Title

A study design for statistical learning technique to predict radiological progression with an application of idiopathic pulmonary fibrosis using chest CT images

Authors

Kim, Grace Hyun J
Shi, Yu
Yu, Wenxi
et al.

Publication Date

2021-05-01

DOI

10.1016/j.cct.2021.106333

Peer reviewed

HHS Public Access

Author manuscript

Contemp Clin Trials. Author manuscript; available in PMC 2022 May 01.

Published in final edited form as:

Contemp Clin Trials. 2021 May ; 104: 106333. doi:10.1016/j.cct.2021.106333.

A study design for statistical learning technique to predict radiological progression with an application of idiopathic pulmonary fibrosis using chest CT images

Grace Hyun J. Kim[1,2], **Yu Shi**[1,2], **Wenxi Yu**[1,2], **Weng Kee Wong**[1]

[1]Biostatistics, Fielding School of Public Health, University of California, Los Angeles.

[2]Radiological Science, David Geffen School of Medicine, University of California, Los Angeles.

Abstract

Background: Idiopathic pulmonary fibrosis (IPF) is a fatal interstitial lung disease characterized by an unpredictable decline in lung function. Predicting IPF progression from the early changes in lung function tests have known to be a challenge due to acute exacerbation. Although it is unpredictable, the neighboring regions of fibrotic reticulation increase during IPF's progression. With this clinical information, quantitative characteristics of high-resolution computed tomography (HRCT) and a statistical learning paradigm, the aim is to build a model to predict IPF progression.

Design: A paired set of anonymized 193 HRCT images from IPF subjects with 6–12 month intervals were collected retrospectively. The study was conducted in two parts: (1) Part A collects the ground truth in small regions of interest (ROIs) with labels of "expected to progress" or "expected to be stable" at baseline HRCT and develop a statistical learning model to classify voxels in the ROIs. (2) Part B uses the voxel-level classifier from Part A to produce whole-lung level scores of a single-scan total probability's (STP) baseline.

Methods: Using annotated ROIs from 71 subjects' HRCT scans in Part A, we applied Quantum Particle Swarm Optimization–Random Forest (QPSO-RF) to build the classifier. Then, 122 subjects' HRCT scans were used to test the prediction. Using Spearman rank correlations and survival analyses, we ascertained STP associations with 6–12 month changes in quantitative lung fibrosis and forced vital capacity.

Conclusion: This study can serve as a reference for collecting ground truth, and developing statistical learning techniques to predict progression in medical imaging.

gracekim@mednet.ucla.edu.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Keywords

particle swap optimization; quantitative lung fibrosis; machine learning; random forest; medical image

1. Introduction and background

Statistical Learning is an integral component of Artificial Intelligence (AI) in medical imaging [1]. However, broadly implementing AI in medical imaging can result in several issues, such as, lack of reproducibility, generalizability, and computational power [1]–[3]. Thus, it is important to develop algorithms that maintain adequate repeatability and reproducibility by using one or more independent sets of data for clinical validation [4]. Generally, there are three steps involved in developing an algorithm: (1) model development, (2) analytic validation, and (3) clinical validation [2]. The first two steps can be achieved by using an appropriate statistical design and model. The last step of clinical validation involves multiple aspects, including reproducibility, limit of agreement, and minimal clinical difference. The last step requires collecting repeated measurements from independent cohorts [5] and consequently, it usually requires a longer time to complete the evaluation.

Statistical learning algorithms broadly fall into the category of supervised learning and unsupervised learning. The former directly utilizes a ground truth or reference typically provided by a medical expert to develop a model. In contrast, most of the unsupervised learning methods, such as a clustering approach, build a model without using the reference or ground truth. There are pros and cons of both approaches and many statistical methods have been utilized jointly with them in feature selection, segmentation, and classification problems [6]–[11].

Developing effective feature selection methods are increasingly important to identify the right variables for accurate predictions. Popular model-based methods for feature selection are least absolute shrinkage and selection operator (LASSO), and smoothly clipped absolute deviation (SCAD) using a penalized likelihood as a loss function [7], [12], [13]. An increasingly powerful class of optimization tools is nature-inspired metaheuristic algorithms and quantum particle swarm optimization (QPSO) is a popular member of this class. QPSO is inspired by particle movements in the quantum mechanics and is a global optimization algorithm [14] with superior searching capabilities. Its performance has been compared favorably with other evolutionary algorithms, for tackling a wide variety of high-dimensional and complex optimization problems in the real-world [15]–[17]. Two drawbacks of QPSO are its longer running time relative to other evolutionary methods and open source codes for QPSO are not easily available for implementation. In our work, we used QPSO, which is not based on the penalized likelihood function, and hybridized it with statistical learning methods to select features for prediction. This is a common and modern approach in engineering, where we hybridize different types of algorithms for enhanced effectiveness. For example, classification models, a kernel-based support vector machine (SVM) and random forest (RF) for classification and regression tree techniques were applied to tackle a few challenging problems [18]–[20]. The synchronized feature selection methods

and classification models can create a generalizable and robust model utilizing statistical learning.

A statistical learning algorithm is typically conducted in two stages: a training stage and a test stage. The training stage is an exploratory phase and uses statistical learning to constantly improve the quality of the model fit by employing an optimization algorithm to update model parameters iteratively. During the model building process, an n-fold cross-validation procedure is usually used to check the adequacy of the fit and the robustness properties of the training model. Afterwards, an independent test set is used to evaluate the model performance. The subjects in the test set are expected to have similar characteristics of the cohort and not from the training set [1], [4].

The major challenge is to build a model that accurately predicts important outcomes that are not readily available. Our aim is to develop a state-of-the-art algorithm and a model using HRCT baseline scans to predict progression in idiopathic pulmonary fibrosis (IPF) in follow-up measurements.

2. Research design and methods

2.1. Objectives

The purpose of this study is to predict disease progression in the natural follow-up of 6–12 months in subjects with IPF using the baseline HRCT. IPF is a rare and fatal interstitial lung disease (ILD) with a median survival time of 3 to 5 years after diagnosis [21]. Progression in IPF is known to have heterogeneous and unpredictable patterns of progression – stable, slow progression or, rapid progression [21], [22]. Although gender, age, and pulmonary function tests (GAP) and usual interstitial pneumonia on HRCT images are known to be prognostic factors for overall survival, it is difficult to reliably predict disease progression in subjects with IPF even with the GAP index [23]–[25]. For IPF subjects, a clinically acceptable outcome is progression free survival (PFS), which is defined as the duration of time after baseline and prior to progression, where progression is defined by 10% or more decline in lung function as measured by the variable of forced vital capacity (FVC). However, it is well known that the clinical outcome of FVC teds to show substantial intra-subject variability in subjects with IPF [26]. Changes in the percent predicted FVC are not necessarily linear over time.

Our work is partly motivated by a recent study that shows the quantitative changes from HRCT scans, measured by quantitative lung fibrosis, from 6 to 9 months predicts PFS within 2 years [27]. However, this approach requires paired HRCT scans of 6–9 months apart, which may not be commonly available. Our innovative approach is to use baseline HRCT scans to predict IPF progression. Because we are using information at single time point (baseline prognosis), it is helpful to have follow up clinical validation, which are both feasible and more realistic than requiring two or more HRCT scans [4]. The potential impact of our works is that it can be used easier in practice to stratify subjects into a different risk group; stable, moderate, rapid progressing groups and apply drug therapies to each subgroup and estimate the relative therapeutic effect. Results from our classification can also be used for early referrals of patients for lung transplants.

This study describes the development and evaluation of the model using a separate dataset to predict the progression of IPF based on the prognosis of CT images at single time point. Our description follows the guideline from Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) [4].

2.2. Design

The study is to build a supervised statistical learning approach using a collection of retrospective HRCT image data in subjects with IPF who were naïve to the anti-fibrotic treatments for 6–12 months [28], [29]. HRCT image is a three-dimensional matrix of size 512-by-512-by the number of slices, where elements of matrix are called as voxels. Each voxel is of size (< 1.0 mm^3) and has a quantitative gray scaled-intensity value of radiodensity ranging from −1024 HU to 1024 HU. The Hounsfield unit (HU) scale is named after Sir Godfrey Hounsfield who invented CT [30].

Our study for predicting a progression using HRCT imaging has two parts: (A) build a model at a small region of interest (ROI) level by optimizing a collection of voxel classification within the ROI and (B) expand the prediction using the model from Part A to all the voxels in the whole lung and produce a metric of single-scan total probability (STP) in predicting progression in ILD (See our study schema in Figure 1). The outcome of the prediction model is a STP score in a percent scale, ranging from 0 to 100. The model performance of STP classification in Part A was assessed in small ROIs compared with the visual reference truth provided by a thoracic radiologist. In Part B, STP scores for the whole were lung compared with the change in QLF score from CT and the percent predicted FVC from pulmonary function tests.

In Part A of our algorithm, our study design collects data by taking advantages of retrospective data. A pair of HRCT images from baseline and at a follow up visit between 6–12 months were presented to a thoracic radiologist. The expert was instructed to contour a region of interest *in the baseline image* and label them as ‘expected to progress’ or ‘expected not to progress’ (i.e. stable) based on the changes in the follow-up HRCT. After collecting the training set of the reference truth of ROI status of being stable or progression, a statistical learning algorithm was built to classify voxels within each ROI into two types: (1) expected to progress or (2) stable. In particular, every voxel within a ROI is assumed to have same labels, because we instructed a radiologist to contour a homogeneous and representative region as a ROI. Our algorithm then searches for a classifier with a combination of features selection iteratively. Upon convergence, the voxels from the test set (i.e. not part of training set) were used to evaluate the model performance for predicting progression as a part of supervised learning. If more than 50% of voxels within a ROI are classified as progression, the ROI is classified as expected to progress. Similarly, if more than 50% of voxels within a ROI are classified as stable, the ROI is classified as expect to be stable.

In Part B, the algorithm classified each voxel in a 4 voxels-by-4 voxels of grid per slice from the whole lung into two types, i.e. whether they are expected to progress or not. We use the baseline scan as a percentage scale to calculate the predicted progression in the follow up visit and call this metric the STP score. The score was compared with the change in lung

function test, FVC, and changes in radiographic outcomes. Change in the percent predicted FVC, measured by the difference between two visits, is commonly used as a primary endpoint in many clinical trials and 10% changes is used as a threshold of defining progression [28], [29], [31]. A radiological outcome, quantitative lung fibrosis (QLF) score is a extent of fibrotic reticular patterns on HRCT images, which is generated by a statistical learning technique using denoised texture features, SCAD feature selection, and SVM classifier [32], [33]. Changes in QLF scores, measured by the difference between two visits, have been utilized as secondary and exploratory imaging endpoints, and the primary endpoint (NCT01979952) in several NIH and industry sponsored clinical trials [34]–[41]. Figure 1 provides a schema of Part A and B in our study design, which is a type 2a study per TRIPOD criteria, where training and test sets are randomly split into samples for development and validation [4].

Patient Selection

We retrospectively collected anonymized longitudinal HRCT images from 215 IPF subjects in multiple studies and the dates of baseline scans ranged from May 2011 to March 2015, which included research time for utilizing existing data to investigate a new imaging model (NIH-NHLBI R21HL 140465–01). The use of anonymous image data was approved by a local institutional review board. A paired HRCT scans (baseline and 6 months to 1 year follow up) from each subject were required for building a model and evaluating its performance. Most of the collected HRCT scans were for IPF diagnosis and from follow up visits without active treatment between the study intervals. Of the 215 IPF subjects, 22 were excluded because of image quality issues, lack of follow-up visits, or the follow-up visits were before 5 months or after 13 months from the baseline visits. The eligible cohort of 193 IPF subjects had a mean age of 70.0 years (SD ±7.5 years), 73% male/27% female, with the percent predictive forced vital capacity (FVC) of 67.8% (SD ±12.3%). The average time from baseline to follow-up visits was 7.6 months (SD ±1.8 months). The baseline quantitative lung fibrosis (QLF) score is 15.4% (SD ±8.7%).

We divided the total sample randomly into a training and a test set. The sample sizes were 71 and 122, respectively. A direct calculation shows the 71 subjects in the training set can provide approximately 93% of the population with 95% tolerance internal [42] and the 122 subjects in the test set can provide approximately 85% power to detect a normalized hazard ratio (HZ) of 2.0 using our proposed STP cut-off score between two groups of high and low values with a two-sided test and at the 5% significance level [43]. To ensure that the training and test sets have about the same distribution of patients' stratification with stable and progressed ROIs, both training and test sets have about 40% subjects with stable ROIs and 60% subjects with at least one progressed ROI in the follow up HRCT scans.

2.2.1. Study Design for Part A—To predict a progression at follow-up, we design the reading paradigm in the traditional supervised approach using the baseline and 6–9 months follow-up HRCT scans from subjects with IPF. The dominant area of usual interstitial pneumonia (UIP) in IPF is located peripherally in the lower or middle lobes of the right and left lungs. As part of disease progression assessment in patients with ILD on HRCT images, an expert thoracic radiologist provided a reference truth (See Figure 1 for details of the

process). The radiologist contoured the classic homogeneous patterns of ROIs with various sizes and labeled them as ‘expected to progress’ or ‘expected not to progress’ (i.e. stable) at baseline scans. Each ROI contains elemental voxels in the HRCT imaging matrix and all voxels received the same label as its own ROI’s label (i.e. either all expected to be stable or all expected to progress). Typically, the radiologist contoured approximately 5 ROIs in a subject’s HRCT image if representative regions were available.

In the training set, there were 434 annotated ROIs from 71 subjects to build the classifier. Out of 434 ROIs, 193 (44.5%) of the annotated ROIs were labeled as expected to progress and 241 (55.5%) ROIs were labeled as expected not to progress; 149 (34%) ROIs are from the upper lung, 185 (43%) ROIs are from the middle lung, and 100 (23%) ROIs are from the lower lung. There are 423 ROIs (97%) that contain the partial or full peripheral of the lung (within 1cm from the chest wall), which is consistent with the nature of the disease. In the test set, there were 549 annotated ROIs from 122 subjects to evaluate the classifier. Out of 549 ROIs, 208 (37.9%) of the annotated ROIs were labeled as expected to progress and 341 (62.1%) ROIs were labeled as expected not to progress.

2.2.2. Measurements for Part A of the study—The study was conducted at the UCLA Computer Vision and Imaging Biomarker (CVIB) in-house workstation that has been built and upgraded since 1997 [44]. Quantitative imaging analysis system has many models and one of many modules is texture features. The types of texture features are : (1) statistical features, which are summary measurements from histogram (e.g. mean, standard deviation, skewness and kurtosis), (2) co-occurrence texture features, which are measures of contrast or uniformity in neighboring voxels by estimating the probabilities of changes gray-level, and (3) run-length parameters, which are the estimated length of uniform gray-level [45], [46]. These features have been frequently used for classifying the patterns of diffuse lung disease on HRCT images since early 1990s in medical imaging [32], [47]. Recently texture features are called radiomic, a type of the many omic features from radiological images.

Pre-processing Image data

Prior to obtaining quantitative texture features, we mitigated the heterogeneous HRCT imaging quality from different acquisition parameters using an image denoising technique [32]. To this end, we applied the denoised technique by Aujol and Gilles. It uses total variation methods [48–50] and for space consideration, we briefly describe the method for our HRCT application. The original CT image (f) is decomposed into the denoised images (u) and noised images (w), respectively. Noise can be modeled using the dual norm in the Besov space with a nonlinear projection ($\boldsymbol{P}_{\mathrm{BG}}$) where the norm of noise δ is less than the norm of signal. The elements of $\boldsymbol{P}$ are denoted by $\{p_{i,j}\}$. The method ensures that when the algorithm converges, the sum of denoised (u) and noise images (w) is approximately equal to original CT image (f). There are parameters in the algorithm and they include δ, λ, and τ, which are, respectively, for noise, residual, and the step size of the gradient descent (fixed point) algorithm. We refer technical details and further notational definitions to [32] and highlight the key steps in the algorithm as follows:

1. Initialization: Set $u_0 = 0$

1. Iterations: Define $w_{n+1} = \mathbf{P}_{\delta BG}(f - u_n)$, where $\mathbf{P}_{\delta BG}$ is an orthogonal matrix, $f = u + v\, u_{n+1} = f - w_{n+1} - \mathbf{P}_{\lambda B_G}(f - w_{n+1})$, and the gradient descent (fixed point) algorithm runs iteratively using

$$p_{i,j}^{n+1} = \frac{p_{i,j}^{n} + \tau\left(\nabla\left(div(p^n) - f/\lambda\right)\right)_{i,j}}{1 + \tau\left|\nabla(\mathrm{div}(p^n) - f/\lambda)\right)_{i,j}\right|}$$

2. Stopping the test: if $\max(|u_{n+1} - u_n|$ and, $|w_{n+1} - w_n|) \leq \varepsilon$, where ε is a user pre-selected constant.

Using the local adaptively variation of uniform region of air in the background, trachea, and aorta on HRCT image, we set the noise parameter (δ) as 50. The residual parameter (λ) was set to 1, which controls the convergence of the algorithm (See the detail [32]). The parameter τ is set to 0.25 [48–50] and ε is set to 1 for the model convergence. By the standard convex duality theory, the projection of $\mathbf{P}$ is an orthogonal projection matrix onto B_G and it involves in a fixed point method and the numerical divergence operator. Figure 2 shows an example of the original and denoised axial slice and a ROI of normal lung patterns from HRCT before and after the grainy scatter noises were removed by Aujol and Chamboll's image decomposition methods [51].

Extracting Quantitative Texture Features from Image data

We extracted texture features from the original HRCT and denoised HRCT images for comparison using our models. Texture features quantify the levels of grayness and contrast of the images. A reading of −1024 HU appears as black on the image and indicates air in the lung, and a reading of −700HU to −200HU appears as white on the image and indicates fibrotic reticulation. A square of approximately 4×4 (approximately 4mm-by-4mm) grid sampling on the contoured slice was implemented to generate voxel instances within each ROI. A ROI is a collection of voxels where a size of voxel is typically < $1mm^3$. We denote that f_{ij} and u_{ij} are the intensity, radiodensity, of voxel at i^{th} and j^{th} *on image,* respectively. And y_{ij} is the label of voxel at i^{th} and j^{th} *within a ROI, where* y_k *is the label of* k^{th} *ROI by a radiologist.* The notation of slice number is omitted for simplicity. The neighboring of each voxel is used to calculate the 191 texture features, which includes a set of statistical features, run-length parameters, and co-occurrence parameters, see for example, [45], [46] A size of the window for calculating the texture feature was 12×12 voxels of neighboring voxels, which is a typical size in ILD classification. Texture features were computed per voxel within a ROI. A same label of ROI is assigned to the label for every voxel within the ROI. To clarify ideas, we define two useful notations, where each voxel f_{ij} from HRCT image:

$$ROI\ expected\ to\ be\ stable\{y_k = 0\} = \bigcup_{i,j}\left\{label\ of\ f_{ij} = 0 \middle| all\ voxels\ in\ f_{ij} \in ROI\right\}$$

$$ROI\ expected\ to\ progress\{y_k = 1\} = \bigcup_{i,j}\left\{label\ of\ f_{ij} = 1 \middle| all\ voxels\ in\ f_{ij} \in ROI\right\}$$

Similarly, each voxel u_{ij} was from a denoised HRCT image:

$$ROI\ expected\ to\ be\ stable\{y_k = 0\} = \bigcup_{i,j} \{label\ of\ u_{ij} = 0 \mid all\ voxels\ in\ u_{ij} \in ROI\}$$

$$ROI\ expected\ to\ progress\{y_k = 1\} = \cup_{i,j} \{label\ of\ u_{ij} = 1 \mid all\ voxels\ in\ u_{ij} \in ROI\}.$$

All the labeled voxels within the ROI is equal to the labels provided by the expert radiologist. If a label of ROI *is 'expected to be stable'* ($y_k = 0$ *for any* k^{th} *ROI*), *all* the labels of voxels within the ROI *are 'expected to be stable' (labels of* f_{ij} *or* $u_{ij} = y_{ij} = 0$ for all i and j in the ROI*).* Similarly, if a label of ROI *is 'expected to progress* ($y_k = 1$ *for any* k^{th} *ROI*) *, all* the labels of voxels within the ROI *to be 'expected to progress' (labels of* f_{ij} *or* $u_{ij} = y_{ij} = 1$ for all i and j in the ROI).

2.2.3. Model Building of Part A—A usual approach in statistical learning is to build an initial classifier model and then use it to build the next model with better classification ability. These iterative steps, along with powerful optimization algorithms, can lead to an effective and robust model for accurate classification. Our models used the traditional ROIs and the labels from the baseline HRCT information to predict the ROIs at follow-up scans by classifying them as progression ($y_i = 1$) or stable ($y_i = 0$) in the follow-up visit. Each voxel within a ROI was classified as expect to progress or to be stable. The classification of the ROI is then determined by the majority votes in the voxels' classification. For example if there are more than 50% voxels in the ROI are expected to progress, then the ROI is classified with the same label as expected to progress (>50%).

Quantum PSO - Random Forest (QPSO-RF) is an integrated algorithm that selects the best HRCT texture features to predict imaging patterns optimally. Previously, we used Quantum PSO to select variables/features from the original HRCT images, and along with RF, built a classification model with satisfactory performance. We plan to compare two QPSO-RF classifiers using texture features from the original HRCT with the denoised HRCT images. The QPSO algorithm selects a candidate feature subsets and to train RF classifiers iteratively until it optimizes the objective function. QPSO then iteratively searches for a better feature subset to train the RF classifier to attain a better objective function value than the one found from the previously selected feature subset by QPSO (see the details of the comparisons with other classifiers [52]). Throughout, we adjust for the imbalanced rates of a classifier (e.g. the rate of progression vs. no-progression) by using a synthetic minority over-sampling technique (SMOTE). Specifically, we used QPSO as the optimizer to search the feature subsets and build the RF from the selected subsets; the built RF produces the evaluation metrics of sensitivity and specificity. The QPSO-RF searches for the feature space iteratively and returns the global best solution at the last iteration as the best feature subset that gives the best classification performance. The figure below illustrates our process and the objective for optimization:

$$\text{maxmize}_{F\theta}\{\min(sensitiviyt, sepcificity)\},$$

where

$$sensitivity = \frac{\sum_{ij}^{N} y_{ij} * \hat{p}_{ij}}{\sum_{i}^{N} I(y_{ij} = 1)},$$

$$specificity = \frac{\sum_{ij}^{N} (1 - y_{ij}) * (1 - \hat{p}_{i}j)}{\sum_{ij}^{N} I(y_{ij} = 1)},$$

and $$accuracy = \frac{\sum \hat{p}_{ij} I(y_{ij} = 1) + (1 - \hat{p}_{ij}) I(y_{ij} = 0)}{N}.$$

Here, we use QPSO-RF to optimize the objective function by first searching all over possible subsets of texture features for the minimum value of either sensitivity (true positive rate) or specificity (true negative rate). Let y_{ij} be a binary variable that takes the value of 1 ($y_{ij} = 1$) if the voxel of i^{th} and j^{th} location shows ground truth of progression and a value 0 if the voxel is stable ($y_{ij} = 0$); let N be the total number of voxels and let $\hat{p}_{ij}$ be the binary probability of progression in the next visit at location of i^{th} and j^{th} voxel based on the QPSO-RF model. We ran QPSO-RP at a voxel-level using five-fold cross-validation procedure, 4 folds for building a model and 1 fold for validation.

The sensitivity and specificity at ROI level was determined by majority voting from the classification labels on the voxel in the ROI. Letting I be the indicator function, we have

$$Esimate\ of\ the\ k^{th}\ ROI\ being\ stable(y_k = 0) = \hat{p}_k = 0 = 1 - I(\frac{\sum_{i,j} I(\hat{p}_{kij} = 0)}{n_k} > 0.5)$$

$$Esimate\ of\ the\ k^{th}\ ROI\ being\ progressed(y_k = 1) = \hat{p}_k = 1 = I(\frac{\sum_{i,j} I(\hat{p}_{kij} = 1)}{n_k} > 0.5),$$

Here, n_k is the total number of voxels in the kth ROI and $\hat{p}_{kij}$ is equal to 0 if the voxel f_{ij} in k^{th} ROI is expected to be stable by our classification model. Similarly, $\hat{p}_{kij}$ is equal to 1 if the voxel f_{ij} in k^{th} ROI is expected to be progressed by our classification model. The estimated label for the k^{th} ROI of $\hat{p}_k$ is 1 (i.e. expected to be progressed) if the 50% or more voxels are classified as being progressed. Similarly, the estimated label for the k^{th} ROI of $\hat{p}_k$ is 0 (i.e. expected to be stable) if the 50% or more voxels are classified as being stable. Here we recall that the diagnostic measure of the sensitivity, specificity, and accuracy of k^{th} ROI follows:

$$sensitivity = \frac{\sum_k^{N1} y_k * \hat{p}_k}{\sum_k^{N1} I(y_k = 1)}, \quad speciicity = \frac{\sum_k^{N2} (1 - y_k) * (1 - \hat{p}_k)}{\sum_k^{N2} I(y_k = 1)},$$

and $$accuracy = \frac{\sum \hat{p}_k I(y_k = 1) + (1 - \hat{p}_k) I(y_k = 0)}{N1 + N2}.$$

Here, N_1 is the total number of ROIs, expected to progress and N_2 is the total number of ROIs expected to be stable, and $\hat{p}_k$ is the binary outcome of k^{th} ROI from the majority voting of voxel-wise classification of QPSO-RF model.

Two models of QPSO-RF statistical learning were developed using two types of texture features from the original and denoised HRCT images. The texture features were derived at voxel-level and the models were optimized at ROI-level. The model performance was compared using metrics, such as, sensitivity, specificity and accuracy at ROI-level to ensure consistency with the unit of visual labels of our reference truth.

2.2.4. Study Design and Measurements for Part B—We next apply the training model built at voxel-level to the whole lung. The classifier model from Part A is designed as a function to be easily integrated into a quantitative imaging analysis system, which averts importing high dimensional CT images (Fig 1B). In the test set, there were 122 independent subjects' HRCT images to test the generated baseline metric at the whole lung level.

To expand and apply the prediction model from Part A to the whole lung level, we first had to determine the boundary of lung segmentation. We used our in-house developed software for semi-automated lung segmentation [44]. The semi-automated tool serves two purposes: (1) to review the results of the segmentation and edit the edge of the lung, when the automated lung segmentation failed to include the parenchymal area of the lung; and (2) to approve the lung segmentation by an experienced radiologist. The algorithm is programmed to classify the voxels within the segmented parenchymal regions, thus the accuracy can be dependent on the quality of the lung segmentation.

After implementing the voxel-level classifier into quantitative imaging analysis system, we estimate the prediction probability using a single-scan total probability (STP) for the whole lung, where we recall STP was created to predict the disease progression in 6 to 9 month of follow up using a baseline HRCT.

There are 5 steps for obtaining a STP metric in quantitative imaging analysis after lung segmentation: (1) denoise HRCT image; (2) conduct grid sampling within a whole lung; (3) calculate texture features selected by QPSO; (4) run random forest classifier, which was built on the QPSO selected features, to obtain classification results for each sampled voxel; (5) record the number of progression voxels predicted by the classifier and total number of voxels, and calculate the STP metric by dividing the former with the latter (See Part B of Figure 1 for graphical explanation). These integrated STP codes with the quantitative

imaging analyses preclude the unnecessary imaging transfers for Part B. The metric of STP is in a percent scale and derived only from baseline imaging information:

$$\text{STP} = \frac{Number\ of\ voxels\ that\ classified\ as\ expected\ to\ be\ progressed\ in\ whole\ lung}{Number\ of\ total\ voxels\ in\ whole\ lung} \times 100.$$

2.2.5. Statistical Analysis for Part B—The STP metric was compared with two outcomes of functional and radiological changes, which are the percent predicted FVC and QLF scores, respectively. We used the STP score driven from a single baseline scan to compare the changes in clinical outcomes: (a) computing the association in changes by Spearman rank test (rho) with the first available follow-ups, and (b) comparing the predictability of high and low STP groups in progression-free-survival in the functional and radiological changes. Complete-case analysis was used. No imputations were used in the missing data.

Continuous scale of scores of STP and changes of QLF and changes in the percent predicted FVC in the first follow-up will be used for testing the association. STP at baseline ranges from 0% to 100%, where higher scores are indicative of high probability of being progressed in the next 6–9 months of visits. The QLF scores represent the extent of the fibrotic reticular patterns on a HRCT scan, which range from 0 to 100% [54]. Higher quantitative scores are indicative of more severe disease. Increasing QLF score indicates the worsening of fibrotic reticulation over time. The percent predicted FVC also ranges 0% to approximately 120%, where 100% predicted FVC indicates the individual lung function is close to normal population adjusted by race, ethnicity, sex, and height [55]. Decreasing in the percent predicted FVC indicates the worsening lung function.

We defined an increase of 4% in QLF between baseline and the follow-up scans as radiographic progression in IPF [56], [57]. Based the previous studies, the minimal clinically important differences was determined to be between 2–4% in the most severe lobe, where the most severe lobe was defined as the one with the largest QLF score at baseline [57] and subjects with >4% changes in QLF score had associated with clinical progression after 6–8 months [58]. In functional changes, we used the clinical definition of PFS, which is the reduction in the percent predicted FVC of 10% [28], [29]. Cox proportional hazard regressions and ad-hoc log-rank analyses were performed to compare two groups high and low baseline STP scores in predicting progression where the outcomes of PFS were defined by the change in QLF and in the percent predicted FVC. We divided two groups based on the approximate of mean: >40% and ≤ 40% for high and low baseline STP scores, respectively.

3. Results

We now present results from the three steps in our algorithm in the following order: (1) model development, (2) analytic validation, and (3) clinical validation.

3.1 Result of Model Development

The five-fold cross-validation method yields a parsimonious model with 23 features from the original HRCT images, whereas our algorithm yields a set of 18 texture features from the denoised HRCT images using QPSO-RF. The latter model is favorable due to the less number of important features, which indicate a more parsimonious model that saves computation time and prevents a model from overfitting [59].

3.2 Analytic Evaluation and Validation from Part A

Table 1 shows the sensitivity, specificity, and accuracy of the QPSO-RF statistical learning approach using two types of texture features from the original and denoised HRCT images from Part A. The performance of the model ranges approximately 70% in the training set and 65%−68% in the test set in terms of sensitivity, specificity, and accuracy using the texture features from the original HRCT images. Moreover, the performance of the model from the denoised texture features reached approximately 60–73% in the training set and 70% in the test set in terms of sensitivity, specificity, and accuracy. Model performance in accuracy was numerically higher as 70% when using the denoised texture features, compared to 67% for the original texture features.

We produced a set of texture features from the denoised HRCT images, which had superior robust results in the test set. At the same time, the model performed reasonably well in ROIs matching with visual reference in predicting progression in the next HRCT scans given that the prediction using only the baseline characteristics only is a challenging problem, it remains approximately 30% (i.e. 100–70% accuracy) of non-deterministic factors

3.3 Clinical Evaluation from Part B

In Part B of the study, we integrated the selected denoised features from Part A of the study and the classification algorithm of QPSO-RF that was deployed for prediction in the whole lung level to obtain a STP score. Table 2 provides the baseline characteristics of the 122 IPF subjects in the test set, by their follow-up progression status, namely, whether or not subjects had QLF scores increased by more than 4%. The non-progression group had a higher percentage of female, were slightly older, had lower QLF and higher percentage predicted FVC at baseline.

Figures 3 and 4 contain four subfigures and represent the examples of the expected to be progressed and stable cases, respectively. The figures display the representative axial HRCT images at baseline and the corresponding the dichotomized results and probability of predictive progression, as well as the follow-up axial HRCT images.

Figure 3 is a representative sample of IPF subject with a STP of 49.1%. This subject had an increased QLF of 12.6% from baseline to 17.3% at 7-month follow-up, which was more than 4% QLF change in the follow-up scan, and had experienced a more than 10% drop (from 69% to 55%) in FVC percent predicted value. Figure 3 shows that the majority of voxels were classified as expected to progress. Overall STP consists of the voxels that classified as expected to progress in both dichotomized and continuous probabilities (Fig 3.b and Fig 3.c), which can be confirmed in baseline and 7 month HRCT scans (Fig 3.a, and Fig

3.d). This model is able to pick up many voxels expected to progress in their exact locations at the follow-up HRCT.

Figure 4 is a representative stable case of IPF subject with a STP of 14.9%. The subject had the same QLF score of 4% from baseline and 12-month follow-up, which indicates progression. Figure 4 shows that the majority of voxels were classified as expected to be stable and the overall STP consists of the signals with being stable (Fig 4.b and Fig 4.c), which can be confirmed in baseline and 12 month HRCT scans (Fig 4.a, and Fig 4.d). The prediction model is able to pick up many stable voxels in their exact locations at the follow-up HRCT.

3.4 STP Associations with QLF and FVC

We used the STP score from a single baseline scan to compare the changes in clinical outcomes: (a) association correlation coefficient (rho) in changes, and (b) testing predictability of progression-free-survival (PFS). Disease progression are defined by (i) the lung functional change in the percent predicted FVC of 10% or more reduction; this is a common criteria for evaluating a therapeutic effect on the lung function [28], [29], and (ii) an increase in radiological outcome QLF of 4% or more [58]. In our study, STP had a weak trend with the changes in QLF at 6–9 month follow-ups (rho ρ=0.1295; p=0.155) with no significant association with the percent predicted FVC scores (rho ρ=−0.0115; p=0.90). However, a moderate association was found between changes in QLF and the percent predicted FVC (rho ρ=−0.25; p=0.0074).

In PFS analyses, high STP score at baseline scans had poor prognosis with QLF changes, and PFS defined by FVC changes did not reach statistical significance. In a univariate analysis, higher STP is associated with higher risk of progression in a univariate analysis, with a normalized hazard ratio of 1.45 (p = 0.027). In a multivariate analysis after adjusting for subjects' age and gender, the normalized hazard ratio is 1.53 (p = 0.041). No statistically significant trend was found in PFS using FVC percent predicted outcome; the normalized hazard ratio is 0.88, p=0.49 for univariate analysis, and the normalized hazard ratio is 0.92 p=0.70 for multivariate analysis adjusting for gender and age. Table 3 summarizes the results from Cox regression using the two types of outcomes with the baseline STP as a covariate.

The mean (± SE) follow up time was 7.6 (± 0.2) months for HRCT and 8.0 (± 0.4) months for pulmonary function test. Figure 5 shows that the STP was well correlated with QLF based progression with the median of 6–7 month follow-ups. Mean of STP at baseline was high in the group with progression compared with group without progression, as is shown in Table 3. Subjects with 40 or higher in STP had earlier PFS than those with STP score below than 40 (log-rank test, p=0.0196). However, STP was not significantly related to levels in functional progression-free survival using FVC, even after examining the longitudinal observations beyond 1 year follow-ups (log-rank test, p=0.65). Of note that 75% of population for those who had >40% STP (~=33 subjects) were missing after 12 months. No significant result in STP was found in PFS using FVC as an outcome, after we right censored the data at 400 days, which is about the last days of QLF follow-ups (log-rank test, p=0.50).

4. Discussion

Building a robust model to produce results for clinical validation requires careful consideration of the processes that include the targeted population for a training data set [60], [61]. Overall, our proposed STP score is significantly associated with the changes in QLF scores, where both scores are derived from HRCT images. However, the STP score was not associated with the expected changes in the percent predicted FVC. There are several possible reasons for this observation. One of the main reason for the disassociation is that the data set was collected from subjects with paired HRCT scans to build a classifier model. The inclusion criteria in this retrospective study were: (a) subjects who have not yet undergone a lung transplant (b) available HRCT within 6–12 months (c) available percent predicted FVC measurements who met conditions (a) and (b). Thus, the sequence of inclusion criteria may have resulted in more missing FVC during data collection than the other scenario of FVC data first collected and then checked for available HRCT data. Furthermore, there were many incidences where the duration of percent predicted FVC in the follow up visit were less than a year, which is deemed unreliable in the natural follow up. The main reason for a missing FVC is that the subject could no longer be in the study; for subjects who had more than 40% STP, it is likely that he or she had moved on to an anti-fibrotic treatment or opted into a lung transplant program.

To build a robust model with good classification results, it is also important to clearly identify the target population [60]. First, the characteristics of the independent test set have to be similar to the training set. For example, both the prevalence of population and inclusion criteria for model building should be similar. Second, the generalizability of a model is an important attribute of a successful model; to achieve greater generalizability, it is crucial that we understand the sources of measurement variations in data collection and able to resolve or control the sources of variation. Beyond the statistical techniques of feature selection and classification for developing a model, it is important to consider the data collection method. In particular, careful attention must be given to the imaging data collection, because the quality of the lung segmentation can affect accuracy in the classification model. For example, the choice of the segmentation model and setting a lung boundary for the domains of data elements can have an impact on the next steps of analytic validation in calculating texture features and clinical validation. The data collection method should also be developed specific to the intended population of the training set. This is critical to maintain the quality and characteristics of the training data set and mitigate factors that require normalization of the heterogeneous data for generalizability in a test set.

The significance of this study can be summarized as follows. Our integrated procedure uses statistical learning to collect data and development for predicting disease progression along with an analytic evaluation methodology, and a practical model for clinical evaluation. The proposed methods requires multidisciplinary approaches, such as statistical learning and analytic tools from the mathematical and engineering disciplines to solve some of the problems that typically arise in imaging studies. The problems include situations when there are (a) unbalanced rates of classifier (e.g. different prevalence rates), (b) different sources of measurement variations from multicenter studies, and (c) requirements to simultaneous to identify and process the important feature selections. In our work, we solve (a) by using a

synthetic minority over-sampling technique, (b) by using features from denoised images, and (c) by using particle swamp optimization with random forest classifier to overcome the inter-dependency between feature selection and the classification model. Because the characteristics of the training data set can influence the accuracy of prediction in a test set for a prediction model, , we emphasize the source of the training and test data sets have to be the same as much as possible to facilitate algorithm or AI implementation.

The performance of our model in the analytic evaluation of ROIs ranged between 65% and 70% in sensitivity, specificity, and accuracy (Table 1). This indicates that prediction is a difficult problem and the rest of the 30% to 35% remains unpredictable. There may be due to genetic, environmental, and clinical factors. Prediction using medical imaging can play a role in precision medicine as additional laboratory results [28], [29] . The clinical evaluation of STP model performs well in accordance with radiological outcomes. However, the model performance was dragged down by the pulmonary function test, FVC. Considering that the early changes of QLF predict the FVC changes and the duration of FVC follow-up was less than a year, we expect to have higher concordance of STP with the lung function for longer follow-ups [25]. Normally the duration of 1 year changes in FVC has been commonly used in clinical trials; a longer follow-up can be explored in the next planning and data collection [28], [29].

Overall, our approach seems promising but note that there are practical limitations. First, this study is a retrospective study. Evaluating whether the proposed algorithm is correctly predicting disease progression can be limited due to missing data who are no longer in the natural follow up without treatment, where active anti-fibrotic treatments are available since year 2014 [28], [29]. There could be multiple reasons for missing data in FVC measurement. Clinical measurements of pulmonary function tests may not be routinely performed and electrically stored. Secondly this supervised statistical learning approach may require a long time in data collection for gathering the reference truth from an expert and for examining and deriving a set of standardized features from approved lung segmentation. In contrast, deep learning or unsupervised learning may requires short time in a model development. Another limitation of our approach is that the current training model did not incorporate directly clinical information of FVC decline. We used only radiographic worsening as a reference truth in the training model.

After STP score is validated through several clinical studies, another potential application is to provide a counterfactual scenario to estimate the probability of progression in ILD if a subject decides not to take or continue with an effective treatment. We note that a subject who experienced disease progression may not undergo the HRCT scanning in a clinic. In this scenario, prediction of disease progression using STP can be derived from baseline or the prior scans. Our work in STP is initial work in predicting the radiological progression or worsening in a short term of 6–12 months with only a single HRCT scan. This is the first study to use the baseline HRCT scan to predict the progression in whole lung. A new study is currently underway for evaluation and validation.

5. Conclusions

Recent medical research is completed to analyze and make an inference because of increasingly huge data set. Standard statistical methods may no longer be adequate and modern analytic tools can be used. Increasingly, this involves statistical learning techniques and nature-inspired metaheuristic algorithms, such as quantum particle swam optimization, or some hybridization thereof. Our initial work of a statistical learning paradigm is an integrate approach, coupled with a hybrid of quantum particle swamp optimization and a random forest algorithm built upon a reference truth of visual assessment. The classifier model is designed to easily integrate into quantitative imaging analysis system which averts importing high dimensional CT images. The results of statistical learning model is to quantify a score of progression in IPF patients using a single HRCT scan. To attain clinical utility, clinical evaluation is further required.

Reference

[1]. James G, Witten D, Hastie T, and Tibshirani R, An introduction to statistical learning, vol. 112. Springer, 2013.

[2]. Wernick MN, Yang Y, Brankov JG, Yourganov G, and Strother SC, “Machine learning in medical imaging,” IEEE Signal Process. Mag, vol. 27, no. 4, Art. no. 4, 2010.

[3]. Menzies T, “Guest Editor? s Introduction: 21st Century AI–Proud, Not Smug,” IEEE Intell. Syst, no. 3, Art. no. 3, 2003.

[4]. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation vol. 13.

[5]. Obuchowski NA et al., “Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons,” Stat. Methods Med. Res, vol. 24, no. 1, pp. 68–106, 2015. [PubMed: 24919829]

[6]. Donoho DL and Johnstone IM, “Adapting to unknown smoothness via wavelet shrinkage,” J. Am. Stat. Assoc, vol. 90, no. 432, pp. 1200–1224, 1995.

[7]. Fan J and Li R, “Variable selection via nonconcave penalized likelihood and its oracle properties,” J. Am. Stat. Assoc, vol. 96, no. 456, pp. 1348–1360, 2001.

[8]. Heinze G, Wallisch C, and Dunkler D, “Variable selection–a review and recommendations for the practicing statistician,” Biom. J, vol. 60, no. 3, pp. 431–449, 2018. [PubMed: 29292533]

[9]. Liu Y, Zhang HH, and Wu Y, “Hard or soft classification? large-margin unified machines,” J. Am. Stat. Assoc, vol. 106, no. 493, pp. 166–177, 2011. [PubMed: 22162896]

[10]. Juan-Albarracín J et al., “Automated glioblastoma segmentation based on a multiparametric structured unsupervised classification,” PLoS One, vol. 10, no. 5, Art. no. 5, 2015.

[11]. Wang P et al., “Machine learning models for diagnosing glaucoma from retinal nerve fiber layer thickness maps,” Ophthalmol. Glaucoma, vol. 2, no. 6, Art. no. 6, 2019.

[12]. Tibshirani R, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. Ser. B Methodol, vol. 58, no. 1, Art. no. 1, 1996.

[13]. Tibshirani R, Saunders M, Rosset S, Zhu J, and Knight K, “Sparsity and smoothness via the fused lasso,” J. R. Stat. Soc. Ser. B Stat. Methodol, vol. 67, no. 1, pp. 91–108, 2005.

[14]. Sun J, Feng B, and Xu W, “Particle swarm optimization with particles having quantum behavior,” in Proceedings of the 2004 congress on evolutionary computation (IEEE Cat. No. 04TH8753), 2004, vol. 1, pp. 325–331.

[15]. Sun J, Lai C-H, and Wu X-J, Particle swarm optimisation: classical and quantum perspectives Crc Press, 2016.

[16]. Jin C and Jin S-W, "Prediction approach of software fault-proneness based on hybrid artificial neural network and quantum particle swarm optimization," Appl. Soft Comput, vol. 35, pp. 717–725, 2015.

[17]. Li Y, Jiao L, Shang R, and Stolkin R, "Dynamic-context cooperative quantum-behaved particle swarm optimization based on multilevel thresholding applied to medical image segmentation," Inf. Sci, vol. 294, pp. 408–422, 2015.

[18]. Vapnik V, The nature of statistical learning theory Springer science & business media, 2013.

[19]. Breiman L, "Manual on setting up, using, and understanding random forests v3. 1," Stat. Dep. Univ. Calif. Berkeley CA USA, vol. 1, p. 58, 2002.

[20]. Strobl C, Boulesteix A-L, Zeileis A, and Hothorn T, "Bias in random forest variable importance measures: Illustrations, sources and a solution," BMC Bioinformatics, vol. 8, no. 1, p. 25, 2007. [PubMed: 17254353]

[21]. Raghu G 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, vol. 183, no. 6, Art. no. 6, 2011.

[22]. Raghu G et al., "Diagnosis of idiopathic pulmonary fibrosis. An official ATS/ERS/JRS/ALAT clinical practice guideline," Am. J. Respir. Crit. Care Med, vol. 198, no. 5, Art. no. 5, 2018.

[23]. Ley B et al., "A multidimensional index and staging system for idiopathic pulmonary fibrosis," Ann. Intern. Med, vol. 156, no. 10, pp. 684–691, 2012. [PubMed: 22586007]

[24]. Flaherty KR et al., "Radiological versus histological diagnosis in UIP and NSIP: survival implications," Thorax, vol. 58, no. 2, pp. 143–148, 2003. [PubMed: 12554898]

[25]. Salisbury ML et al., "Idiopathic pulmonary fibrosis: gender-age-physiology index stage for predicting future lung function decline," Chest, vol. 149, no. 2, pp. 491–498, 2016. [PubMed: 26425858]

[26]. Nathan SD 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, vol. 71, no. 5, pp. 429–435, 2016. [PubMed: 26968970]

[27]. Kim GHJ et al., "Prediction of idiopathic pulmonary fibrosis progression using early quantitative changes on CT imaging for a short term of clinical 18–24-month follow-ups," Eur. Radiol, pp. 1–9, 2019.

[28]. King TE Jr et al., "A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis," N. Engl. J. Med, vol. 370, no. 22, pp. 2083–2092, 2014. [PubMed: 24836312]

[29]. Richeldi L et al., "Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis," N. Engl. J. Med, vol. 370, no. 22, pp. 2071–2082, 2014. [PubMed: 24836310]

[30]. Webb WR, Brant WE, and Major NM, Fundamentals of Body CT E-Book Elsevier Health Sciences, 2019.

[31]. Raghu G 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, vol. 183, no. 6, Art. no. 6, 2011.

[32]. Kim HJ et al., "Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study," Acad. Radiol, vol. 15, no. 8, pp. 1004–1016, 2008. [PubMed: 18620121]

[33]. Kim HJ 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, vol. 22, no. 1, pp. 70–80, 2015. [PubMed: 25262954]

[34]. Kim HJ et al., "Quantitative texture-based assessment of one-year changes in fibrotic reticular patterns on HRCT in scleroderma lung disease treated with oral cyclophosphamide," Eur. Radiol, vol. 21, no. 12, pp. 2455–2465, 2011. [PubMed: 21927793]

[35]. Tashkin DP et al., "Mycophenolate mofetil versus oral cyclophosphamide in scleroderma-related interstitial lung disease (SLS II): a randomised controlled, double-blind, parallel group trial," Lancet Respir. Med, vol. 4, no. 9, Art. no. 9, 2016.

[36]. Raghu G et al., "FG-3019 anti-connective tissue growth factor monoclonal antibody: results of an open-label clinical trial in idiopathic pulmonary fibrosis," Eur. Respir. J, vol. 47, no. 5, Art. no. 5, 2016.

[37]. Richeldi L et al., “Pamrevlumab, an anti-connective tissue growth factor therapy, for idiopathic pulmonary fibrosis (PRAISE): a phase 2, randomised, double-blind, placebo-controlled trial,” Lancet Respir. Med, vol. 8, no. 1, Art. no. 1, 2020.

[38]. Fishman J, Kim G, Kyeong N, Goldin J, and Glassberg M, “Intravenous stem cell dose and changes in quantitative lung fibrosis and DLCO in the AETHER trial: A pilot study,” Eur Rev Med Pharmacol Sci, vol. 23, pp. 7568–7572, 2019. [PubMed: 31539148]

[39]. Palmer SM et al., “Randomized, double-blind, placebo-controlled, phase 2 trial of BMS-986020, a lysophosphatidic acid receptor antagonist for the treatment of idiopathic pulmonary fibrosis,” Chest, vol. 154, no. 5, Art. no. 5, 2018.

[40]. Denton CP et al., Lung function preservation in a phase 3 trial of tocilizumab (TCZ) in systemic sclerosis (SSc). Eur Respiratory Soc, 2019.

[41]. Martyanov V et al., “Novel lung imaging biomarkers and skin gene expression subsetting in dasatinib treatment of systemic sclerosis-associated interstitial lung disease,” PloS One, vol. 12, no. 11, Art. no. 11, 2017.

[42]. Hahn GJ and Meeker WQ, Statistical intervals: a guide for practitioners, vol. 92. John Wiley & Sons, 2011.

[43]. Schoenfeld DA, “Sample-size formula for the proportional-hazards regression model,” Biometrics, pp. 499–503, 1983. [PubMed: 6354290]

[44]. Brown MS et al., “Automated measurement of single and total lung volume from CT,” J. Comput. Assist. Tomogr, vol. 23, no. 4, pp. 632–640, 1999. [PubMed: 10433299]

[45]. Haralick RM, Shanmugam K, and Dinstein IH, “Textural features for image classification,” IEEE Trans. Syst. Man Cybern, no. 6, pp. 610–621, 1973.

[46]. Sonka M, Hlavac V, and Boyle R, Image processing, analysis, and machine vision Cengage Learning, 2014.

[47]. Chabat F, Yang G-Z, and Hansell DM, “Obstructive lung diseases: texture classification for differentiation at CT,” Radiology, vol. 228, no. 3, pp. 871–877, 2003. [PubMed: 12869685]

[48]. Aujol J-F and Chambolle A, “Dual norms and image decomposition models,” Int. J. Comput. Vis, vol. 63, no. 1, Art. no. 1, 2005.

[49]. Gilles J, “Noisy image decomposition: a new structure, texture and noise model based on local adaptivity,” J. Math. Imaging Vis, vol. 28, no. 3, Art. no. 3, 2007.

[50]. Aujol J-F, Gilboa G, Chan T, and Osher S, “Structure-texture image decomposition—modeling, algorithms, and parameter selection,” Int. J. Comput. Vis, vol. 67, no. 1, Art. no. 1, 2006.

[51]. Aujol J-F and Chambolle A, “Dual norms and image decomposition models,” Int. J. Comput. Vis, vol. 63, no. 1, pp. 85–104, 2005.

[52]. Shi Y, Wong WK, Goldin JG, Brown MS, and Kim GHJ, “Prediction of progression in idiopathic pulmonary fibrosis using CT scans at baseline: A quantum particle swarm optimization-Random forest approach,” Artif. Intell. Med, vol. 100, p. 101709, 2019. [PubMed: 31607341]

[53]. Brown MS et al., “Automated measurement of single and total lung volume from CT,” J. Comput. Assist. Tomogr, vol. 23, no. 4, Art. no. 4, 1999.

[54]. Wu X et al., “Computed tomographic biomarkers in idiopathic pulmonary fibrosis. The future of quantitative analysis,” Am. J. Respir. Crit. Care Med, vol. 199, no. 1, Art. no. 1, 2019.

[55]. Hankinson JL, Odencrantz JR, and Fedan KB, “Spirometric reference values from a sample of the general US population,” Am. J. Respir. Crit. Care Med, vol. 159, no. 1, pp. 179–187, 1999. [PubMed: 9872837]

[56]. Wu X et al., “Computed tomographic biomarkers in idiopathic pulmonary fibrosis. The future of quantitative analysis,” Am. J. Respir. Crit. Care Med, vol. 199, no. 1, pp. 12–21, 2019. [PubMed: 29986154]

[57]. Kafaja S 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, vol. 197, no. 5, pp. 644–652, 2018. [PubMed: 29099620]

[58]. Kim GHJ et al., “Prediction of idiopathic pulmonary fibrosis progression using early quantitative changes on CT imaging for a short term of clinical 18–24-month follow-ups,” Eur. Radiol, pp. 1–9, 2019.

[59]. Ng AY, “On feature selection: learning with exponentially many irreverent features as training examples,” PhD Thesis, Massachusetts Institute of Technology, 1998.

[60]. Campbell MJ and Machin D, Medical statisticsa commonsense approach 1993.

[61]. Graham JW, “Missing data analysis: Making it work in the real world,” Annu. Rev. Psychol, vol. 60, pp. 549–576, 2009. [PubMed: 18652544]

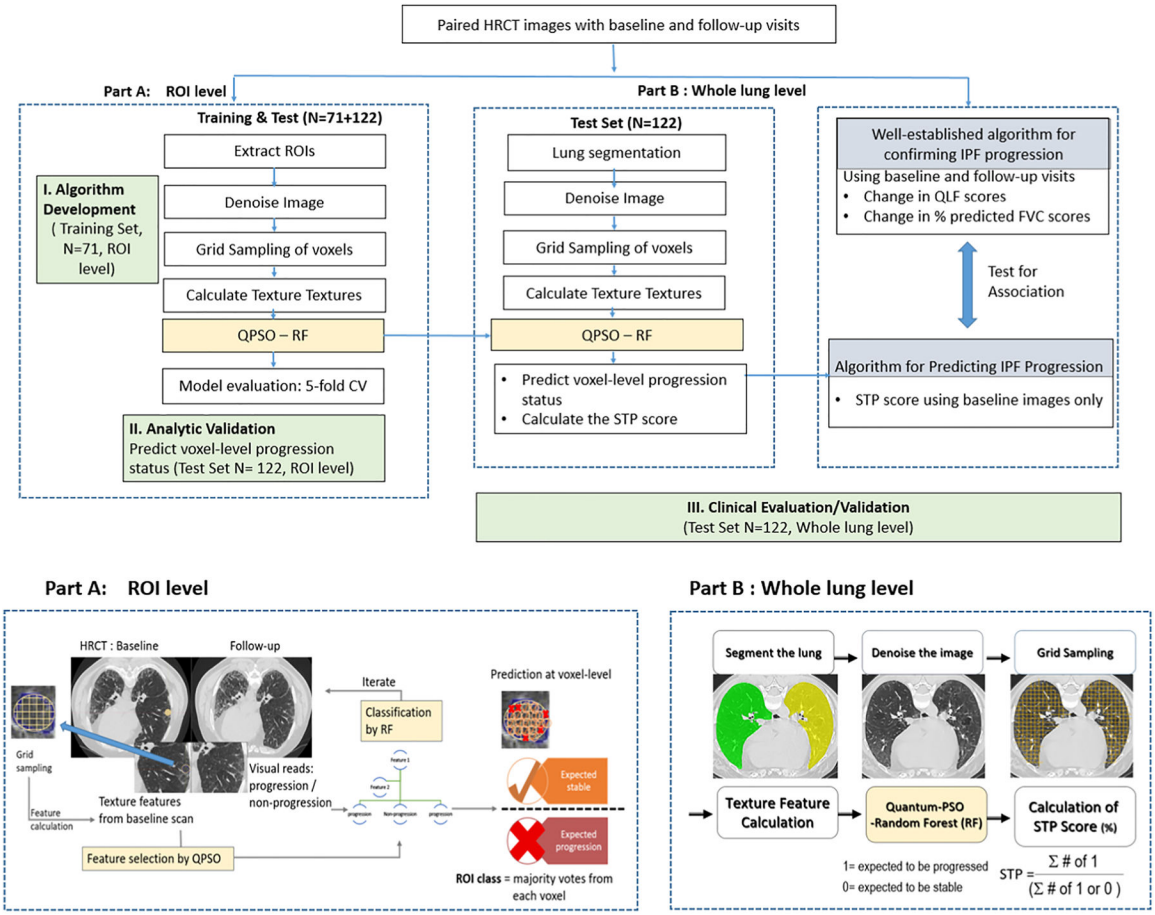

Figure 1:

Study Design of Flow Chart and Graphical Illustration of ROI and Whole Lung Level. HRCT: high-resolution computed tomography, ROI: region of interest, QPSO: quantum particle swarm optimization, RF: random forest, CV: cross validation, STP: single-scan total probability, QLF: quantitative lung fibrosis

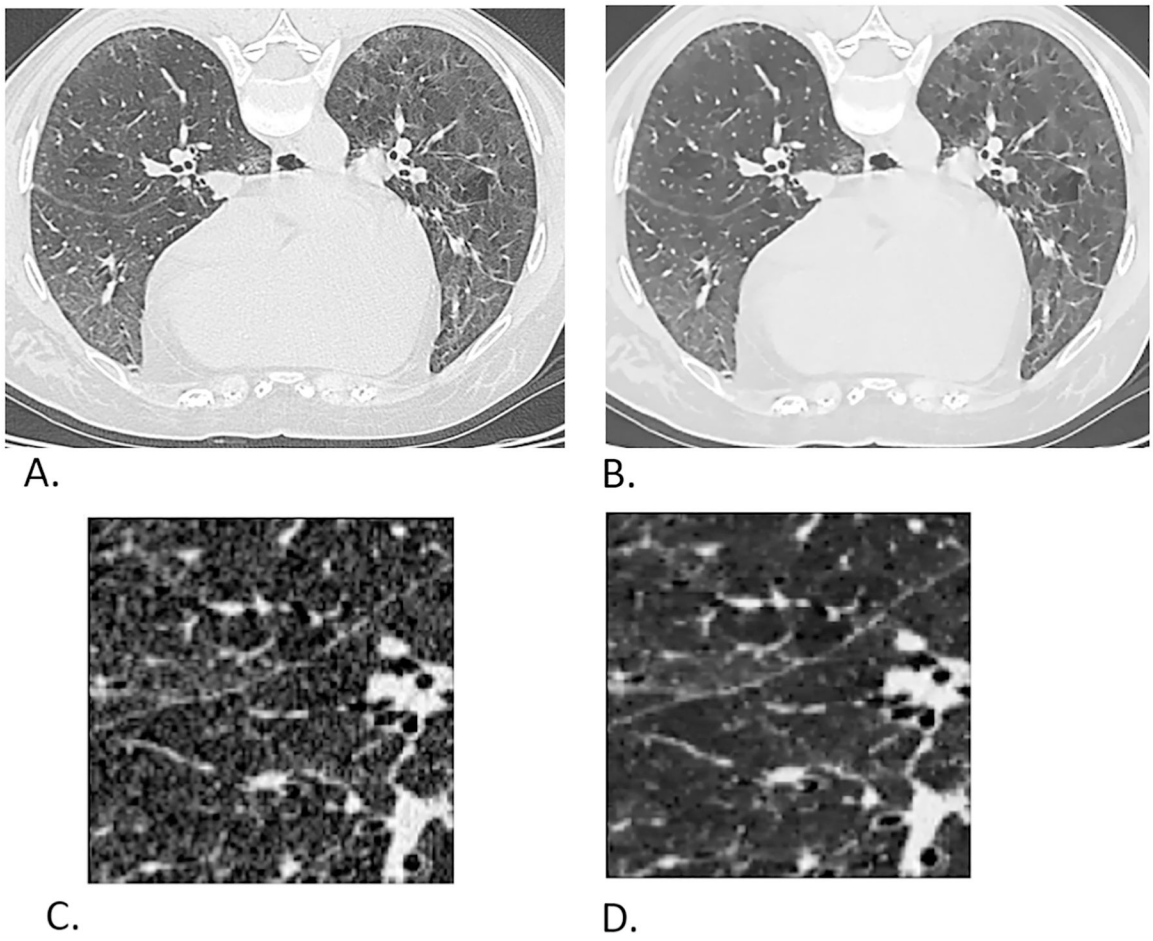

Figure 2.
Example of original axial slice HRCT image and its denoised image with enlarged region of interest (ROI): **(A)** original axial HRCT; **(B)** corresponding denoised image of (A); **(C)** original ROI; **(D)** corresponding denoised image of (C).

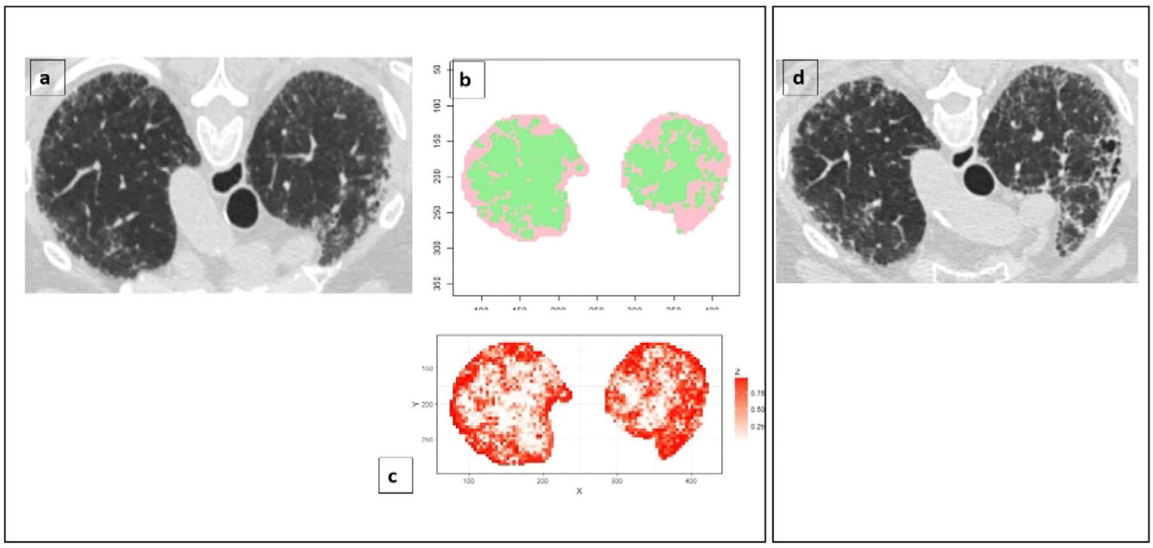

Figure 3.

Whole Lung Predictive Progression from Part B (Progressed case): The subject has a QLF of 12.6% at baseline and 17.3% in a 7-month follow-up and the percent predicted FVC was 69% and 55% at baseline and 12-months, respectively: **(a)** baseline HRCT; **(b)** dichotomized classification results of (a) (green dots = voxels expected not to progress, red dots = voxels expected to progress, with 0.5 probability cutoff), the percentage of predicted progression voxel is 52.4% on this slide; **(c)** classification result of predictive probability of progression of (a); **(d)** 7-month follow-up HRCT.

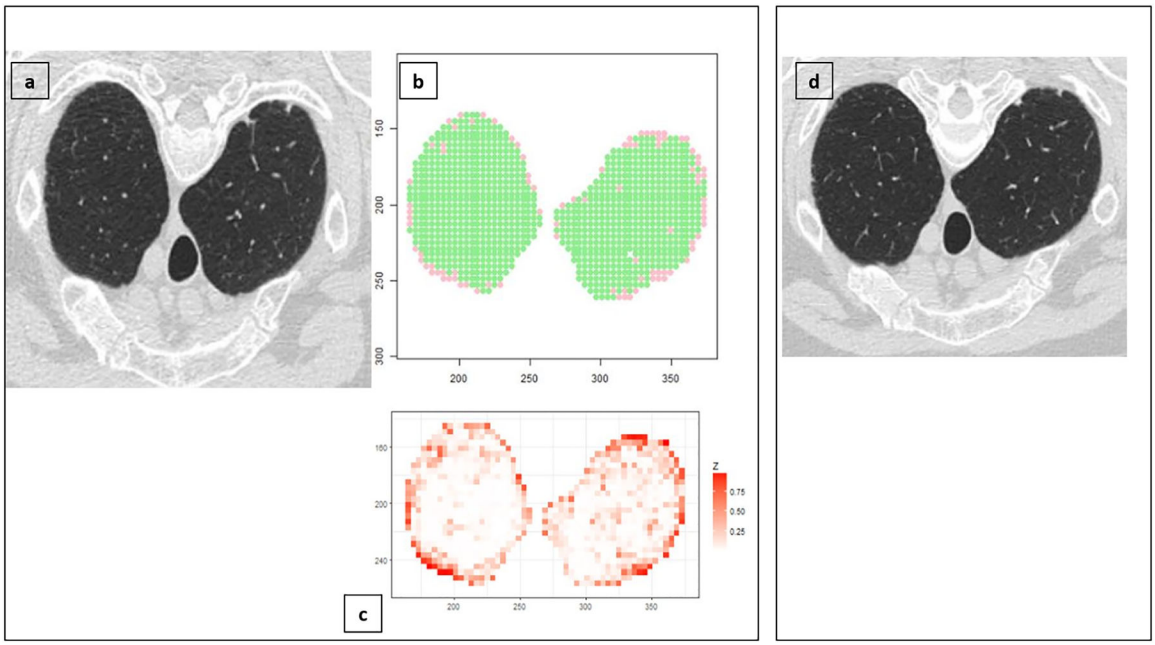

Figure 4.

Whole Lung Predictive Progression from Part B (Stable case): The subject has a QLF of 4% at baseline and 4% in a 12-month follow-up and the percent predicted FVC was 74.6% and 76.0% at baseline and 12-months, respectively: **(a)** baseline HRCT; **(b)** dichotomized classification results (a) (green dots = voxels expected not to progress (stable), red dots = voxels expected to progress, with 0.5 probability cutoff), the STP of predicting progression voxel is 8.7% on this slide; **(c)** classification result of predictive probability of progression of (a); **(d)** 12-month follow-up HRCT.

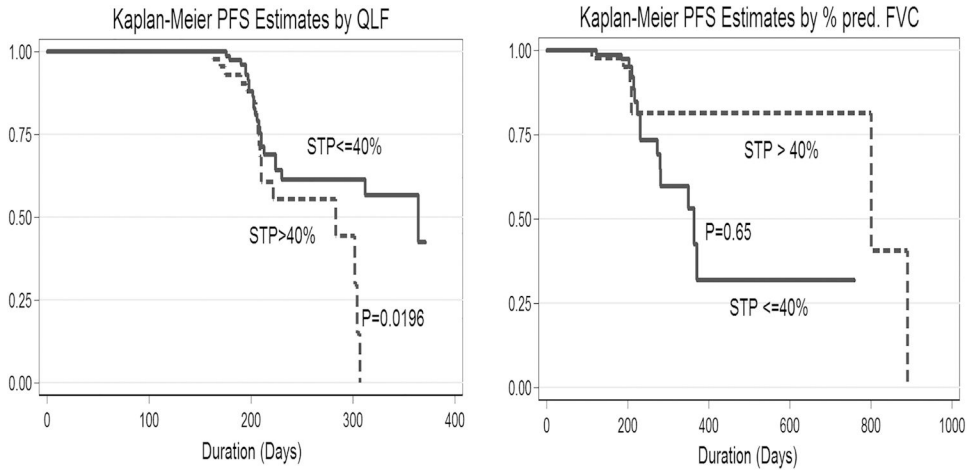

Figure5:

Kaplan-Meier Curve by the Single-scan Total Prediction (STP) in interstitial lung disease: (A) Progression defined by QLF score more than 4% increase; (B) Progression defined by the percent predicted FVC 10% or more reduction

Table 1.

Evaluation of Classification in the Region of Interests (ROIs) from Part A of the study.

QPSO-RF model Machine Learning					
Texture Features from the original images			**Texture Features from denoised images**		
For 5-fold Cross-Validation (71 subjects)			For 5-fold Cross-Validation (71 subjects)		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
0.71	0.70	0.70	0.73	0.60	0.65
For independent test set (122 subjects)			For independent test set (122 subjects)		
Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
0.68	0.65	0.67	0.70	0.70	0.70

Table 2.

Baseline characteristics of the 122 subjects in the test set, by their progression status in the follow-up visits

	ΔQLF > 4% (progression group)	**ΔQLF ≤ 4% (non-progression group)**	**Total**
# Subjects (%)	38 (31%)	84 (69%)	122
% Female	13%	34%	27%
Age, mean in years (SE)	68.1 (± 1.3)	71.0 (± 0.9)	70.0 (± 0.7)
Quantitative Lung Fibrosis, %, mean (SE)	17.6 (± 1.2)	14.5 (± 1.0)	15.4 (± 0.8)
Total Lung Capacity *, L, mean (SE)	3.82 (± 0.113)	3.88 (± 1.31)	3.84 (± 0.874)
FVC predicted percentage,%, mean (SE)	64.2 (± 1.7)	69.4 (± 1.4)	67.8 (± 1.1)

* Total Lung volume from HRCT

Table 3.

Summary results of baseline STP by the progression status in the follow-up visits

			Baseline STP
QLF changes	Baseline mean (±SE)	ΔQLF > 4%	40.0 (± 1.8)
		ΔQLF ≤ 4%	35.4 (± 1.2)
		Total	36.8 (± 1.0)
	Univariate normalized hazard ratio (p-value)		1.45 (p=0.027 *)
	Dichotomized STP # of subjects (%)	> 40%	44 (36%)
		≤ 40%	78 (64%)
		log-rank test	p=0.020 *
	Multivariate normalized hazard ratio (p-value)		1.53 (p=0.041 *)
Changes in % predicted FVC (ppFVC)	Baseline mean (±SE)	ΔppFVC < −10%	35.2 (± 2.3)
		ΔppFVC ≥ −10%	37.3 (± 1.1)
		Total	36.7 (± 1.0)
	Univariate normalized hazard ratio (p-value)		0.88 (p=0.49)
	Dichotomized STP # of subjects (%)	> 40%	44 (36%)
		≤ 40%	78 (64%)
		log-rank test	p=0.974
	Multivariate normalized hazard ratio (p-value)		0.92 (p=0.70)

Asterisk(*) indicates significance at 0.05 alpha level; ΔQLF: changes in QLF score , where a threshold of 10% reduction based on [58]; ΔppFVC: changes in the percent predicted FVC score, where a threshold of 10% reduction based on [28], [29]; Dichotomized Single-scan total probability (STP) at 40%

#: the threshold is an approximate of 36.8% mean and 37.6% median STP at baseline scan.