A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients

H.J. Kim[1], D.P. Tashkin[2], P. Clements[3], G. Li[6], M.S. Brown[4], R. Elashoff[5], D.W. Gjertson[7], F. Abtin[1], D.A. Lynch[8], D.C. Strollo[9], J.G. Goldin[1]

[1]Department of Radiological Sciences, [2]Department of Med-Pul & Critical Care, [3]Department of Med-Rheum, [4]Department of Radiological Sciences and [5]Department of Biostatistics and Biomathematics, David Geffen School of Medicine, UCLA, Los Angeles, CA; [6]Department of Biostatistics and [7]Department of Biostatistics and Pathology, School of Public Health, UCLA, Los Angeles, CA; [8]Radiology, Department National Jewish Health, Denver, CO; [9]Radiology Department, UPMC Presbyterian, Pittsburgh, PA, USA.

Hyun J. Kim, PhD
Donald P. Tashkin, MD
Philip Clements, MD
Gang Li, PhD
Matthew S. Brown, PhD
Robert Elashoff, PhD
David W. Gjertson, PhD
Fereidoun Abtin, MD
David A. Lynch, MD
Diane C. Strollo, MD
Jonathan G. Goldin, MD, PhD

Please address correspondence and reprint requests to:
Hyun J. Kim, PhD,
Department of Radiological Sciences, David Geffen School of Medicine, UCLA, 924 Westwood Blvd., Suite 650, Los Angeles, CA 90024-2926, USA.
E-mail: gracekim@mednet.ucla.edu

Received on April 16, 2010; accepted in revised form on August 7, 2010.

Clin Exp Rheumatol 2010; 28 (Suppl. 62): S26-S35.

Key words: texture feature, classification, scleroderma, interstitial lung disease, CAD

Conflict of interest: Dr Clements is a consultant to Gilead. Dr Lynch is a consultant to Intermune, Gilead, Centocor, Perspective Imaging, and Novartis; a member of the Advisory Board for the BUILD-3 study sponsored by Actelion; and is an independent contractor for Siemens Inc. Dr Goldin received funding for the study from NIH ROI Funding.
The other co-authors have declared no competing interests.

ABSTRACT

Objectives. *To evaluate an improved quantitative lung fibrosis score based on a computer-aided diagnosis (CAD) system that classifies CT pixels with the visual semi-quantitative pulmonary fibrosis score in patients with scleroderma-related interstitial lung disease (SSc-ILD).*

Methods. *High-resolution, thin-section CT images were obtained and analysed on 129 subjects with SSc-ILD (36 men, 93 women; mean age 48.8±12.1 years) who underwent baseline CT in the prone position at full inspiration. The CAD system segmented each lung of each patient into 3 zones. A quantitative lung fibrosis (QLF) score was established via 5 steps: 1) images were denoised; 2) images were grid sampled; 3) the characteristics of grid intensities were converted into texture features; 4) texture features classified pixels as fibrotic or non-fibrotic, with fibrosis defined by a reticular pattern with architectural distortion; and 5) fibrotic pixels were reported as percentages. Quantitative scores were obtained from 709 zones with complete data and then compared with ordinal scores from two independent expert radiologists. ROC curve analyses were used to measure performance.*

Results. *When the two radiologists agreed that fibrosis affected more than 1% or 25% of a zone or zones, the areas under the ROC curves for QLF score were 0.86 and 0.96, respectively.*

Conclusion. *Our technique exhibited good accuracy for detecting fibrosis at a threshold of both 1% (i.e. presence or absence of pulmonary fibrosis) and a clinically meaningful threshold of 25% extent of fibrosis in patients with SSc-ILD.*

Introduction

Scleroderma lung disease is the leading cause of death in patients with scleroderma (1-2). A recent study has suggested that an important predictor of survival is the extent of disease and extent of reticular pattern, which are visually scored on computer tomography (CT) (3-4). In another study, the visual score of pulmonary fibrosis, which was defined as reticular opacities with architectural distortion (*i.e.* traction bronchiectasis and bronchiolectasis) (5), alone was shown to be predictive of the therapeutic response to cyclophosphamide (6-7). CT is important in detecting and quantifying interstitial lung disease (ILD) in the management of scleroderma patients (8).

Visual scoring systems are limited by intra- and inter-reader variations (9-11). Development of a computer-based scoring system offers the potential for both reducing reader variation and standardising data across multiple sites. Though quantitative scoring of other lung diseases, such as emphysema, has been achieved (12, 13), several computer-aided diagnosis (CAD)-based systems that have been developed for assessing ILD or obstructive lung disease using texture features have not been applied in studies of large numbers of subjects (14-18). These methods have the potential to provide a score for abnormal patterns with respect to the extent of whole lung involvement, which can be beneficial for research applications and for facilitating clinical care (3-4, 6-8, 11, 19).

Development of an effective classifier model that accurately detects and grades fibrosis in whole lung imaging faces mainly two challenges. To date, most computer-generated texture features have used only small areas of lung to categorise ILD patterns (14-18). When applied to the whole lung, the computer-based model depending upon the intensities of pixels tends to

misclassify anatomical structures such as airways, fissures, and vessels or lung abnormalities into ground glass opacity or pulmonary fibrosis or other abnormalities. Another challenge is to obtain the sufficient image data with semi-quantitative scoring by expert radiologists for evaluation of whole lung fibrosis.

This paper reports the development of an automated fibrosis classifier and quantitative scoring system in whole lung imaging, which is then evaluated by comparison with semi-quantitative visual CT-based lung fibrosis scoring by expert radiologists in patients with scleroderma ILD.

Materials and methods

Patient selection

The Scleroderma Lung Study (SLS) was a multicentre NIH-sponsored randomised controlled trial comparing cyclophosphamide with placebo. The study was conducted between September 2000 and June 2006, involving 13 clinical centres throughout the United States (NCT 000004563, U01 HL60587-01A1, for detail, see Tashkin *et al*. (6), R01 HL072424). The use of anonymous image data from the clinical trial was approved by each local institutional review board. Briefly, baseline thoracic high-resolution (HR) CT was used to scan patients in the prone position at total lung capacity (TLC). Of these 158 randomised patients, 129 were analysed (Fig. 1, Table I). CT imaging of 29 patients could not be evaluated due to supine positioning (n=2), performance of CT scans outside of the protocol for routine clinical assessment (n=8), arms at side (n=1), 5mm-collimation (n=1), motion artifact and compromised image quality (n=2), or non-digitised format of CT images (n=15). Lung segmentation was of diagnostic quality in all evaluable cases. CT images were acquired from 4 manufacturers (Elscint, Haifa, Israel; General Electric, Milwaukee, USA; Picker International Inc., Highland Heights, USA; Siemens, Munich, Germany). The radiation exposure parameters ranged from 80 to 380mAs (mean of 245mAs±79) and the peak tube current potentials ranged from 120 to 140kVp. Non-volumetric CT scans of 1–2 mm slice thickness were acquired at 10mm increments and were typically reconstructed with sharp or over-enhancing reconstruction filters.

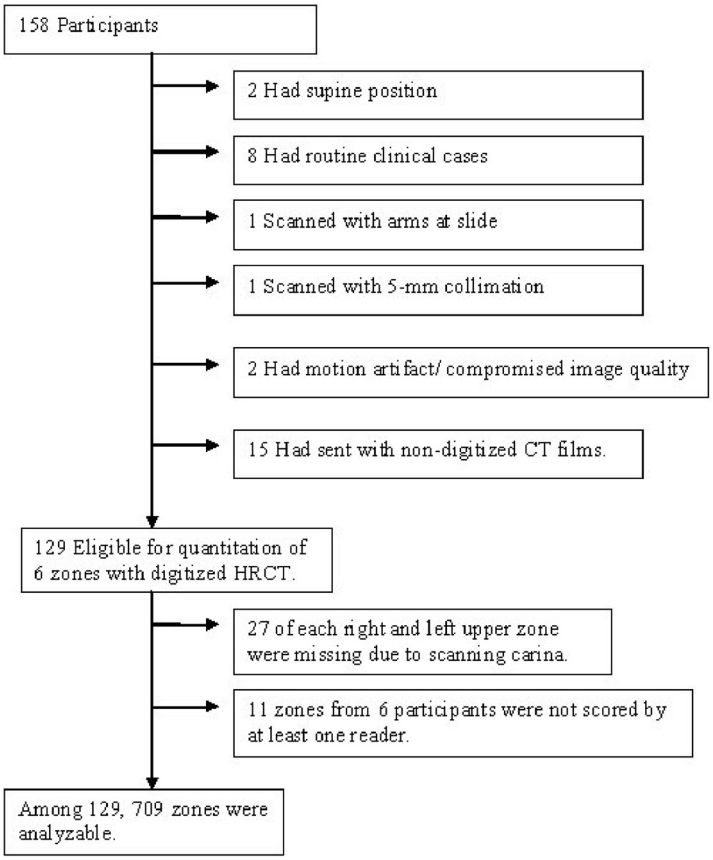

Fig. 1. Flow chart of patients

Table I. Patient characteristics.

n=129	Mean (SD)	Range
Age, yrs	48.8 (12.1)	22.3–83.1
Female sex (% of patients)	93 (72.1%)	
Duration of scleroderma, yrs	3.0 (2.1)	0.05–12.0
FVC (% of predicted)	68.3 (11.8)	29.4–90.5
FEV_1/FVC (%)*	83.2 (6.8)	61.0–99.0
Total lung capacity (% of predicted)#	69.5 (13.2)	24.0–100.0
Residual volume (% predicted)#	69.0 (25.9)	9.0–166.0
DLCO (% of predicted)	46.6 (14.0)	17.0–100.0
Cough, n. (% of patients)##	88 (70.4%)	
Focal score for the Mahler Dyspnea Index+	5. (1.8)	0–10.0
Skin-thickening score**	14.2 (10.5)	0–45

*n=126, #n=127, ##n=125, +n=122, **n=128; for all other characteristics, n=129.

Semi-quantitative lung fibrosis score

As part of visual assessment in SLS that has been published previously, two SLS thoracic radiologists (DAL and DCS) with 21 and 16 years experience assessed the CTs for extent of pure ground glass opacity (pGGO), pulmonary fibrosis (PF), honeycomb cysts (HC), and emphysema (6, 7). In this study, we only emphasise evaluation of PF using the CAD system for reasons listed below. In the visual assessment, each of interstitial lung disease component was scored using Likert scale semi-quantitative scores, ranging 0–4 (0=absent, 1=1–25%, 2=26–50%, 3=51–75%, and 4=76–100% extent of involvement) in three lung zones (upper, middle and lower) in a blinded fashion (20) (for detail, see Tashkin *et*

al. (6) and Goldin *et al*. (7)). The upper zones covered the apices to the aortic arch, the middle zone spanned from the aortic arch to the pulmonary veins, and the lower zone started at or below the pulmonary veins. As subjects from one site (n=27) were only scanned below the carina, the upper zones in these cases were not evaluated. Zones degraded by breathing artifact or limited image quality were not scored.

We focused on PF (*i.e*. reticular pattern with architectural distortion) in whole lung evaluation for the following reasons: 1) Good agreement for the presence or absence of visually scored PF had been found between the expert readers (7) whose scores were used in the present study for evaluation of the computer-based scoring system; 2) only very few cases of emphysema (1.2%) were visually noted by either reader; 3) only fair interobserver agreement for visually scored HC was noted, thus failing to provide a good "truth" for CAD evaluation; and 4) poor inter-reader agreement was found between pGGO and GGO with or without associated PF (so-called "any GGO", an eligibility criterion for the SLS), thus again making it difficult to establish "truth" for the computer-assisted classification where CAD GGO indicates any GGO. Semi-quantitative fibrosis (semi-QLF) score was defined as each radiologist's visual PF score using a 5-point Likert scale when both radiologists registered non-missing scores. Zones from six participants were partially excluded (*i.e*. not scored by at least one radiologist) due to nondiagnostic images in the upper zones (n=3), right middle zone (n=2), or remaining three zones (n=1) (Table II).

Table II. Summary statistics of marginal distributions of visual semi-Quantitative Lung Fibrosis (QLF) scores and Computer-Aided Diagnosis (CAD) QLF scores.

		Statistics				
Zone (Total n=129)	Likert Scale of fibrosis Score	Reader 1 semi-QLF Scores	Reader2 semi-QLF Scores	CAD QLF Scores		
		n	n	n	Mean ± SD	(min, max)
Right Upper	0 = <1%	60	34	31	0.53 ± 0.26	(0.10, 0.99)
	1 = (1-25%)	39	55	65	4.76 ± 4.68	(1.00, 23.82)
	2 = (26-50%)	0	10	3	35.20 ± 3.10	(33.15, 38.77)
	3 = (51-75%)	0	0	0	*NA*	*NA*
	4 = (76-100%)	0	0	0	*NA*	*NA*
	missing*	30	30	30	*NA*	*NA*
Right Middle	0 = <1%	44	33	23	0.60 ± 0.24	(0.20, 0.99)
	1 = (1-25%)	70	64	97	5.87 ± 4.52	(1.04, 20.90)
	2 = (26-50%)	12	26	7	36.54 ± 6.96	(28.48, 45.50)
	3 = (51-75%)	1	4	0	*NA*	*NA*
	4 = (76-100%)	0	0	0	*NA*	*NA*
	missing	2	2	2	*NA*	*NA*
Right Lower	0 = <1%	22	16	12	0.56 ± 0.31	(0.11, 0.98)
	1 = (1-25%)	56	32	72	10.92 ± 7.57	(1.04, 25.47)
	2 = (26-50%)	31	31	32	36.51 ± 6.88	(25.95, 48.91)
	3 = (51-75%)	17	43	9	64.24 ± 4.69	(56.92, 69.72)
	4 = (76-100%)	2	6	3	81.40 ± 6.76	(76.68, 89.15)
	missing	1	1	1	*NA*	*NA*
Left Upper	0 = <1%	58	36	40	0.49 ± 0.28	(0.00, 0.93)
	1 = (1-25%)	41	53	57	4.49 ± 3.42	(1.10, 18.17)
	2 = (26-50%)	0	10	1	30.12 ± .	(30.12, 30.12)
	3 = (51-75%)	0	0	1	54.18 ± .	(54.18, 54.18)
	4 = (76-100%)	0	0	0	*NA*	*NA*
	missing*	30	30	30	*NA*	*NA*
Left Middle	0 = <1%	41	36	11	0.63 ± 0.28	(0.00, 0.99)
	1 = (1-25%)	80	63	111	6.59 ± 5.01	(1.02, 24.85)
	2 = (26-50%)	7	26	5	31.93 ± 5.50	(26.21, 40.18)
	3 = (51-75%)	0	3	1	63.07 ± .	(63.07, 63.07)
	4 = (76-100%)	0	0	0	*NA*	*NA*
	missing	1	1	1	*NA*	*NA*
Left Lower	0 = <1%	22	13	5	0.50 ± 0.31	(0.00, 0.84)
	1 = (1-25%)	52	31	81	12.02 ± 7.73	(1.21, 25.00)
	2 = (26-50%)	35	35	24	37.64 ± 6.99	(27.96, 49.21)
	3 = (51-75%)	19	42	15	59.20 ± 5.46	(51.83, 67.99)
	4 = (76-100%)	0	7	3	82.84 ± 3.05	(79.39, 85.17)
	missing	1	1	1	*NA*	*NA*
total	0 = <1%	247	168	122	0.54 ± 0.27	(0.00, 0.99)
	1 = (1-25%)	338	298	483	7.51 ± 6.36	(1.00, 25.47)
	2 = (26-50%)	85	138	72	36.43 ± 6.72	(25.95, 49.21)
	3 = (51-75%)	37	92	26	60.90 ± 5.61	(51.83, 69.72)
	4 = (76-100%)	2	13	6	82.12 ± 4.76	(76.68, 89.15)
	All	709	709	709	11.84 ± 16.12	(0.00, 89.15)
	missing	65	65	65	*NA*	*NA*

*Total n=99; 27 subjects were scanned at carina and below instead of whole lung, 3 subjects were scored as missing at least by one radiologist due to breathing artifact and poor image quality.
NA: not applicable.

Small regions of data for CAD whole lung development

To effectively apply a PF classification from small regions of interest to the whole lung, we included normal anatomical structures from the LIDC in the classifier training and test data sets. The training set for the classification model was composed of 52 CTs: baseline CTs from consecutive SLS patients (n=38) and CTs from randomly selected patients from the Lung Image Database Consortium (LIDC) (n=14). From the SLS patients, 148 regions of interest (ROIs) exhibited classic, homogeneous and unambiguous features of scleroderma lung disease patterns and normal lung parenchyma, which were contoured by another thoracic radiologist (JGG, 12 years experience) (18). Regions included 46 PF, 85 GGO, 4 HC, and 13 normal lung (NL) patterns. From the LIDC data set, markings from 74 ROIs were used to delineate PF and other abnormalities from anatomical components in non-

volumetric scans (21). The markings of each patient were chosen with the minimum distance of 3 slices. Regions from 15 airways (1st to 6th generation), 15 major fissures, 14 minor fissures and 30 vessels (hilum to peripheral) were included as NL (disease free). For assessing the classification ability of the model built on the training set, the test set was composed of 199 ROIs from 47 patients using identical criteria: 132 contoured ROIs from 33 independent SLS participants and 67 marked ROIs from 14 independent LIDC test set subjects. Test regions included 44 PF, 72 GGO, 4 HC, and 12 NL patterns, in addition to 67 NL regions that included 14 airways, 14 major fissures, 13 minor fissures, 13 hilar large vessels, and 13 small lung vessels.

CT image analysis and CAD classification model

– Development of a fibrosis classifier for whole lung using small ROI

In our upgraded classification model, we included the robust texture features extracted from cleaned images, oracle features selection, and a support vector machine (SVM) with few assumptions on data distribution and dependency (18, 22-25). Oracle feature selection was used to avoid over fitting by maximising a penalised likelihood function. The non-concave penalised likelihood function was composed of two parts: a regular likelihood function and a penalty function for adding the number of features. Logistic likelihood was used with NL as the reference group and smoothly clipped absolute deviation (SCAD) as the penalty function (22). Matlab, Version 7.3.0. (R2006b) was used. The model was extended for application to the entire lung field by including features from anatomical structures from the LIDC in the classifier training and test data sets. In the small ROI test set, classification of PF by CAD yielded 94.4% sensitivity and 94.7% specificity (Of note additional classifying model of PF including HC, "any" GGO, and all types of patterns in interstitial lung disease yielded sensitivities of 95.1%, 82.4%, and 95.3% and specificities of 96.8%, 98.0%, and 96.9%, respectively).

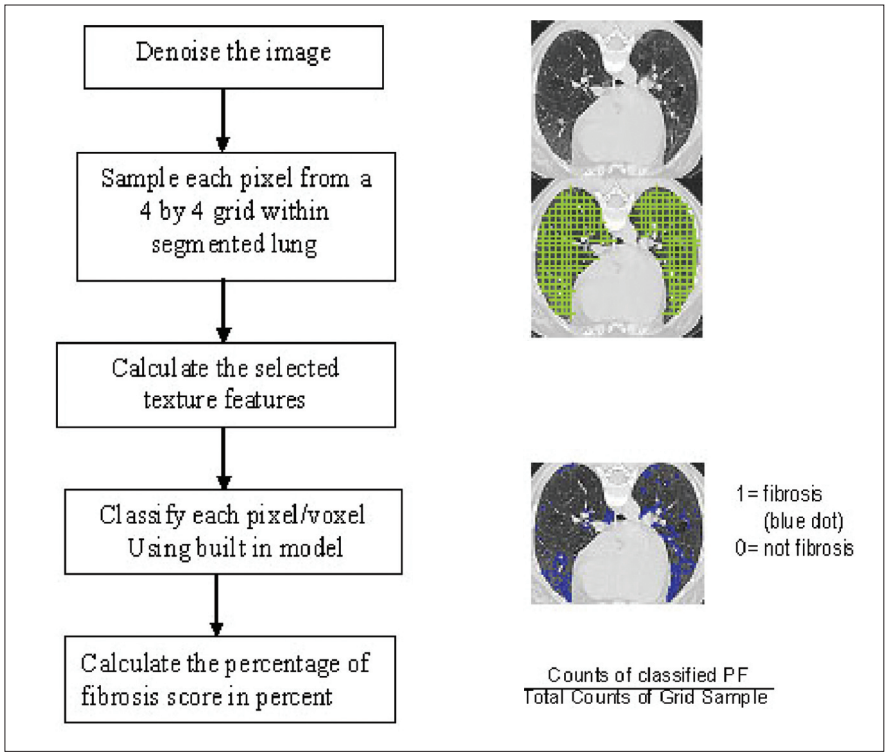

Fig. 2. Procedure for Automated Development of Computer-Aided Diagnosis (CAD) Quantitative Lung Fibrosis (QLF) Score.

Procedure for automated quantitative Fibrosis (QLF) scoring in whole lung

A semi-automated 3D lung segmentation program was applied (26), and the automated QLF scoring was run, consisting of the following steps:

1 *Cleaning (De-noising) the CT image*. To reduce variation of texture features across different scanners, Gilles' and Aujol's de-noise algorithm (23, 24) was implemented with the noise parameter based on the standard deviation (SD) of the aorta (18). Details of the algorithm are given in Appendix 1 and 2.

2 *Sampling each pixel from a 4-by-4 grid within segmented lung*.

3 *Calculating the texture features from de-noised CT image for the sampled pixel* (27, 28).

4 *Integrating database with the previously built SVM classifier to predict PF using the same selected texture features*. Features from Step (3) were used to predict PF or non-PF (*i.e*. NL, GGO and/or HC) by built-in classifier using SVM from R software version 2.2.1(The R Foundation for Statistical Computing, Vienna, Austria) with connecting an image work station. This integration between this classifier from R software to the image work station (JAVA language) was a key factor for the automated score.

5 *Calculating the PF score percentage by zones*. For comparison with semi-QLF zonal scores, we used the z-axis of the pixel location to register upper, middle, and lower zones. One-third of the total number of slices (*i.e*. maximum of z-axis – minimum of z-axis +1) were mapped to the upper, middle, and lower zones, respectively. When the upper zone data were not available, half of the total number of slices was mapped to middle and lower zones, respectively. The formula is below:

$$QLF = \frac{\text{Counts of classified PF}}{\text{Total Counts of Grid Sample}}$$

We used the five steps indicated in Figure 2 to develop an automated QLF score (Fig. 2).

Statistical analysis

Means (±SD) of QLF scores and counts of semi-QLF scores by each radiologist for each lung zone were reported.

Spearman rank correlations were used to compare continuous QLF scores and semi-QLF scores of the three zones in each lung. The linear mixed effects model was used to accommodate the dependency from six zones per subject in the comparison of QLF scores with semi-QLF scores (Appendix 3). In a sensitivity analysis, the Kappa (κ) statistics between two radiologists' semi-QLF scores were estimated to find the threshold in which the best agreement (highest kappa) was seen. This threshold was then chosen to evaluate CAD QLF scores. Receiver operating curve (ROC) analyses were also performed on the most agreed-on score categories by the two radiologists. We did the same analysis for the proposed clinically meaningful threshold of >25% (3-4). For determination of statistical significance, we took into account the potential intra-dependency of scores from the six different zones in each patient (29). Kendall's correlations between QLF score and PFT, physiological score, symptom scores were performed. Stata V.10.0 (College Station, Texas 77845 USA) and R Version 2.2.1 (The R Foundation for Statistical Computing, Vienna, Austria) were used for this analysis.

Fig. 3. Box-plot of Quantitative Lung Fibrosis (QLF) Scores (%) over visual scores using only the agreed-on scores by both radiologists (n=399 zones).

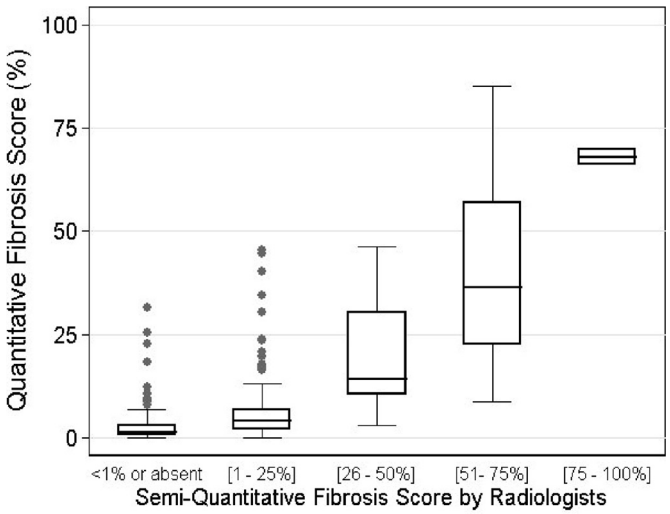

Table III. Correlation between semi-Quantitative Lung fibrosis (QLF) scores by readers and Computer-Aided Diagnosis (CAD) QLF scores.

	Spearman rank correlations		
Zone	Between two readers' scores	CAD QLF score & Reader One's score	CAD QLF score & Reader Two's score
Right Upper	0.62 (p<0.0001)	0.55 (p<0.0001)	0.54 (p<0.0001)
Right Middle	0.66 (p<0.0001)	0.53 (p<.0001)	0.58 (p<.0001)
Right Lower	0.67 (p<0.0001)	0.61 (p<.0001)	0.71 (p<.0001)
Left Upper	0.58 (p<0.0001)	0.45 (p<.0001)	0.50 (p<.0001)
Left Middle	0.54 (p<.0001)	0.28 (p=0.0012)	0.53 (p<.0001)
Left Lower	0.65 (p<0.0001)	0.50 (p<0.0001)	0.69 (p<0.0001)
Average of correlation	0.62	0.49	0.60

n=129 subjects.

Results

Lung fibrosis scores from two readers vs. CAD

The counts of semi-QLF scores by Likert scale from 0 to 4 were recorded within each zone (Table II). High Likert scores found in lower zones compared with upper zone, indicating moderate and severe extent of PF were located in lower zones. In the right upper zone, the number of zones in which the semi-QLF scores were zero were 60 for Reader One, and 34 for Reader Two, respectively, whereas the number of zones that had QLF scores <1% were 31 by CAD. The overall means (SD) of the two readers' semi-QLF scores using the Likert scale were 0.93 (±0.86) and 1.27 (±1.03), and the mean QLF score by CAD was 11.84% (±16.12). Figure 3 shows the box plots of QLF CAD scores by the visual Likert scores only in the cases for which the two radiologists agreed. The QLF scoring systemis sensitive for detecting PF when the semi-QLF scores are either zero or 1 (*i.e.* range of 0–25%), but relatively underestimate PF when the semi-QLF scores are ≥2. Of 146 zones with 0 on the Likert scale (*i.e.* no or <1% PF), the median QLF score was 1.25%, indicating that more than half of 146 zones had greater QLF scores than 1%. Of 55 zones with ≥2 on the Likert scale, the majority of QLF scores were lower than the corresponding range. When the visual semi-QLF scores were 2 (*i.e.* range of 26–50%), the mean (SD) CAD QLF scores were 19.0% (±13.6). The association between QLF scores and visual semi-QLF scores was significant (p<0.001) based on the model fit of the linear mixed effects model.

Correlations between CAD and each of the two readers by zones

The Spearman rank correlations were determined for each of the 6 zones between QLF scores and each of the reader's semi-QLF scores (Table III). Correlations between readers (0.54 to 0.67) and between each of the readers and the QLF scores (0.28 to 0.61 and 0.50 to 0.71) were comparable and significant (all eighteen p-values <0.002).

Evaluation

At the 1% threshold, substantial and moderate agreement occurred in all six zones between the two radiologists (bootstrap κ=0.59, 95% CI=(0.52, 0.65)), whereas slightly less agreement occurred at the 25 % threshold (bootstrap κ=0.49, 95% CI=(0.43, 0.56)). Agreement between the two radiologists in semi-QLF scoring decreased progressively with an increase in the threshold from 1% to 75%. ROC analyses were performed both for the threshold with the best agreement (1% or above 1%), and for the clinically

meaningful threshold for PF (above 25%) (3-4). AUC is depicted by defining "truth" from the interpretation of each of the two radiologists and by assessing only agreed-on cases by the two radiologists using these two thresholds (Table IV). For the 1% threshold, the AUCs were 0.80 and 0.83 for each of the two radiologists and 0.86 for the agreed-on cases; for the 25% threshold, the AUCs were 0.90 and 0.91 for each of the radiologists, and 0.96 for the agreed-on cases. The ROC plot for agreed-on cases shows an AUC of 0.86 for the 1% threshold (*i.e.* visual score ≥1) and an AUC of 0.96 for the 25% threshold (*i.e.* visual score ≥2) (Fig. 4). The QLF scores showed good agreement with the corresponding HRCT images in most of cases (Fig. 5 A-B: visual score of 1=QLF score near 5%, Fig. C-D: visual score of 2=QLF score near 30%), but in a few cases the QLF scores varied from being higher than the visual semi-QLF scores in mild PF, and lower than the semi-QLF scores in moderate to severe PF (Fig. 5 E-F: visual score of 0 vs. QLF near 5%, Fig. G-H contains streak artifact: visual score of 0 *vs*. QLF near 5%, and Fig. I-J: outlying disagreed case, visual score of 2 *vs*. QLF near 5%, respectively). Several outlying zones (10/146), which were scored by both radiologists as zero (meaning non-PF lung) were registered as minimal PF by the QLF scoring system (Fig. 3, Fig. 5 E-F, and G-H).

Table IV. Area Under Curve (AUC) from ROC analysis of visual semi-Quantitative Lung fibrosis (QLF) Score on CAD QLF score.

Threshold of semi-QLF Score in ROC analyses	AUC (95% CI)
Reader One ≥1 (n=709 zones)	0.83 (0.80, 0.86)
Reader Two ≥1 (n=709 zones)	0.80 (0.75, 0.85)
Reader One ≥1 and Reader Two ≥1 (n=594 zones)	0.86 (0.83, 0.89)
Reader One ≥2 (n=709 zones)	0.90 (0.84, 0.92)
Reader Two ≥2 (n=709 zones)	0.91 (0.86, 0.96)
Reader One ≥2 and Reader Two ≥2 (n=576 zones)	0.96 (0.94, 0.98)

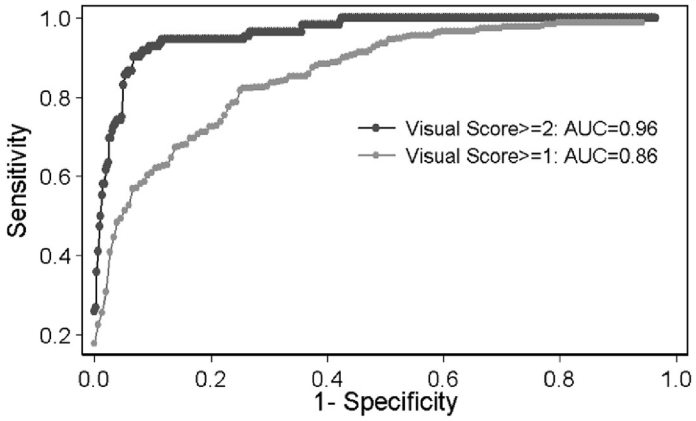

Fig. 4. ROC analysis of semi-Quantitative Lung fibrosis (QLF) Score on CAD QLF scores for the cases agreed-on by the two radiologists (n=594 for 1% threshold A≥1 and B≥1; n=576 for 25% threshold A≥2 and B≥2).

Correlations between CAD and pulmonary function test, other physiological measurements

Significant inverse associations were found between severity of whole lung CAD QLF and pulmonary function measurements of FVC (-0.31; $p<0.001$), TLC (-0.34; $p<0.001$), RV (-0.22; $p=0.0003$), DLCO (-0.35; $p<0.0001$), and FEV_1 (-0.23, $p=0.0001$). Severity of cough and frequency of cough were associated positively with severity of QLF score (0.22; $p=0.0017$) and (0.19, $p=0.02$), respectively, as well as with dyspnea in the domains of magnitude of task (0.16; $p=0.02$) and magnitude of effort (0.17; $p=0.01$). Insignificant correlations were found between skin score and the Health Assessment Questionnaire and whole lung CAD QLF score.

Discussion

We have shown that automated CAD-based scoring systems of PF can be developed using data from a multicenter clinical trial to assess the whole lung rather than limited regions of the lung and that QLF scoring has high discerning ability for detection of PF, as well as for the recently proposed clinically meaningful threshold of 25% (0.96 AUC) for predicting mortality (3). The present work evaluates PF quantification of the entire lung rather than smaller regions described in previous systems (18). In this study, we extended the previous classification model by including vessel, fissures, and airways and implementing a novel classification model within the CAD system. Moreover, with visual scores of whole lung from two independent expert radiologists, who had not served for contouring small regions of interests as part of developing CAD model, we evaluated the agreement of the findings from the CAD-based scoring system in a large number of participants from the Scleroderma Lung Study.

The CAD system of whole lungs involves two major processes to detect and quantify abnormalities. Detection is based on pixel classification from a methodological model, while the quantification is a simple but powerful bookkeeping operation that assesses large image data sets. The visual detection rate for lung pathology increases with knowledge and experience, whereas CAD can improve this rate as soon as it is applied. Concerning quantification, visual quantification is associated with intra- and inter-observer variation, especially in non-cubical or non-ellipsoidal topology, such as the thoracic cage. The scoring of pulmonary fibrosis has been hampered by intra- and inter-reader variation (9-10). When CAD is applied in well-segmented lung regions and is developed with input from experienced radiologists, "truth" may significantly improve CAD's ability to classify and quantify the extent of interstitial lung disease.

Quantification of whole-lung fibrosis faces challenges in both the development and evaluation of a classification model. Most regions in the training set were constructed from well-defined lung regions. In contrast, evaluation

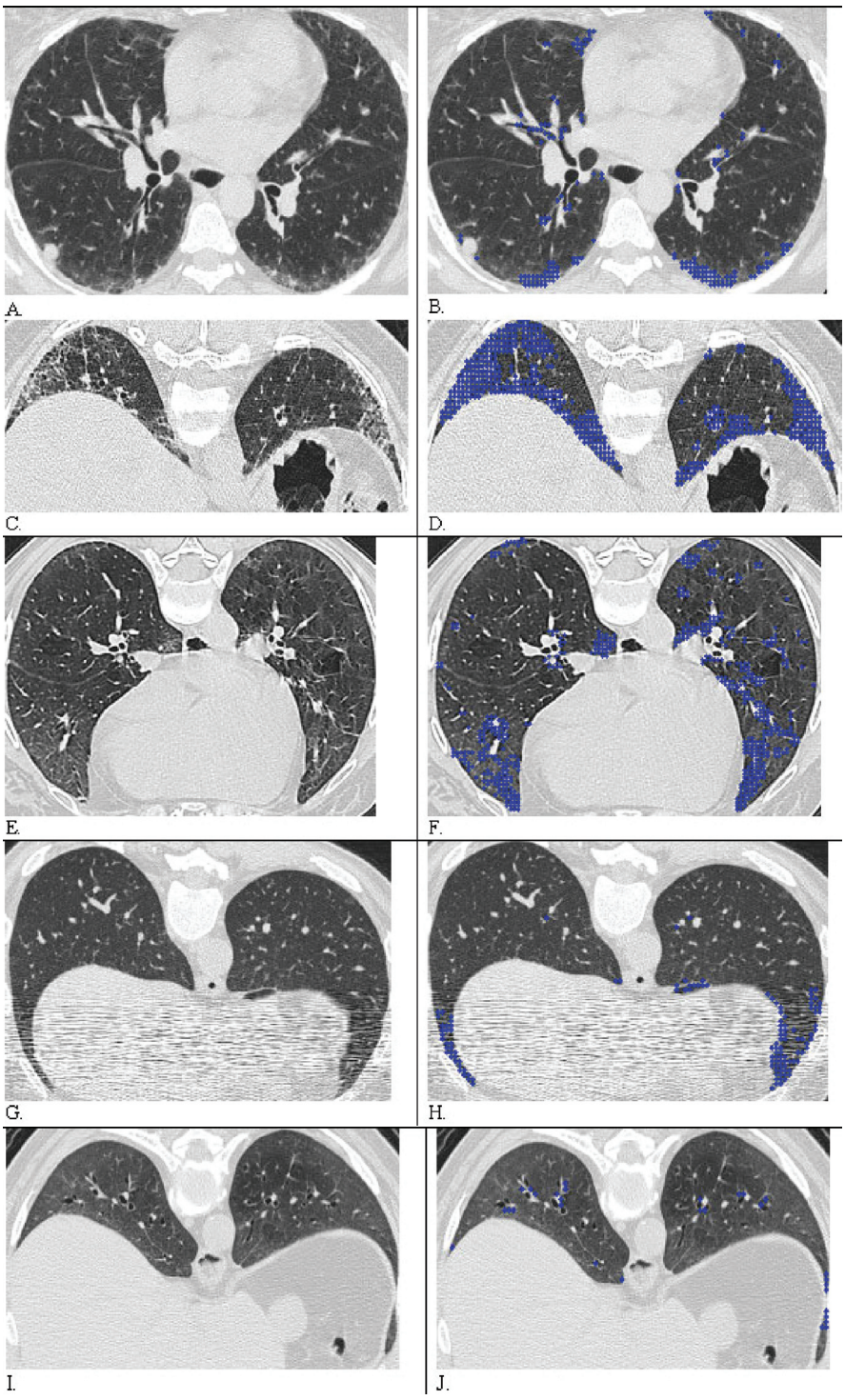

Fig. 5. Result of Automated Classification of Quantitative Fibrosis (QLF) and scores: A, C, E, G, and I were original CT images and are coupled with their overlaid images B, D, F, H, and J, respectively. Blue dots indicate classified Pulmonary Fibrosis (PF). **A**. Both radiologist scored as 1 = (1, 25%) in both zones. **B**. CAD Quantitative lung fibrosis (QLF) score were 4% and 5% of in both zones and agreed with visual semi-QLF score at 1% thresholds. **C**. Both radiologists scored as 2 = (26, 50%) in both zones. **D**. QLF score were 30% and 29% in the right and left zones. **E**. Both radiologists scored as 0 in right and left zones. **F**. QLF detected and scores were 4.4% and 6.0% of in the right and left middle zones, where bilateral peripheral fibrosis is detected in dependent lung. **G**. When CT images were degraded by streak artifact, two radiologists scored as 0 in both zones. **H**. De-noised CAD-based QLF score improved detection of PF as 5% and 6% of in the right and left lower zones. **I**. Both radiologists scored as 2= (26, 50%) in both zones. **J**. QLF score underscored PF as 5% and 4% of in both zones compared with both radiologists' scores. CAD classified the abnormal region as GGO when both radiologists might have scored the abnormality as PF.

of the whole lung includes lung parenchyma and additional anatomical structures such as vessels, fissures, and airways and a partial volume effect from the heart. Most lung segmentations do not perfectly separate these other anatomical components from the lung. For non-volumetric scan data with a 10mm gap between slices, we chose to address these confounding structural problems by adding the anatomical components into the classification model via the LIDC.

Comparison of the QLF scores on a continuous scale from the automated classifier algorithm with the ordinal scale of semi-QLF scoring by two thoracic radiologists is also challenging. Figure 3 shows that in this comparison, visual assessment systemically underestimated the presence (*i.e.* detection) and the amount of disease (*i.e.* quantification). The underestimation of the presence can be due to a) the broad range of Likert scale of 1 indicating 1–25%. (When the reader found a minimal PF, the reader may not score this as 1 unless the zone had a minimal amount of PF with clinical significance); and to b) noisy or degraded CT image. Whereas the CAD system is forced to calculate a score regardless of image quality, a radiologist can filter-out different types of noises and assign a score of no PF or determine that the scan is not-readable (*e.g.* Fig. 5. G and H). For moderate or severe cases, this underestimating phenomenon of a visual scoring *vs.* computer-based scores is not a new concept and has already been reported in studies involving scoring the extent of emphysema (30, 31). The underestimation might be due to the different approaches of summing computer-based scores versus the visual reader's subtracting the disease extent from 100%. Whereas the CAD system summed up at the pixel level in each slice, the visual readers scrolled up and down and found PF and/or started from a representative PF across slices from the zone and subtracted the amount. Thus, the QLF scoring evaluation requires the utilisation of well agreed-on cases between radiologists (1% threshold) and clinically meaningful guidelines

(25% threshold) (3-4); the latter was the threshold predictive of a therapeutic response to cyclophosphamide in the SLS (6) and is close to the 20%–30% threshold predictive of patients' survival (3). Between-reader agreement in semi-QLF scoring was fair in the upper lung zones and moderately good in the middle and lower zones.

We have shown that a CAD-based scoring system of PF can be performed and evaluated against visual scoring by highly experienced radiologists and has several advantages. First, our QLF model was sensitive in detecting mild PF and QLF scores were appropriately conservative by not overestimating PF in more severely affected areas (Fig. 3). Both radiologists showed good correlations in Likert scale for detection of PF (score of ≥1). However, QLF scores were better correlated with the detection of PF (score of ≥1) by Reader Two than Reader One (Table III). It seems that Reader Two was sensitive in detecting minimum PF, while Reader One was conservative in detecting and scoring PF (Table II). From detection to quantitation, Wells *et al*. have suggested that it may be more clinically relevant to discriminate between those cases with a visual semi-quantitative PF score >25% *versus* ≤25% level (rather than simply the presence or absence of PF) since they demonstrated that subjects with an extent of disease of 20–30% are at a higher risk of mortality than those with less extensive PF (3-4). In our study, the AUC showed significantly greater accuracy between QLF scores and the semi-quantitative scores at the 25% threshold (95 CI% (0.94, 0.98)) than at the 1% threshold (95 CI% (0.83, 0.89)). Thus, our classifier should be applicable to an assessment of the extent of PF as a predictor of mortality risk.

A second advantage of the CAD-based PF measurement is that it uses a continuous percentage scale, rather than a categorical Likert scale. As a result, the CAD system can provide higher statistical power for detecting the extent of PF on the HRCT scan (32). In a future study, we will address the sensitivity of changes in QLF score over time in the presence and absence of therapeutic intervention as a necessary validation step.

A third advantage is that the CAD is reproducible and traceable on CT images. The system shows regions that are classified as PF, as in Figure 5, which may be visually confirmed for accuracy. Additionally, texture features from de-noised images may be a potential way to reduce noise variation considering the effect of HU measurements that may vary across different scanners, kernel reconstructions, and exposure parameters (16, 18). Even in images with streak artifact and a semi-QLF score of 0, QLF scoring can identify PF (Fig. 5G-H), thus obviating the problem of obscuring of PF by streak artifact in visual scoring.

There were five main limitations of this study. First, the CT image scores are non-anatomical, and the registration of lung zone may differ slightly from the zone visually identified by the radiologists. While the radiologists used anatomical landmarks to define each zone for semi-QLF scoring, QLF scoring evenly divided each axial image into equal thirds. The worst correlation between QLF scores and visual scores occurred in the left middle lung zone, whereas the correlations were consistent in the upper and lower zones (Table III). A second limitation was the use of non-volumetric CT data with a 10mm gap between slices. While the CAD system can only analyze scanned slices, radiologists may impute a score between slices. A third limitation was that our training and test data sets of abnormal patterns were from a single clinical trial, the Scleroderma Lung Study. We are currently planning to apply PF classification to a new clinical trial of scleroderma ILD for validation (33). Another possible limitation was that the QLF scores may overestimate PF due to breathing artifact, partial volume effect or cardio-respiratory motion. Lastly, we did not have a visual assessment of overall extent of interstitial lung involvement (any ground glass + reticular changes + honeycombing), so that we could not evaluate the overall extent of interstitial lung involvement by CAD in comparison with that assessed visually.

Conclusion

We have developed a 5-step automated classifier of whole lung fibrosis in patients with scleroderma interstitial lung disease using HRCT. Our technique exhibited good accuracy for detecting fibrosis at a threshold level of both 1% (*i.e*., presence or absence of PF) and at a clinically meaningful threshold of 25% extent of fibrosis. Our findings suggest that this automated classifier is potentially useful for reproducible objective measurements of fibrosis in clinical trials of interventions in ILD.

Appendices

Appendix 1

Gilles' and extension of Aujol's Algorithm (23-24)**:**

1. Initialisation:
   $u_0 = v_0 = 0$

2. Iterations:
   $w_{n+1} = P_{\delta B_G}(f - u_n - v_n))$
   $v_{n+1} = P_{\mu B_G}(f - u_n - w_{n+1})$
   $u_{n+1} = f - v_{n+1} - w_{n+1} - P_{\lambda B_G}(f - v_{n+1} - w_{n+1})$

3. Stopping test: we stop if
   $\max(|u_{n+1} - u_n|, |v_{n+1} - v_n|, |w_{n+1} - w_n|) \leq \varepsilon$,

where *u, v* and *w* represents the geometric, texture, and noised images, respectively. And the sum of *u* and *v* image is denoised image. P_{B_G} is a non-linear projection described in appendix 2, and δ represents the amount of noise and λ represents the accuracy of algorithm. The sum of *u, v,* and *w* is approximately equal to original CT image if the algorithm converges. Here, for the sake of simplicity and consistency, we set the noise parameter (δ) as 50 and texture parameter (μ) as 450, which were the upper bound of standard deviation in aorta and in CT image across patients. Because the parameter has a certain threshold, the results of denoised images are similar to the values above the threshold. The residual parameter (λ) was set to 1, which controls the convergence of the algorithm.

Appendix 2

Any computerised image can be digitalised into N by N vectors. And each element of a matrix is a pixel. We denote by X by Euclidian space R^{NxN} and de note Y=X×X. In CT image, the window size is 512 by 512.

Projection

Each element of P, projection matrix is below. And it was solved by a fixed point method (23):

$$p^0=0 \text{ and}$$

$$p^{n+1}_{i,j} = \frac{p^n_{ij} + \tau(\nabla(\mathrm{div}(p^n) - f / \lambda))_{i,j}}{1 + \tau \mid (\nabla(\mathrm{div}(p^n) - f / \lambda))_{i,j}}$$

Theoretically, this projection converge $\tau \leq 1/8$. Practically, the author used ¼ and he stated that it worked better (23).

Gradient operator

Defining a discrete total variation, they introduced a discrete version of the gradient operator. If $u \in X$, the gradient ∇u is a vector in Y given by: $(\nabla u)_{i,j} = ((\nabla u)^1_{i,j}, (\nabla u)^2_{i,j}))$.

Using

$$(\nabla u)^1_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } i < N \\ 0 & \text{if } i = N \end{cases}$$

and

$$(\nabla u)^2_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } j < N \\ 0 & \text{if } i = N \end{cases},$$

Divergence operator

They defined it by analogy with the continuous setting by div = - ∇^*, where ∇^* is the adjoint of ∇: that is, for every $p \in Y$, and $u \in X$, $(-\mathrm{div}\ p, u)_X = (p, \nabla u)_Y$.

$$(\mathrm{div}(p))_{i,j} = \begin{cases} p^1_{i,j} - p^1_{i-1,j} & \text{if } 1 < i < N \\ p^1_{i,j} & \text{if } i = 1 \\ -p^1_{i-1,j} & \text{if } i = N \end{cases} + \begin{cases} p^2_{i,j} - p^2_{ij-1} & \text{if } 1 < j < N \\ p^2_{ij} & \text{if } j = 1 \\ -p^2_{ij-1} & \text{if } j = N \end{cases}$$

Appendix 3

Due to the dependency of six zones per subject, the mixed effects model was used. The automated computer-aided diagnosis (CAD) quantitative lung Fibrosis (QLF) score was the response variable. The 5 Likert scales of the ordinal semi-QLF scores were used as 4 dichotomised fixed-effect regressors with the reference group having a zero score and subjects and the zones that nested to the subject being used as random-intercept and random-slope (coefficient) in the model. The regression model is expressed for subject *i* and zone *j* as below:

$$\text{CAD QLF Score}_{ij} = \sum_{p=1}^{4} \beta_p \text{semi-QLF Score}_{p\,ij} + \sum_{q=1} b_{i\,q}\, \text{subject}^{q}_{ij} \mid \text{zones}_{ij} + \varepsilon_{ij}$$

$$b_i \sim N_q(0, \Psi)$$
$$\varepsilon_i \sim N_{ni}(0, \sigma^2 \Lambda_i)$$

where β is fixed-effect coefficient, b_i is random-effect coefficient for subject i, ε_{ij} is error term of subject and zone. Ψ is the 6 × 6 covariance matrix for the random effects. $\sigma^2\Lambda_i$ is the $n_i \times n_i$ covariance matrix for the errors in subject *i*. The small counts in semi-QLF scores at 4 were excluded in the final regression model to avoid influential points, although including them did not change the overall conclusion.

References

1. CONTE C, OWENS GR, MEDSGER TA JR: Severe restrictive lung disease in systemic sclerosis. *Arthritis Rheum* 1994; 37: 1283-9.
2. KARASSA FB, IOANNIDIS JP: Mortality in systemic sclerosis. *Clin Exp Rheumatol* 2008; 26 (Suppl. 51): S85-93.
3. GOH NS, DESAI SR, VEERARAGHAVAN S *et al*.: Interstitial lung disease in systemic sclerosis: a simple staging system. *Am J Respir Crit Care Med* 2008; 177: 1248-54
4. WELLS AU, BEHR J, SILVER R: Outcome measures in the lung. *Rheumatology* (Oxford) 2008; 47: v48-50.
5. SAHIN H, BROWN K, CURRAN-EVERETT *et al*.: Chronic hypersensitivity pneumonitis: CT features–comparison with pathologic evidence of fibrosis and survival. *Radiology* 2007; 244: 591-8.
6. TASHKIN DP, ELASHOFF R, CLEMENTS PJ *et al*.: Cyclophosphamide versus placebo in scleroderma lung disease. *N Engl J Med* 2006; 354: 2655-66.
7. GOLDIN JG, LYNCH DA, STROLL DC *et al*.: High-resolution CT scan findings in patients with symptomatic scleroderma-related interstitial lung disease. *Chest* 2008; 134: 358-67.
8. STRANGE C, SEIBOLD JR: Scleroderma lung disease: "if you don't know where you are going, any road will take you there". *Am. J. Respir Crit Care Med* 2008; 177: 1178-9
9. COLLINS CD, WELLS AU, HANSELL DM *et al*.: Observer variation in pattern type and extent of disease in fibrosing alveolitis on thin section computed tomography and chest radiography. *Clin Radiol* 1994; 49: 236-40.
10. CAMICIOTTOLI G, ORLANDI I, BARTOLUCCI M *et al*.: Lung CT densitometry in systemic sclerosis: correlation with lung function, exercise testing, and quality of life. *Chest* 2007; 131: 672-81.
11. LYNCH DA: Quantitative CT of fibrotic interstitial lung disease. *Chest* 2007; 131: 643-4.
12. MÜLLER NL, STAPLES CA, MILLER RR, ABBOUD RT: "Density mask". An objective method to quantitate emphysema using computed tomography. *Chest* 1988; 94: 782-7.
13. GEVENOIS PA, DE MAERTELAER V, DE VUYST P, ZANEN J, YERNAULT JC: Comparison of computed density and macroscopic morphometry in pulmonary emphysema. *Am J Respir Crit Care Med* 1995; 152: 653-7.
14. CHABAT F, YANG GZ, HANSELL DM: Obstructive lung diseases: texture classification for differentiation at CT. *Radiology* 2003; 228: 871-7.
15. UPPALURI R, HOFFMAN EA, SONKA M *et al*.: Computer recognition of regional lung disease patterns. *Am J Respir Crit Care Med* 1999; 160: 648-54.
16. XU Y, VAN BEEK EJ, HWANJO Y *et al*.: Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). *Acad Radiol* 2006; 13: 969-78.
17. ZAVALETTA VA, BARTHOLMAI BJ, ROBB RA: High resolution multidetector CT-aided tissue analysis and quantification of lung fibrosis. *Acad Radiol* 2007; 14: 772-87.
18. KIM HJ, LI G, GJERTSON DW *et al*.: Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study. *Acad Radiol* 2008; 15: 1004-16.
19. 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: 935-40.
20. KAZEROONI EA, MARTINEZ FJ, FLINT A *et al*.: Thin-section CT obtained at 10-mm increments versus limited three-level thin-section CT for idiopathic pulmonary fibrosis: correlation with pathologic scoring. *Am J Roentgenol* 1997; 169: 977-83.
21. OCHS RA, GOLDIN JG, ABTIN F *et al*.: Automated classification of lung bronchovascular anatomy in CT using AdaBoost. *Med Image Anal* 2007; 11: 315-24.
22. FAN J, LI J: Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; 96: 1348-60.
23. AUJOL J., GILBOA G., CHAN T, OSHER S: Structure-Texture Image Decomposition-Modeling, Algorithm, and Parameter Selection. *Int J Comput Vis* 2006; 67: 111-36.
24. GILLES J: Noisy decomposition: a new structure, texture and noise model based on local adaptivity, *J Math Imaging Vis* 2007; 28: 285-295.
25. VAPNIK VN: The Nature of Statistical Learning Theory. New York, Springer, 2nd edition; 1999.
26. BROWN MS, MCNITT-GRAY MF, GOLDIN JG

et al.: Automated measurement of single and total lung volume from CT. *J Comput Assist Tomogr* 1999; 23: 632-40.

27. HARALICK RM: Statistical and structural approaches to texture. Proc IEEE 1979; 67: 786-804.
28. SONKA M, HLAVAC V, BOYLE R: Image processing, analysis and machine vision London, England: *Chapman & Hall*, 1993.
29. LI G, ZHOU KA: Unified approach to nonparametric comparison of receiver operating characteristic curves for longitudinal and clustered data. *J Am Stat Assoc* 2008; 103: 705-13.
30. ZOMPATORI M, BATTAGLIA M, RIMONDI MR *et al*.: Quantitative assessment of pulmonary emphysema with computerized tomography. Comparison of the visual score and high resolution computerized tomography, expiratory density mask with spiral computerized tomography and respiratory function tests, *Radiol Med* (Torino) 1997; 93: 374-81.
31. PARK KJ, COLLEEN JB, CLAUSEN JL: Quantitation of emphysema with three-dimensional CT densitometry: comparison with two-dimensional analysis, visual emphysema scores, and pulmonary function test results. *Radiology* 1999; 211: 541-7.
32. GOLDIN J, ELASHOFF R, KIM HJ *et al*.: Treatment of scleroderma-interstitial lung disease with cyclophosphamide is associated with less progressive fibrosis on serial thoracic high-resolution CT scan than placebo: findings from the scleroderma lung study. *Chest* 2009; 136: 1333-40
33. ALTAR CA, AMAKYE D, BOUNOS D *et al*.: A prototypical process for creating evidentiary standards for biomarkers and diagnostics, *Clin Pharmacol Ther* 2008; 83: 368-71.