceptable reliability (ICC=0.78–0.91) in DTA fibrosis scores in subjects who remained stable, based on each anchor variable, during the follow-up period. We also observed that greater DTA fibrosis scores corresponded with greater degree of physiological impairment, reduced exercise tolerance and reduced patient-reported quality of life, and that changes in DTA fibrosis scores were moderately correlated with changes in external anchors. Confirmation in a separate population would be ideal; however, an appropriate, independent dataset with sequential scans and physiology was not available at the time of this analysis. As quantitative imaging using machine learning continues to advance, there is an increasingly urgent need for standardised imaging cohorts in IPF and other fibrotic lung diseases that can be used to develop, test and validate methods. It is likely that available datasets would help drive innovation in the field.

Estimates of MCID are useful for consistent interpretation of results and for sample size calculations in clinical trial design. Distribution-based methods are more straightforward to calculate and provide an estimate of the degree of change in an outcome that is unlikely to be attributable to random measurement variation [19]. However, they lack the context provided by external anchors. Anchor-based methods are generally preferred [23], because they determine MCID as the degree of change in the outcome that is associated with a clinically relevant change in an external variable. We chose FVC, D_{LCO}, 6MWD and SGRQ as external anchors because they are well-known indices of severity in IPF, and are routinely measured in clinical care and therapeutic trials. Of these, only FVC could be considered a validated outcome and this may be the best anchor. However, we included D_{LCO}, 6MWD and SGRQ because, despite showing greater variability, they met minimum criteria for appropriateness and have been used as anchors in estimates of MCID of FVC [25]. Progression of morphological fibrosis on HRCT may be relatively independent of physiological progression, which may explain why the correlations between these measures in our study are not very strong. In fact, DTA may function best as a complementary measure rather than a substitute for physiological evaluation.

Strengths of the present study include use of pooled data, acquired prospectively in clinical trials, and the use of four external anchor variables in testing DTA's performance characteristics and estimation of its MCID. There are also several limitations to be noted. First, this was a *post hoc* analysis and data beyond 60 weeks were not available. Follow-up HRCT was available on only a small fraction of total subjects enrolled in each trial and this may represent a selection bias toward subjects with less aggressive disease progression. Second, there was variation in HRCT parameters and a slightly different follow-up interval in the trials. Differences in HRCT acquisition and reconstruction parameters, and in the level of lung inflation during a scan, are well-known sources of variation in quantitative HRCT of the lungs [40]. These effects can be alleviated by using standardised HRCT protocols that require only short breath holds and coaching subjects on the importance of reaching full inspiration for the scan [41]. We speculate that improved consistency in HRCT characteristics would reduce variability and improve performance of DTA or any quantitative image analysis method. Third, repeat HRCT over a short time interval was not available for test–retest analysis. Fourth, subjects either remained clinically stable or declined, so our MCID estimates are for worsening only. Finally, other methods for fibrosis quantification on HRCT have been proposed, but we did not perform direct comparisons of different algorithms.

This study demonstrates quantitative measurement of lung fibrosis on HRCT is a reliable, valid and responsive measure of disease severity in a cohort combining subjects with IPF from two clinical trial populations. We estimated that DTA's MCID for worsening in IPF is in the range of 3.4–6.4%. This work suggests that quantitative HRCT using DTA, an image-based measure of morphology, may be a valuable additional tool for assessing outcomes in IPF that should be tested in prospective clinical trials.

Author contributions: Concept and design: S.M. Humphries, J.J. Swigris, K.K. Brown and D.A. Lynch. Data acquisition, analysis and interpretation: S.M. Humphries, J.J. Swigris, Qi Gong, J.S. Sundy, G. Raghu, M. Strand, M.I. Schwarz, K.K. Brown, K.R. Flaherty, R. Sood, T.G. O'Riordan and D.A. Lynch. Drafted manuscript for important intellectual contribution: S.M. Humphries, J.J. Swigris, K.K. Brown and D.A. Lynch. Review and finalising of the manuscript: all authors

Support statement: Analysis of PANTHER-IPF study data was partially supported by NIH/NHLBI R01 HL091743 (K.R. Flaherty). Gilead Sciences funded quantitative analysis of HRCT in the RAINIER study. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: S.M. Humphries reports service contract for quantitative analysis of RAINIER HRCT scans from Gilead Sciences, during the conduct of the study; personal fees from Boehringer Ingelheim, grants from NHLBI, and service contract from PAREXEL Informatics, outside the submitted work; in addition, S.M. Humphries has a patent "Systems and methods for automatic detection and quantification of pathology using dynamic feature classification" pending to National Jewish Health. J.J. Swigris has nothing to disclose. K.K. Brown reports multiple lung fibrosis grants from NHLBI, personal fees from AstraZeneca, Bayer, Biogen, Fibrogen, Galecto, MedImmune, Novartis, Aeolus,

ProMetic, Patara, Third Pole, aTyr and Boehringer Ingelheim, conversations under CDAs with Genoa, Galapagos and Global Blood Therapeutics, grants and personal fees from Gilead Sciences, and submitted grant from Roche/Genentech, outside the submitted work. M. Strand has nothing to disclose. Q. Gong has nothing to disclose. J.S. Sundy reports being a full-time employee and stockholder in Gilead Sciences, Inc. G. Raghu has been a consultant on IPF and fibrotic lung diseases for Boehringer Ingelheim, BMS, Bellerophan, Roche/Genentech and Veracyte, and a consultant on IPF studies for Biogen, Fibrogen, Gilead Sciences, Nitto, Promedior, Patara and Sanofi, outside the submitted work. M.I. Schwarz has nothing to disclose. K.R. Flaherty reports grants and personal fees from Boehringer Ingelheim and Roche/Genentech, personal fees from Veracyte, Aeolus, Pharmakea, Fibrogen and Sanofi-Genzyme, and grants from Afferent, outside the submitted work. R. Sood reports that Gilead Sciences paid for cost of services for running the IPF clinical trial, during the conduct of the study. T.G. O'Riordan is a full-time employee and stockholder of Gilead Sciences. D.A. Lynch reports grants from NHLBI, personal fees and research support from PAREXEL and Veracyte, personal fees from Boehringer Ingelheim, Genentech/Roche and Acceleron, outside the submitted work; in addition, D.A. Lynch has a patent "Systems and methods for automatic detection and quantification of pathology using dynamic feature classification" pending to National Jewish Health.

References

1 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: 788–824.
2 Pérez ERF, Daniels CE, Schroeder DR, *et al.* Incidence, prevalence, and clinical course of idiopathic pulmonary fibrosis: a population-based study. *Chest* 2010; 137: 129–137.
3 Schmidt SL, Tayob N, Han MK, *et al.* Predicting pulmonary fibrosis disease course from past trends in pulmonary function. *Chest* 2014; 145: 579–585.
4 Martinez FJ, Safrin S, Weycker D, *et al.* The clinical course of patients with idiopathic pulmonary fibrosis. *Ann Intern Med* 2005; 142: 963–967.
5 Ley B, Collard HR, King TE Jr. Clinical course and prediction of survival in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011; 183: 431–440.
6 Robbie H, Daccord C, Chua F, *et al.* Evaluating disease severity in idiopathic pulmonary fibrosis. *Eur Respir Rev* 2017; 26: 170051.
7 Nathan SD, Albera C, Bradford WZ, *et al.* Effect of continued treatment with pirfenidone following clinically meaningful declines in forced vital capacity: analysis of data from three phase 3 trials in patients with idiopathic pulmonary fibrosis. *Thorax* 2016; 71: 429–435.
8 Hansell DM, Goldin JG, King TE, *et al.* 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* 2015; 3: 483–496.
9 Nathan SD, Meyer KC. IPF clinical trial design and endpoints. *Curr Opin Pulm Med* 2014; 20: 463–471.
10 Paterniti MO, Bi Y, Rekić D, *et al.* Acute exacerbation and decline in forced vital capacity are associated with increased mortality in idiopathic pulmonary fibrosis. *Ann Am Thorac Soc* 2017; 14: 1395–1402.
11 Lammi MR, Baughman RP, Birring SS, *et al.* Outcome measures for clinical trials in interstitial lung diseases. *Curr Respir Med Rev* 2015; 11: 163–174.
12 Watadani T, Sakai F, Johkoh T, *et al.* Interobserver variability in the CT assessment of honeycombing in the lungs. *Radiology* 2013; 266: 936–944.
13 Flaherty KR, Mumford JA, Murray S, *et al.* Prognostic implications of physiologic and radiographic changes in idiopathic interstitial pneumonia. *Am J Respir Crit Care Med* 2003; 168: 543–548.
14 Wu X, Kim GH, Salisbury ML, *et al.* Computed tomographic biomarkers in idiopathic pulmonary fibrosis: the future of quantitative analysis. *Am J Respir Crit Care Med* 2018; in press [https://doi.org/10.1164/rccm.201803-0444PP].
15 Salisbury ML, Lynch DA, Van Beek EJ, *et al.* Idiopathic pulmonary fibrosis: the association between the adaptive multiple features method and fibrosis outcomes. *Am J Respir Crit Care Med* 2017; 195: 921–929.
16 Humphries SM, Yagihashi K, Huckleberry J, *et al.* Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. *Radiology* 2017: 285: 270–278.
17 Idiopathic Pulmonary Fibrosis Clinical Research Network. Prednisone, azathioprine, and *N*-acetylcysteine for pulmonary fibrosis. *N Engl J Med* 2012; 2012: 1968–1977.
18 Raghu G, Brown KK, Collard HR, *et al.* Efficacy of simtuzumab versus placebo in patients with idiopathic pulmonary fibrosis: a randomised, double-blind, controlled, phase 2 trial. *Lancet Respir Med* 2017; 5: 22–32.
19 McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA* 2014; 312: 1342–1343.
20 Humphries S, O'Riordan TG, Sundy JS, *et al.* Change in CT-derived fibrosis score correlates with lung function progression in a clinical trial population with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2017; 195: A6783.
21 Swigris JJ, Esser D, Conoscenti CS, *et al.* The psychometric properties of the St George's Respiratory Questionnaire (SGRQ) in patients with idiopathic pulmonary fibrosis: a literature review. *Health Qual Life Outcomes* 2014; 12: 124.
22 Revicki D, Hays RD, Cella D, *et al.* Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; 61: 102–109.
23 Kafaja S, Clements PJ, Wilhalme H, *et al.* Reliability and minimal clinically important differences of FVC. Results from the Scleroderma Lung Studies (SLS-I and SLS-II). *Am J Respir Crit Care Med* 2017; 197: 644–652.
24 du Bois RM, Weycker D, Albera C, *et al.* Forced vital capacity in patients with idiopathic pulmonary fibrosis: test properties and minimal clinically important difference. *Am J Respir Crit Care Med* 2011; 184: 1382–1389.
25 Swigris JJ, Wamboldt FS, Behr J, *et al.* The 6 minute walk in idiopathic pulmonary fibrosis: longitudinal changes and minimum important difference. *Thorax* 2010; 65: 173–177.
26 Swigris JJ, Brown KK, Behr J, *et al.* The SF-36 and SGRQ: validity and first look at minimum important differences in IPF. *Respir Med* 2010; 104: 296–304.

27 Horita N, Miyazawa N, Kojima R, *et al.* Minimum clinically important difference in diffusing capacity of the lungs for carbon monoxide among patients with severe and very severe chronic obstructive pulmonary disease. *COPD* 2015; 12: 31–37.
28 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155–163.
29 Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003; 41: 582–592.
30 R Core Team. R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2017 .
31 Jacob J, Bartholmai BJ, Rajagopalan S, *et al.* Automated quantitative computed tomography versus visual computed tomography scoring in idiopathic pulmonary fibrosis: validation against pulmonary function. *J Thorac Imaging* 2016; 31: 304–311.
32 Park HJ, Lee SM, Song JW, *et al.* 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* 2016; 207: 976–983.
33 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: 70–80.
34 Coates A, Ng A, Lee H. An analysis of single-layer networks in unsupervised feature learning. 2011. http://proceedings.mlr.press/v15/coates11a/coates11a.pdf Date last accessed: August 4, 2018.
35 Karimi-Shah BA, Chowdhury BA. Forced vital capacity in idiopathic pulmonary fibrosis – FDA review of pirfenidone and nintedanib. *N Engl J Med* 2015; 372: 1189–1191.
36 Raghu G, Wells AU, Nicholson AG, *et al.* Effect of nintedanib in subgroups of idiopathic pulmonary fibrosis by diagnostic criteria. *Am J Respir Crit Care Med* 2017; 195: 78–85.
37 Okuda R, Hagiwara E, Baba T, *et al.* Safety and efficacy of pirfenidone in idiopathic pulmonary fibrosis in clinical practice. *Respir Med* 2013; 107: 1431–1437.
38 Raghu G. Idiopathic pulmonary fibrosis: lessons from clinical trials over the past 25 years. *Eur Respir J* 2017; 50: 1701209.
39 Ley B. Clarity on endpoints for clinical trials in idiopathic pulmonary fibrosis. *Ann Am Thorac Soc* 2017; 14: 1383–1384.
40 Coxson HO. Sources of variation in quantitative computed tomography of the lung. *J Thorac Imaging* 2013; 28: 272–279.
41 Newell JD Jr, Sieren J, Hoffman EA. Development of quantitative CT lung protocols. *J Thorac Imaging* 2013; 28: 266–271.