
Alexander J Procter¹, Joseph Jacob²,³*

¹Department of Radiology, University College Hospital, London, UK
²Department of Respiratory Medicine, University College London, London, UK;
³Centre for Medical Image Computing, University College London, London, UK;

*=Corresponding author

Contact details: The Centre for Medical Image Computing, The Front Engineering Building, Floor 3, Malet Place, London, WC1E 7JE. Email: j.jacob@ucl.ac.uk. Tel: 0044 2031081061.

Conflicts of interest: Joseph Jacob reports personal fees from Boehringer Ingelheim and Roche unrelated to the submitted work.

Acknowledgements: Joseph Jacob was supported by Wellcome Trust Clinical Research Career Development Fellowship 209553/Z/17/Z.

Word Count (without ‘key points‘): 2,483
Abstract

Purpose of Review: Computer algorithms possess an intrinsic speed, objectivity, reproducibility and scalability unmatched by visual quantitation methods performed by trained readers. The question of how well quantitative CT (QCT) analysis methods compare to visual CT analysis to predict functional status in fibrosing lung diseases (FLDs) is of increasing relevance to understand the future role QCT may have in prognostication of FLD.

Recent Findings: The latest computer algorithms demonstrate improved performance over visual CT analysis in predicting baseline disease severity as measured by correlations with functional indices of lung damage. QCT analysis may therefore have a role in aiding clinical decision making as well as in the enrichment of drug trial populations. Quantitative analysis on longitudinal CTs has also shown better correlations with changes in functional indices when compared to visual scores of change suggesting the potential of QCT analysis as an imaging biomarker of disease progression in FLD. Importantly, computer algorithms are now able to identify prognostic imaging biomarkers that cannot be quantified visually (e.g. vessel-related structures).

Summary: QCT holds great promise for the evaluation of damage in FLD. Challenges for QCT include accommodating measurement noise from variation in CT acquisition techniques and developing patient-friendly visualisations of quantitative outputs.

Key Words

Quantitative CT analysis
Pulmonary function tests
Fibrosing lung disease
Introduction

High Resolution Computed Tomography (HRCT) of the chest is central to the multidisciplinary approach now advocated for the diagnosis of interstitial lung diseases (ILD) and in particular the fibrosing lung diseases (FLDs) [1–3]. Beyond diagnosis, interest has been growing in the role CT imaging may play as an imaging biomarker of disease severity in FLDs. Traditionally, semi-quantitative CT evaluation was achieved by visually scoring disease patterns with trained readers. However, visual CT analysis is time-consuming, and interpretation is associated with inter-observer variability. The advent of sophisticated computer tools where speed, objectivity, reproducibility and scalability are intrinsic to the analytic methods, has highlighted their potential as candidate prognostic imaging biomarkers.

Research into CT prognostication has principally focused on its application in one of two domains:

1) Providing a measure of disease burden in FLDs using a scan from a single time point.

2) As a longitudinal biomarker, quantifying disease progression on repeated CTs in an individual patient.

This review will discuss the role that visual and quantitative CT (QCT) analysis have in the characterisation of lung damage in FLD. We will highlight recent reports linking visual analyses and/or QCT analyses with pulmonary function test (PFTs), which have traditionally been used to monitor FLD severity and progression.
Visual versus Computer-based CT analysis

Visual CT analysis is typically performed by 2-3 thoracic radiologists. Protocols frequently involve dividing each lung into three lobes (upper, middle, lower) which are considered to contribute equally to total lung volume [4,5]. Readers quantify affected lung parenchyma using binary scores (i.e. fibrosis vs no-fibrosis) [6] or continuous scores of CT patterns: ground-glass opacification (GGO), reticulation, honeycombing, emphysema and the low attenuation component of a mosaic attenuation pattern (all typically estimated to the nearest 5%) [5]. Traction bronchiectasis severity is usually measured with an ordinal score [5].

Whilst a detailed discussion of the different QCT methods used in lung fibrosis evaluation is beyond the scope of this article, a broad overview is provided. Invariably all methods start with lung segmentation, which involves separating the lung parenchyma from the surrounding chest wall and mediastinal structures (Figure 1). This is commonly an automated step with manual intervention when required [4].

Histogram analysis

The earliest QCT methods constructed and analysed lung-density histograms [7]. Analogous to emphysema quantitation using Hounsfield unit thresholds to define low-density structures, thresholds for high-density structures were used to quantify fibrosis. However, as low attenuation structures e.g. traction bronchiectasis and honeycombing can develop secondary to fibrosis, and are strongly linked to poor outcomes, correlations of high-density thresholds with survival were poor [8]. The addition of other parameters (e.g. the distance from the pleural surface or size of low-attenuation areas) into whole lung histogram models has reduced the confounding effect imposed by coexistent emphysema and/or honeycombing [4,9].

Texture analysis

Most modern QCT methods employ texture analysis which evaluates structural and density patterns at a voxel level (Table 1a). These are more computationally demanding with computers learning to recognise CT patterns traditionally described by radiologists (e.g. normal lung, GGO, honeycombing, reticulation, emphysema) [6,7,10–13].
QCT readouts can be overlaid on CT images to improve visualisation of damage measured across sub-regions of the lung [10].

Current evidence

Visual and QCT analysis has utility in baseline and longitudinal assessment of all FLD patients, though research has focussed on idiopathic pulmonary fibrosis (IPF).

Baseline IPF assessment

QCT tools that assess disease severity from a baseline CT should allow identification of patients at higher risk of disease progression. Such a measure could be used as an adjunct in clinical decision making, for example to allow earlier referral for lung transplantation. Alternatively, it could identify patients that, if included in a drug trial, would be more likely to reach a trial endpoint. By selectively recruiting these patients, it may be possible to reduce requisite trial sample sizes, thereby reducing trial costs [6,14].

The Automated Quantitative System (AQS), developed at the Asan Medical Centre in Seoul, quantifies multiple texture subtypes within the lung: normal lung, GGO, emphysema, consolidation, reticulation and honeycombing, with the latter two patterns summed as fibrosis score. Almost all CT pattern subtypes were found to correlate with baseline forced vital capacity (FVC) and diffusing capacity for carbon monoxide (DLco). The strongest FVC correlations were with ‘reticular opacity’ \( (r=-0.435; \ p<0.001) \) and ‘fibrosis score’ \( (r=-0.392; \ p<0.001) \). The strongest DLco correlation was with ‘fibrosis score’ \( (r=-0.506; \ p<0.001) \). On multivariate analysis, only ‘reticular opacity’ remained an independent predictor of FVC decline \( (p=0.012; \ adjusted \ odds \ ratio \ 1.047) \). Receiver operating characteristic curve analysis demonstrated that the area under the curve for ‘reticular opacity’ was 0.641. The optimal reticular opacity extent cut-off value to predict stable FVC at 1-year follow-up was 22.05% of the lung (sensitivity, 50.0%; specificity, 81.4%; negative predictive value, 89.1%) [12].

CALIPER utilises texture-based analysis that labels voxel volumes as: GGO, honeycombing, reticulation, subtypes of low-attenuation and normal lung (Figure 1). CALIPER-derived estimates of ILD extent demonstrated stronger univariate correlations than visual ILD scores (FVC) or were at least comparable to visual ILD scores (DLco, and the composite physiologic index [CPI]) [5]. Studies using CALIPER to predict outcome in IPF have demonstrated that
honeycombing extent independently predicts mortality (hazard ratio 1.18; p=0.002), and that computer variables are more accurate at predicting survival than corresponding visual CT scores [15]. Interestingly, in patients with severe IPF, functional measures (DLco and CPI) were stronger predictors of survival than visual and QCT scores. However, PFTs were less powerful at predicting outcome in patients with less extensive disease. It is in this patient cohort that QCT measures are likely to prove most useful [14].

CALIPER also quantifies a non-traditional CT pattern, pulmonary vessels and associated structures (e.g. perivascular fibrosis), collectively termed vessel-related structures (VRS) which cannot be quantified by the human eye. VRS (previously termed PVV) demonstrated similar correlations to CALIPER ILD extent for FVC, DLco, and CPI (ILD-extent: FVC r=-0.64, DLco r=-0.56, CPI r=0.69; VRS: FVC r=-0.67, DLco r=-0.58, CPI r=0.72; p<0.0001 for all correlations) [5]. VRS also independently predicted mortality in IPF (hazard ratio 1.53; p<0.0001) [15]. Importantly, VRS has shown promise as a measure that could be used to enrich IPF drug trial populations. Patients with a VRS threshold >4.4% of lung volume were shown to be more likely to reach trial endpoints. Selectively including these patients in an IPF drug trial would reduce sample sizes by 26%. Importantly, patients selected using the >4.4% VRS threshold demonstrated an increased life expectancy and reduced rate of FVC decline if receiving antifibrotic medication compared to patients not receiving antifibrotics, suggesting that these patients still had modifiable disease [14].

Vascular quantitation in IPF has also been performed by the Ludwig Boltzmann institute in Graz, Austria (Figure 2). Total vessel volume expressed as a percentage of the total lung volume was shown to independently predict FVC (r=-0.52, p<1x10^{-6}), DLco (r=-0.35, p=3x10^{-5}), and CPI ( r=0.53, p<1x10^{-6}) [16]. On multivariate analysis adjusted for patient age, male gender, smoking status and CT slice thickness, both vessel volume and heterogeneity in vessel tortuosity were found to independently predict DLco (model r=-0.46), and CPI (model r=0.58) [16].

The Adaptive Multiple Features Method (AMFM) primarily examined summed quantitative radiological features e.g. ground-glass reticulation (GGR). GGR was the AMFM feature that demonstrated the strongest correlations with equivalent visually-scored CT features (r=-0.60; p<0.0001) [13].

The Data-driven Textural Analysis (DTA) tool (Figure 3) uses a machine learning approach to develop a unique library of discrete CT features. The DTA tool was trained to recognise CT features that correspond to ‘fibrotic’ or ‘non-
fibrotic’ lung parenchyma. When evaluating CTs in patients with IPF, DTA quantifies the amount of ‘fibrotic’ lung on single CT images, which allows a total lung fibrosis percentage to be estimated. DTA showed moderate correlations with average visual scores of lung fibrosis (r=-0.50; p<0.001) and improved correlations with percent-predicted FVC (r=-0.60, p<0.001) and DLco (r=-0.68, p<0.001) [6].

Longitudinal IPF assessment

QCT analyses on longitudinal CTs from the same patient could be sensitive measures of FLD progression. Currently PFTs are the primary method used to monitor FLD progression. Yet PFTs have deficiencies, for example wide normal ranges may mask small deteriorations. Additionally, genuine physiological deterioration captured as marginal annualised FVC declines (between 5-10%) can be challenging to distinguish from measurement variation. Adjudication of marginal change using a second measure (e.g. QCT) that could confirm genuine disease progression would be useful for guiding management, and could represent a composite end-point in future drug trials [14]. QCT longitudinal CT biomarkers could also be used clinically to measure treatment efficacy of anti-fibrotic medication or rationalise medication change.

Visual analysis of change on longitudinal CTs can employ one of two methods: features on CT pairs are scored independently and differences in scores compared; or both studies are viewed side-by-side and assigned a ‘progression score’ (usually a 5-point ordinal scale with the middle number representing stable disease). Interestingly when both methods were compared, on independent CT examination, no features were predictive of patient survival. However, assessment of ‘global visual change’ on side-by-side CTs did predict survival [11].

Several quantitative systems have examined longitudinal change in CT features. CT texture-derived total lung fibrosis (QLF) score was shown to be more effective in assessing disease burden than CT histogram indices (e.g. kurtosis) [10]. At baseline both QLF (p=0.59; p<0.0001) and kurtosis (p=0.56; p<0.0001) scores correlated with FVC. However, only QLF scores correlated with longitudinal changes in FVC (p=0.57; p<0.0001) [10]. Humphries et al similarly demonstrated that change in DTA’s binary ‘fibrosis/no-fibrosis’ score correlated significantly with longitudinal changes in DLco (p=-0.40; p<0.001) and FVC (p=-0.41; p<0.001) [6]. When change in GGR measured by the AMFM software was compared to FVC change, significant correlations were again identified (r=-0.25; p=0.01) [13].
Interestingly, in this study, correlations with FVC change were stronger using visual CT scores ($r=-0.30$; $p=0.002$) than using the AMFM tool [13].

QCT is sensitive to measurement noise which is accentuated when examining longitudinal CTs (Table 1b). Noise can arise from variations in CT scanner models, scan acquisition protocols and patient factors (e.g. consistency of breath hold) [17]. To evaluate the intrinsic noise associated with longitudinal CT measurements the DTA tool retrospectively evaluated the CTs of patients recruited into two drug trials. Change in fibrosis extent was compared to change in functional measures (FVC and DLco), exercise capacity and patient-reported questionnaires. The study demonstrated that an increase in fibrosis extent by 3.4% of lung volume could be used to define a minimum clinically important amount of fibrosis progression in IPF patients [18].

When assessing the ability of CT scores to predict changes in FVC, CALIPER variables outperformed visual CT scores (e.g. CALIPER ILD extent score: $r=0.73$; $p<0.0001$; visual ILD extent score: $r=0.40$; $p=0.001$) [19]. This is in keeping with other studies suggesting that visual CT scores are poor longitudinal markers. CALIPER analysis also demonstrated that VRS thresholds identified different poor outcome IPF patients to those patients crossing a 10% FVC change threshold. The weak correlations between FVC change and VRS change indicate that both variables represent important yet distinct measures of lung damage, suggesting that both measures could be integrated as co-endpoints in clinical trials [20].

Non-IPF fibrosing lung diseases

There is limited published work focusing on functional correlations of visual CT analysis and QCT variables in fibrosing lung diseases other than IPF. When CALIPER was used to examine CTs in a cohort of 135 consecutive patients with fibrotic hypersensitivity pneumonitis, CALIPER-derived scores of ILD extent correlated more strongly with PFTs than visual ILD extent scores: FVC (CALIPER $r=0.73$, visual $r=0.51$); DLco (CALIPER $r=0.61$, visual $r=0.48$); CPI (CALIPER $r=0.70$, visual $r=0.55$). The CT variable that correlated most strongly with restrictive functional indices was VRS: FVC $r=0.75$, DLCO $r=0.68$ and CPI $r=0.76$. Decreased attenuation lung quantified by CALIPER was found to be a better morphological measure of obstructive lung disease than equivalent visual scores as judged by relationships with total lung capacity (CALIPER $r=0.63$ and visual $r=0.12$), with all results maintained on multivariate analysis [21].
Future Challenges

Many challenges still remain with regard to optimising QCT evaluation in patients with FLD.

Image Acquisition

Measurement noise associated with variations in CT acquisitions are a constraint to QCT analysis [18]. It is conceivable that scanner manufacturers could produce CT reconstructions optimised for quantitative analysis that are generated alongside each clinical CT reconstruction and reduce inter-scanner noise. Alternatively, deep learning algorithms could be developed that harmonise the appearance of CTs acquired using different parameters. As measurement noise cannot be eradicated, factoring it into model performance, testing and optimisation on large varied datasets of FLD CTs will be critical. Whilst cross-centre collaboration and data sharing may produce appropriate datasets, a longer-term solution would be the curation of publicly-available heterogenous longitudinal FLD datasets that could provide a standardised validation mechanism of quantitative tools.

Image characterisation

There remain CT features that are not well categorised by computer-based methods. In the context of FLD, distinguishing emphysema from honeycombing remains a challenging prospect for computer tools. Visually-scored traction bronchiectasis has been shown in several studies to strongly link to PFTs [15], but as yet there are no automated methods to quantify traction bronchiectasis.

Image interpretation

PFTs and visual CT scoring are readily available to physicians working in most healthcare institutions across the world. Computer-based analysis as well as being less readily available, often utilises proprietary methodologies which may incur additional fees to healthcare providers. The novel quantitative metrics of lung damage developed using specific computer-based tools may not be as readily understood by patients, clinicians or researchers around the world, as current descriptors of parenchymal damage such as reticulation and honeycombing. This has the potential to fragment the clinical and research communities.

The challenge of interpretation is likely to increase as deep learning techniques are increasingly applied to FLD CTs and imaging biomarkers with no clear visual correlate become more commonplace. Identifying the biological
plausibility by which the features captured by the ‘black box’ reflects the biological processes responsible for lung damage is a necessary stage of the biomarker development pathway. Establishing appropriate visualisations of lung damage could be considered a necessary key component of future computer tools. An example of a technique to express detailed quantitative data into a format easily deciphered by the non-specialist in a busy clinic setting, is the Glyph, developed by the team at the Mayo Clinic Rochester (Figure 1) [22]. Originally applied to the CALIPER tool [23], a similar principle could easily applied to other computer-based quantitation methods.

Conclusion

Computer-based quantitation of CT imaging in FLD has begun to demonstrate improved performance in predicting baseline and longitudinal functional indices when compared to traditional visual CT analysis. Importantly, newer tools are starting to identify imaging biomarkers that cannot be quantified visually and are therefore expanding the repertoire of clinically important prognostic biomarkers in FLD. Challenges still need to be overcome, including accounting for the noise associated with variations in CT acquisition techniques, and developing patient-friendly visualisations of quantitative outputs. Nevertheless, computer-based quantitation holds great promise as a clinical and research tool in the evaluation of lung damage in FLD.

Key Points

1. Quantitative CT analysis demonstrates improved performance over visual CT analysis in predicting baseline disease severity in fibrosing lung diseases as measured by correlations with functional indices of lung damage.
2. Quantitative analysis of longitudinal CTs demonstrates improved correlations with changes in functional indices when compared to visual scores of CT change suggesting that analysis quantitative CT analysis has potential as an imaging biomarker of disease progression in FLD.
3. Computer algorithms are now able to identify and quantify CT features that cannot be quantified visually, suggesting the emergence of novel prognostic imaging biomarkers.
References


* The baseline CTs of IPF patients were assessed with a novel quantitative metric, only achievable using computer-based analysis. This study is the first to demonstrate how QCT methods could be used to successfully enrich IPF drug trial populations.


* This study assesses computer-based CT quantitation as a longitudinal biomarker for progression in IPF patients. Importantly the study highlights how clinically meaningful thresholds, which suggest disease progression, can be obtained from QCT analysis.


Table 1a: An overview of computer tools quantifying lung damage using texture analysis

<table>
<thead>
<tr>
<th>Analysis tool</th>
<th>Institution(s)</th>
<th>Brief Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMFM (Adaptive Multiple Features Method)</td>
<td>University of Michigan, Ann Arbor, MI, USA</td>
<td>Quantitatively labelled CT features. Trained on volumes of interest annotated on CTs by three expert radiologists [13].</td>
</tr>
<tr>
<td>AQS (Automated Quantification System)</td>
<td>Asan Medical Center, University of Ulsan, Seoul, South Korea</td>
<td>Pixels assigned to a disease pattern or normal lung. Experienced radiologists assigned labels to multiple sampled regions across images obtained from different CT scanners [12].</td>
</tr>
<tr>
<td>CALIPER (Computer-Aided Lung Informatics for Pathology Evaluation and Rating)</td>
<td>Mayo Clinic Rochester, Minnesota, USA</td>
<td>15x15x15 voxel volume units labelled based on traditional radiologically described features. Trained by thoracic radiologist consensus assessment of pathologically confirmed data sets [5].</td>
</tr>
<tr>
<td>DTA (Data-driven Textural Analysis)</td>
<td>National Jewish Health, Denver, Colorado, USA</td>
<td>Unsupervised feature learning constructs dictionary of lung CT patterns. Analysis performed on two-dimensional axial slices. Dictionary quantifies features using training images labelled by radiologists [6].</td>
</tr>
<tr>
<td>QLF (Quantitative Lung Fibrosis)</td>
<td>University of California in Los Angeles, Los Angeles, California, USA</td>
<td>Classical texture features assigned to sample voxels. Trained on datasets with radiologist labelled regions of interest [10].</td>
</tr>
</tbody>
</table>
Table 1b summarises the main differences between the visual and computer-based methods for CT quantitation.

<table>
<thead>
<tr>
<th>CT Analysis Method</th>
<th>Benefits</th>
<th>Negatives</th>
</tr>
</thead>
</table>
| **Visual**         | • Theoretically, easy to perform by anyone.  
                    • Less dependent on image acquisition factors e.g. reconstruction type, contrast administration, interspaced vs volumetric images.  
                    • Often performed on a lobar basis - less susceptible to underestimating the impact of severe fibrosis preferentially shrinking one lobe. | • In practice requires trained subspecialists.  
                    • Expensive – need lots of radiologist time.  
                    • Subjective:  
                      o Kappa value of 0.3 ($p=0.001$, $n=280$) found when scoring fibrosis extent [6].  
                      o Observer variation can compound when used for longitudinal change [4,18,24]. |
| **Computer-based** | • Once trained, algorithm can be quick – easily scalable to thousands of CT examinations [14].  
                    • Cheaper in the long-term than training humans.  
                    • High precision.  
                    • Can identify new pattern/features not appreciated by humans [14]. | • Often assesses parenchymal patterns, as a proportion of the total lung volume – risks underestimating extent of disease [21].  
                    • Potentially instilled with the biases of the radiologists who trained the algorithm [11].  
                    • Inhomogeneity of CT acquisition protocols and scanner technologies [12] can result in the misclassification of honeycombing, reticulation, or ground-glass opacities [14].  
                    • Poor quality scans (e.g. those affected by motion or inadequate inspiration) often have to be omitted from analyses [18].  
                    • Some patterns are challenging (e.g. emphysema vs honeycombing). |
Figure 1 (anticlockwise from top-left). Computed tomography (CT) image of the lung midzones (a) in a patient with idiopathic pulmonary fibrosis with peripheral reticulation and traction bronchiectasis. A three-dimensionally rendered coronal segmentation and extraction of the lungs, trachea and airways has been performed (b), with the two lungs separated into upper, middle and lower zones. The glyph (c) summarises the classification of lung patterns on each voxel volume unit of the lung. The glyph separates the two lungs with a thick vertical black line. Each segment reflects a lung zone (R=right, L=left, U=upper, M=middle, L=lower). Upper and middle lung zones are demarcated using the carina as a landmark, whilst the middle and lower lung zones are separated using a point midway between the diaphragm and the carina. The concentric dotted lines represent quintiles of lung volume. The various colours on the images correspond to different lung patterns. Lung pattern classification is demonstrated as a
three-dimensionally rendered coronal image (d), with equivalent overlays of lung patterns shown on the original axial CT image (e). Fibrosis is evident on the coloured images as ground glass opacity (yellow) and reticulation (orange) with intervening normal lung parenchyma (green).
Figure 2. Axial CT image (a) in a patient with idiopathic pulmonary fibrosis with pulmonary arteries (red) and pulmonary veins (blue) highlighted in different colours on a three-dimensionally rendered image (b). Quantification of total vessel volume (artery and vein) normalised to total lung volume performed at the Ludwig Boltzman Institute was shown to independently correlate with baseline functional indices including forced vital capacity and the diffusion capacity for carbon monoxide.
Figure 3. Axial CT images of a 71-year-old female diagnosed with idiopathic pulmonary fibrosis by a multi-disciplinary team. Pulmonary function tests demonstrated a percent-predicted forced vital capacity of 39.5% and percent-predicted diffusion capacity of carbon monoxide of 13.1%. The CT images (a) demonstrate peripheral reticulation in the right lung and reticulation and traction bronchiectasis, with associated volume loss in the left lung. The data-driven texture analysis (DTA) fibrosis score was 72.2%, with fibrotic areas highlighted in red (b).