Background: The interstitial lung disease (ILD) associated with connective tissue diseases including systemic sclerosis (SSc) is heterogeneous disease characterized by reduced survival of approximately 3 years (1). "Radiomics'' is a field of research which describes the in-depth analysis of tissues by computational retrieval of high-dimensional quantitative features from medical images (2). Our previous study suggested capacity of radiomics features to differentiate between "high'' and "low'' risk groups for lung function decline in two independent cohorts (3).

Objectives: (1) To develop robust, machine learning (ML) workflow for "radiomics'' data in SSc-ILD to select optimal methods for prediction. (2) To predict the time to individual lung function decline defined as defined by the time to a relative decline of ≥ 15% in Forced Vital Capacity (FVC)% as previously proposed (3), using workflow.

Methods: We investigated two cohorts of SSc-ILD: 90 patients (76.7% female, median age 61.0 years) from Oslo University Hospital's. Patients were retrospectively selected if (3): a) diagnosed with early/mild SSc according to the Very Early Diagnosis of Systemic Sclerosis (VEDOSS) criteria, b) presence of ILD on HRCT as determined by a senior radiologist. For every subject, we defined 1,355 robust radiomic features from HRCT images. The follow-up period was defined as the time interval between baseline visit and the last available follow-up visit.

We have developed a systematic computational workflow to build predictive ML models. To reduce the number of redundant radiomic features, we applied correlation thresholds. We applied distinct methods including 1) Lasso Penalized Regression for feature selection, and 2) Random Forest (RF) for modeling using the R package 'caret'. To select the optimal ML model, we randomly divided derivation cohort into Training (70%) and Holdout (30%) sets and applied fivefold cross-validation (5kCV) for feature and classifier selection on Training set only.

Results: We have investigated various methods to select the optimal set of predictive radiomic features. Since the ML model performance is affected by both, feature, and classifier selection, we assessed these factors first. Results from feature filtering and selection, suggested that the combination of correlation threshold of 0.9 with Lasso regression proved best. As we perform feature selection in 5k CV workflow, features present in at least 2 sets entered model optimization step.

During model selection, we selected RF classifier. We detected positive correlation between actual and predicted values with Spearman's \(\rho = 0.313\), \(p = 0.167\) and Spearman's \(\rho = 0.341\), \(p = 0.015\) in Oslo and Holdout sets respectively, as shown on Figure 1. The percentage of variance remained modest for both Holdout (Rsq = 0.104) and Oslo (Rsq = 0.126) datasets.

Conclusion: In summary, we: (1) developed ML workflow that allowed to select optimal methodology for modeling (i.e., feature and classifier selection), and (2) provide models that predicted time to individual lung function decline, characterized by significant correlation between predicted and actual values.

REFERENCES:
(3) Schniering J. et al. Resolving phenotypic and prognostic differences in interstitial lung disease related to systemic sclerosis by computed tomography-based radiomics. https://www.medrxiv.org/content/10.1101/2020.06.09.20124800v1

Disclosure of Interests: None declared

DOI: 10.1136/annrheumdis-2021-eular.2517

Figure 1. Performance of the best, RF classifier shown as scatterplot between actual and predicted values of individual time to lung decline.