Rheumatology, 2023, **00**, 1–8
https://doi.org/10.1093/rheumatology/kead332
Advance access publication 8 July 2023
Original Article

British Society for Rheumatology RHEUMATOLOGY OXFORD

Clinical science

Performance of serum biomarkers reflective of different pathogenic processes in systemic sclerosis-associated interstitial lung disease

Andrea-Hermina Györfi [1,2,*], Tim Filla[1,2], Nicholas Dickel[3], Florian Möller[4,5], Yi-Nan Li[1,2], Christina Bergmann[4,5], Alexandru-Emil Matei[1,2], Thomas Harrer[4,5], Meik Kunz[3,6,7], Georg Schett[4,5], Jörg H. W. Distler[1,2]

[1]Clinic for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine University, Düsseldorf, Germany
[2]Hiller Research Unit, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine University, Düsseldorf, Germany
[3]Chair of Medical Informatics, Friedrich-Alexander University (FAU) of Erlangen-Nürnberg, Erlangen, Germany
[4]Department of Internal Medicine 3, Rheumatology and Clinical Immunology, Friedrich-Alexander-University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany
[5]Deutsches Zentrum Immuntherapie (DZI), Friedrich-Alexander University (FAU) Erlangen-Nürnberg and Universitätsklinikum Erlangen, Erlangen, Germany
[6]Fraunhofer Institute for Toxicology and Experimental Medicine (ITEM), Hanover, Germany
[7]Fraunhofer Cluster of Excellence Immune-Mediated Diseases (CIMD), Hannover, Germany

*Correspondence to: Andrea-Hermina Györfi, Clinic for Rheumatology and Hiller Research Unit, University Hospital Düsseldorf, Heinrich-Heine University, 40225 Düsseldorf, Germany. E-mail: andrea-hermina.gyoerfi@med.uni-duesseldorf.de

Abstract

Objective: Interstitial lung disease (ILD) is the leading cause of mortality in SSc. Novel biomarkers are crucial to improve outcomes in SSc-ILD. We aimed to compare the performance of potential serum biomarkers of SSc-ILD that reflect different pathogenic processes: KL-6 and SP-D (epithelial injury), CCL18 (type 2 immune response), YKL-40 (endothelial injury and matrix remodelling) and MMP-7 (ECM remodelling).

Methods: Baseline and follow-up serum samples from 225 SSc patients were analysed by ELISA. Progressive ILD was defined according to the 2022-ATS/ERS/JRS/ALAT guidelines. Linear mixed models and random forest models were used for statistical analyses.

Results: Serum levels of KL-6 [MD 35.67 (95% CI 22.44–48.89, $P < 0.01$)], SP-D [81.13 (28.46–133.79, $P < 0.01$)], CCL18 [17.07 (6.36–27.77, $P < 0.01$)], YKL-40 [22.81 (7.19–38.44, $P < 0.01$)] and MMP-7 [2.84 (0.88–4.80, $P < 0.01$)] were independently associated with the presence of SSc-ILD. A machine-learning model including all candidates classified patients with or without ILD with an accuracy of 85%. The combination of KL-6 and SP-D was associated with the presence [0.77 (0.53–1.00, $P' < 0.01$)] and previous progression of SSc-ILD [OR 1.28 (1.01–1.61, $P' = 0.047$)]. Higher baseline levels of KL-6 [OR 3.70 (1.52–9.03, $P < 0.01$)] or SP-D [OR 2.00 (1.06–3.78, $P = 0.03$)] increased the odds of future SSc-ILD progression, independent of other conventional risk factors, and the combination of KL-6 and SP-D [1.109 (0.665–1.554, $P < 0.01$)] showed improved performance compared with KL-6 and SP-D alone.

Conclusion: All candidates performed well as diagnostic biomarkers for SSc-ILD. The combination of KL-6 and SP-D might serve as biomarker for the identification of SSc patients at risk of ILD progression.

Keywords: SSc, SSc-associated interstitial lung disease, serum biomarkers

Rheumatology key message

- The combination of KL-6 and SP-D could be used as a marker for progressive SSc-ILD.

Introduction

SSc is the autoimmune systemic rheumatic disease with the highest disease-related mortality and interstitial lung disease (ILD) is the leading cause of death in SSc [1]. ILD is a frequent organ manifestation in SSc; however, SSc-associated ILD (SSc-ILD) is highly heterogeneous [2]. Some patients remain stable without therapy or progress only slowly, while others experience rapid deterioration within few months despite immunosuppressive therapy [2]. The activity of SSc-ILD and its future course are difficult to predict. For instance, patients with earlier stages of ILD may be asymptomatic and might still have normal pulmonary function. Even patients who are symptomatic and progressive may only experience relatively minor changes in symptoms and changes in lung function

Received: 22 February 2023. **Accepted:** 6 June 2023

testing may be within the range of the intra-patient variability. High-resolution CT (HRCT) might be more sensitive in detecting previously progressive ILD, but is associated with X-ray exposure and thus requires careful indication making. Consequently, additional, sensitive markers of disease activity are required to predict future progression of SSc-ILD. Circulating biomarkers represent an attractive tool due to their high accessibility. Several serum biomarker candidates have been investigated for their potential to predict future development or progression of pre-diagnosed ILD in SSc [3–5].

Herein, we asked whether combinations of markers that reflect different underlying molecular processes may provide additional information over assessment of individual markers. We investigated the potential of serum Krebs von den Lungen 6 (KL-6), surfactant protein D (SP-D), CC-chemokine ligand 18 (CCL18), chitinase-3-like protein 1 (YKL-40) and matrix metalloproteinase 7 (MMP-7), either alone, or in combination, as diagnostic and prognostic biomarkers for SSc-ILD. Indeed, the biomarker candidates examined in our study reflect different underlying processes of the pathogenesis of SSc-ILD, which involve epithelial and endothelial cell injury, inflammation and fibrotic remodelling. While KL-6 and SP-D are expressed by lung epithelial cells and are released into the circulation upon lung epithelial injury, CCL18 is predominantly produced by pulmonary M2 macrophages in response to type 2 cytokines and interleukin 6 (IL-6) and acts as a chemotactic factor for other immune cells [3]. YKL-40 can be secreted by a variety of cell types, including airway epithelial cells, monocytes, macrophages, neutrophils and vascular smooth muscle cells and has been previously involved in endothelial cell dysfunction, inflammation and fibrotic tissue remodelling [6]. MMP-7 is secreted by alveolar macrophages and alveolar and bronchiolar epithelial cells and is known to degrade several extracellular matrix (ECM) components [7]. However, results related to their diagnostic or prognostic value in SSc-ILD are conflicting and further biased by the tendency to rather publish studies with positive results, while negative results often end up unpublished [4, 5, 8–11].

Patients and methods

Study design

We included patients over 18 years of age with a diagnosis of SSc according to the 2013 EULAR/ACR classification criteria, who presented to the rheumatology outpatient department of the University Hospital of Erlangen between 01/2018 and 12/2021. We excluded SSc patients with overlap syndromes with other rheumatic diseases. The cohort is described in the Results section and Supplementary Table S1 and S2 (available at *Rheumatology* online). The mean follow-up period was 2.0 ± 1.2 years. Follow-up visits occurred in 0.5 ± 0.3 year intervals. A full clinical examination according to the EUSTAR standards, blood prelevation with evaluation of routine parameters as well as determination of all serum candidate biomarkers were performed at each visit. Pulmonary function tests were performed at baseline as well as at least once per year during the follow-up period. HRCT was performed at baseline and was repeated as clinically indicated. Demographic, clinical and laboratory data were extracted from the clinical documentation system.

ILD was diagnosed by HRCT. Extensive ILD was defined by at least 20% lung involvement in HRCT and an FVC value (predicted) lower than 70% in indeterminate cases [12]. The diagnosis of ILD and the evaluation of the percent involvement in HRCT was performed by a radiology team consisting of one resident fellow and an attendant physician with long-standing experience in lung HRCT. Progression of ILD was defined according to the 2022 ATS/ERS/JRS/ALAT clinical practice guidelines by the presence of at least two of the following three criteria occurring within the past year with no alternative explanation: (i) worsening of respiratory symptoms; (ii) physiological evidence of disease progression, defined by an absolute decline in FVC $\geq$5% predicted or an absolute decline in DL_{CO} (corrected for Hb) $\geq$10% predicted; or (iii) radiological evidence of disease progression [13].

We defined 'previously progressive ILD' ('$prog_{previous}$ ILD') as ILD that progressed during the year prior to the time of biomarker measurement/blood sampling, according to the ATS/ERS/JRS/ALAT guidelines detailed above [13]. In contrast, patients who did not fulfil the criteria of previously progressive ILD were defined as having 'previously stable ILD' ('$stable_{previous}$ ILD'). 'Future progressive ILD' ('$prog_{future}$ ILD') was defined as ILD that progressed during the year after the date of biomarker measurement/blood sampling, according to the ATS/ERS/JRS/ALAT guidelines [13]. In contrast, patients with ILD who did not progress during the year after the date of biomarker measurement according to the ATS/ERS/JRS/ALAT guidelines were classified as having 'future stable ILD' ($stable_{future}$ ILD).

The study was approved by the ethical committee of Friedrich-Alexander University Erlangen-Nürnberg. All patients provided written informed consent to participate in the study.

Serum KL-6, SP-D, CCL18, YKL-40 and MMP-7 measurements

Blood samples were obtained during a routine blood withdrawal according to a standardized procedure and then centrifuged at room temperature for 10 min. Serum aliquots were stored at −70°C until assayed. Commercially available specific enzyme-linked immunosorbent assay (ELISA) kits were used to analyse the serum levels of circulating KL-6, SP-D, CCL18, YKL-40 and MMP-7 according to the manufacturer's protocols.

Statistical analysis

All statistical analyses were performed using R version 4.2.1 [14]. Descriptive statistics for cohort description were presented as absolute numbers and proportion for binary and nominal covariates. For continuous covariates, mean and standard deviation were used.

To assess differences in biomarker levels in SSc patients with ILD *vs* SSc patients without ILD, a linear mixed model with temporal autocorrelation implemented in R package lme [15] was used. In these models, we used one of the five candidate biomarkers as the dependent variable and ILD as the independent variable. We adjusted for the following confounders: age, sex, CRP, diffuse cutaneous involvement and anti-Scl70 antibodies [for YKL-40 additionally for pulmonary arterial hypertension (PAH) based on a previous report on increased YKL-40 levels in patients with PH] [16]. The set of confounders was chosen based on expert clinical knowledge and literature research [16, 17]. For evaluating the strength of candidate biomarker differences between groups,

95% Wald confidence intervals as well as *P*-values were calculated. All results were adjusted for multiple testing by using the Bonferroni–Holm method. We refer to results as statistically significant if the adjusted confidence intervals do not include 0 or 1 (for mean differences or odds ratios, respectively) and if the adjusted *P*-value is below 0.05.

A random forest model was trained for the prediction of ILD using our set of mechanistically complementary biomarkers. We used the R-package caret (version 6.0.93) to implement the model. Standard preprocessing methods like centering and scaling the data were applied. We split the data in 80% of samples for training the model and 20% of samples to test the model performance on unseen data. Repeated 10-fold cross validation was used to ensure a more robust training.

The association of each candidate biomarker with FVC or DL_{CO} (% predicted) was evaluated in a linear mixed effect model with FVC or DL_{CO} as the dependent variable and the biomarker candidate as the independent variable, adjusted for autocorrelation, age, sex, anti-Scl70 antibodies, diffuse cutaneous involvement, CRP and ILD.

For assessing the effect of each candidate biomarker value on previous or future progression of SSc-ILD, a logistic mixed effect model with temporal autocorrelation implemented in R package MASS was used [18]. Results were adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, anti-Scl70 antibodies, CRP and extensive ILD as well as for multiple testing [19].

In an exploratory analysis, we further analysed if a combination of two biomarker candidates is associated with the presence of ILD, previous or future progression of SSc-ILD, respectively. As these are measured on different scales, we divided each biomarker candidate by its root mean squared error. Then, we used all combinations of the product of two scaled candidate biomarkers as the dependent variable in the linear mixed model and adjusted for autocorrelation, age, sex, CRP, diffuse cutaneous involvement, anti-Scl70 antibodies (and for extensive ILD in the case of progression of SSc-ILD). To show that results were exploratory and thus not adjusted for multiple testing, we used the *P*-value notation *P*'. 95% confidence intervals were in this case also marked by '.

To analyse if there is any correlation between SSc therapy and the biomarker values, we used a linear mixed model approach with the biomarker value as the dependent variable and 'any therapy' (from MMF, nintedanib, methotrexate, tocilizumab and rituximab) or 'MMF therapy' as the independent variable and adjusted for autocorrelation, sex and age.

Further, we performed a sensitivity analysis for the models described above, in which we expanded the set of confounders with a variable defining treatment status.

We assumed all missing values as missing completely at random and decided to analyse all models using a complete case analysis for each biomarker candidate separately.

Results

Study cohort

We included 225 SSc patients, with a total of 1259 visits: 148 patients without ILD and 94 patients with ILD, of which 17 (18.1%) patients developed ILD during the course of this study. Almost one third of the patients with ILD (27.4%) had extensive disease. One third of the SSc-ILD patients progressed during the study despite immunosuppressive and/or antifibrotic therapy. Demographic, clinical and treatment characteristics of the patients at baseline are illustrated in Supplementary Table S1, available at *Rheumatology* online. The clinical characteristics of the patients with previously progressive SSc-ILD ($prog_{previous}$ ILD) and of patients with previously stable SSc-ILD ($stable_{previous}$ ILD) are presented in Supplementary Table S2, available at *Rheumatology* online.

Higher serum levels of KL-6, SP-D, CCL18, YKL-40 and MMP-7 are associated with SSc-ILD

We first evaluated the serum levels of each biomarker candidate in SSc patients with ILD and compared them with patients without ILD. We found significantly higher serum levels of all five biomarker candidates in SSc-ILD patients compared with SSc-non ILD patients [KL-6: mean difference 35.67 (95% confidence interval 22.44–48.89, $P < 0.01$); SP-D: 81.13 (28.46–133.79, $P < 0.01$); CCL18: 17.07 (6.36–27.77, $P < 0.01$); YKL-40: 22.81 (7.19–38.44, $P < 0.01$) and MMP-7: 2.84 (0.88–4.80, $P < 0.01$)] (Fig. 1 and Supplementary Fig. S1, available at *Rheumatology* online).

We further evaluated whether the serum levels of the five biomarker candidates are influenced by the treatment that the SSc patients received during the follow-up period. We observed a statistically significant correlation between the serum levels of YKL-40 and immunosuppressive or antifibrotic therapies in general (any of MMF, methotrexate, nintedanib, tocilizumab and rituximab), as well as between YKL-40 and MMF in particular (Supplementary Table S3, available at *Rheumatology* online). The serum levels of the other biomarker candidates were not influenced by therapy. Furthermore, we compared the serum candidate biomarker levels between patients with and without ILD while additionally adjusting for therapy and observed very similar, statistically significant differences in all five candidate biomarkers between the two groups, highlighting that the therapy does not confound the association of all five biomarkers with the presence of SSc-ILD (Supplementary Table S4, available at *Rheumatology* online).

Previous studies showed higher serum levels of YKL-40 in SSc patients with PAH, or with arthritis, compared with SSc patients without PAH or arthritis, respectively [16, 20]. We also observed significantly higher serum levels of YKL-40 in SSc patients with PAH compared with patients without PAH. However, SSc-ILD patients had significantly higher levels of YKL-40 compared with patients without ILD, even after adjusting for PAH.

We further trained a random forest model using repeated 10-fold cross validation and a data split of 80/20 to predict the presence of ILD using all five biomarker candidates. Our model was statistically significant and showed an accuracy of 85% in the prediction of unseen test data. Consistent with our previous results, KL-6 and SP-D were the two candidate biomarkers that contributed most to prediction in the model (Supplementary Fig. S2, available at *Rheumatology* online).

We also evaluated the association between every possible combination of two biomarker candidates and the presence of ILD. We observed a significant association between the combination of KL-6 and SP-D [0.77 (0.53–1.00, $P' < 0.01$)] and ILD (Supplementary Table S5, available at *Rheumatology* online).

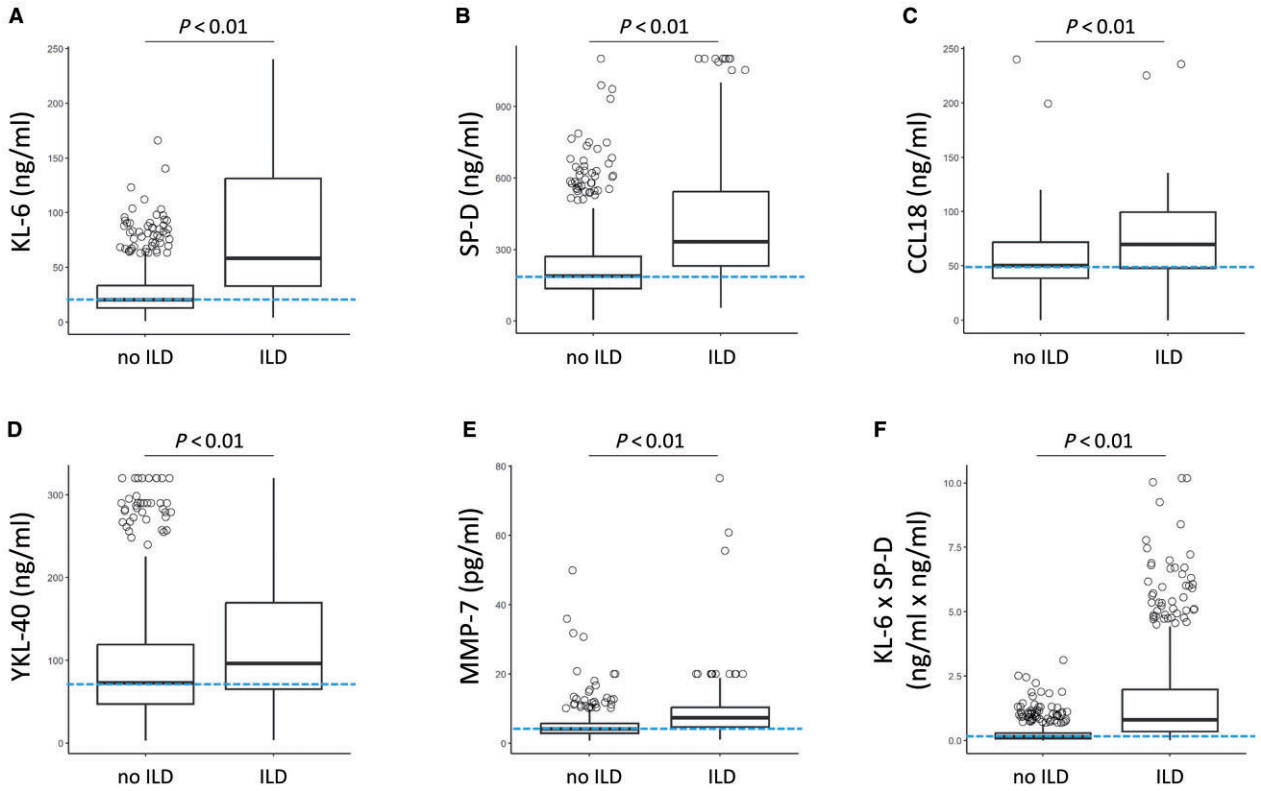

Figure 1. Serum levels of KL-6, SP-D, CCL18, YKL-40 and MMP-7 in SSc patients with and without ILD. Boxplots showing serum levels of KL-6 (ng/ml) (**A**), SP-D (ng/ml) (**B**), CCL18 (ng/ml) (**C**), YKL-40 (ng/ml) (**D**), MMP-7 (pg/ml) and the combination of KL-6 and SP-D (ng/ml × ng/ml) (**E**) in SSc patients without and with interstitial lung disease (ILD). A linear mixed model analysis adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, CRP and the presence of anti-Scl70 antibodies was performed [for YKL-40 additionally for pulmonary arterial hypertension (PAH)]. Adjusted *P*-values are provided. CCL18: CC chemokine ligand 18; KL-6: Krebs-von-den-Lungen-6; MMP-7: matrix metalloproteinase 7; SP-D: surfactant protein D; YKL-40: chitinase-3-like-protein 1

We did not find associations of any of the five biomarker candidates with extensive SSc-ILD, in a linear mixed model adjusted for anti-topoisomerase I antibodies (ATA), diffuse cutaneous involvement and FVC. However, we did find an inverse association between YKL-40 [−0.024 (−0.036, −0.011, $P < 0.01$)] and FVC values (predicted) as well as between SP-D [−0.005 (−0.009, −0.001, $P = 0.02$)] or MMP-7 [−0.132 (−0.254, −0.009, $P = 0.03$)] and DL_{CO} values (predicted) in a linear mixed effect model with FVC or DL_{CO} as the dependent variable and the biomarker candidates as the independent variable, adjusted for autocorrelation, age, sex, ATA, diffuse cutaneous involvement and the presence of ILD (Supplementary Table S6, available at *Rheumatology* online).

Of note, we observed a low to moderate correlation between all five biomarker candidates, the highest of which was observed between KL-6 and SP-D (r = 0.5) (Supplementary Fig. S3, available at *Rheumatology* online).

Higher serum levels of KL-6 and SP-D are associated with previous progression of SSc-ILD

We next tested whether these markers are associated with the presence of $prog_{previous}$ ILD in SSc. Higher serum levels of KL-6 [OR = 1.68 (0.87–3.25, $P = 0.23$)] or SP-D [OR= 1.25 (0.75–2.07, *P*=ns)] were associated with trends towards higher odds of $prog_{previous}$ ILD (Fig. 2 and Supplementary Fig. S4, available at *Rheumatology* online).

Furthermore, we compared the serum candidate biomarker levels between patients with $prog_{previous}$ ILD and patients with $stable_{previous}$ ILD, while additionally adjusting for therapy and also observed similar trends of higher serum levels of KL-6 or SP-D towards higher odds of previous progression of ILD (Supplementary Table S7, available at *Rheumatology* online).

In a logistic mixed model adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, ATA, CRP and extensive ILD, an increase in the combination of KL-6 and SP-D significantly increased the odds of $prog_{previous}$ ILD [OR 1.28 (1.01–1.61, *P*'=0.047)] (Fig. 2 and Supplementary Table S8, available at *Rheumatology* online).

Higher baseline serum levels of KL-6 and SP-D, individually and in combination, are associated with future progression of SSc-ILD

We observed that higher baseline levels of KL-6 [OR 3.70 (1.52–9.03, $P < 0.01$)] or SP-D [OR 2.00 (1.06–3.78, $P = 0.03$)] were significantly associated with future SSc-ILD progression (Fig. 3 and Supplementary Fig. S5, available at *Rheumatology* online).

We also compared the serum candidate biomarker levels between patients with future progressive SSc-ILD ($prog_{future}$ ILD) and patients with future stable SSc-ILD ($stable_{future}$ ILD), while additionally adjusting for therapy and observed similar results (Supplementary Table S9, available at *Rheumatology* online).

Furthermore, the combination of KL-6 and SP-D showed an additional effect compared with KL-6 and SP-D alone

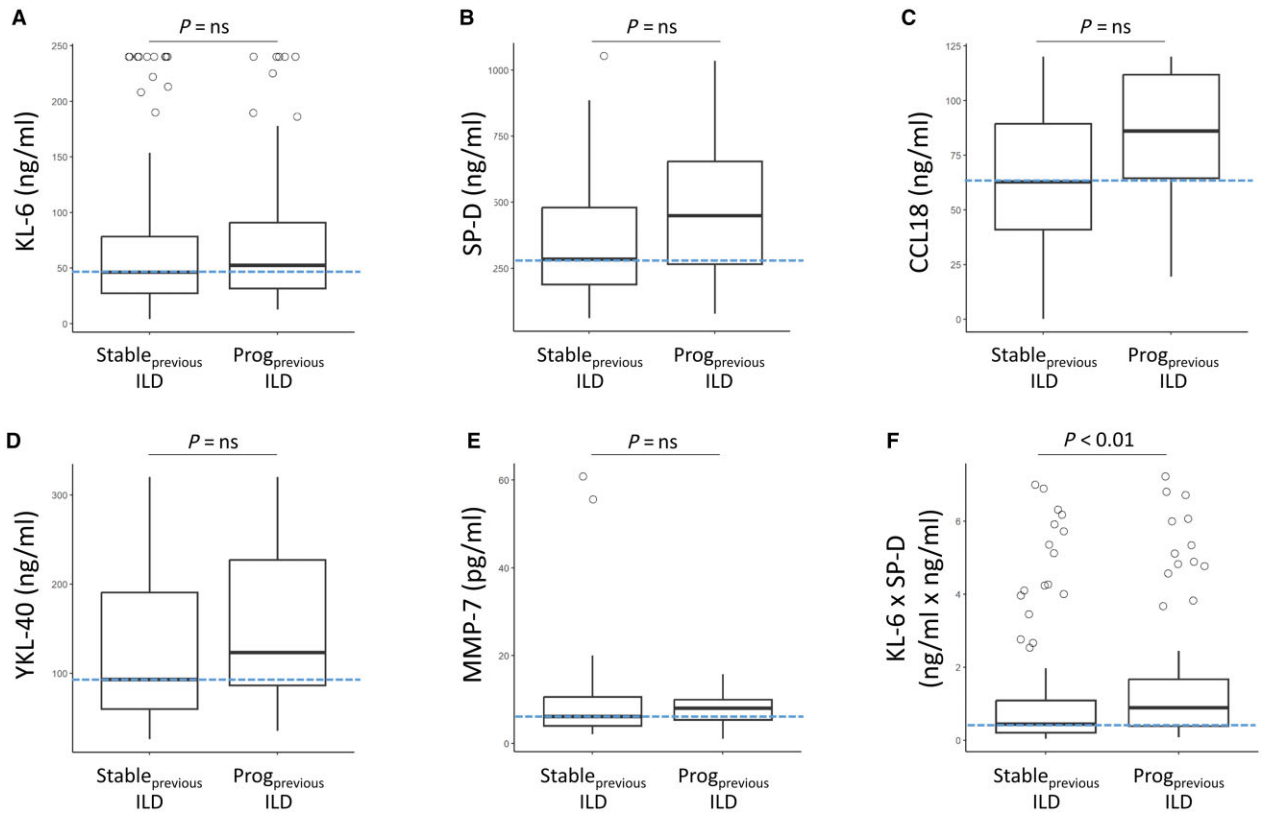

Figure 2. Serum levels of KL-6, SP-D, CCL18, YKL-40 and MMP-7 in SSc patients with previously progressive *vs* previously stable SSc-ILD. Boxplots showing serum levels of KL-6 (ng/ml) (**A**), SP-D (ng/ml) (**B**), CCL18 (ng/ml) (**C**), YKL-40 (ng/ml) (**D**), MMP-7 (pg/ml) (**E**) and of the combination of KL-6 and SP-D (ng/ml × ng/ml) (**F**) in SSc patients with previously progressive ILD ($prog_{previous}$ ILD) compared with patients with previously stable ILD ($stable_{previous}$ ILD). (**A–E**) A logistic mixed model analysis adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, CRP, extensive disease and the presence of anti-Scl70/topoisomerase I antibodies was performed. (**F**) A logistic mixed model adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, CRP, extensive disease and anti-Scl70/topoisomerase I antibodies was performed. Adjusted *P*-values are provided. CCL18: CC chemokine ligand 18; ILD: interstitial lung disease; KL-6: Krebs-von-den-Lungen-6; MMP-7: matrix metalloproteinase 7; SP-D: surfactant protein D; YKL-40: chitinase-3-like-protein 1

(Fig. 3 and Supplementary Table S10, available at *Rheumatology* online).

A total of 16.1% of all patients had progression of skin fibrosis during our study. Of these, 20.5% developed progression of both skin and lung fibrosis. Three patients had progressive myocardial fibrosis and concomitant progression of either skin or lung fibrosis. However, we did not observe differences in the serum levels of any of the investigated biomarker candidates between patients with progressive skin fibrosis compared with patients with stable skin disease (Supplementary Fig. S6, available at *Rheumatology* online).

Discussion

In this study we investigated the potential of five serum candidate biomarkers (KL-6, SP-D, CCL18, YKL-40 and MMP-7), reflective of different pathogenic processes of SSc-ILD, either alone, or in combination, as diagnostic and prognostic biomarkers for SSc-ILD in a well characterized single centre cohort of 225 SSc patients, with a total of 1259 visits, over a follow-up period of up to 3.2 years.

We identified that levels of all five circulating biomarker candidates (KL-6, SP-D, CCL18, YKL-40, MMP-7) were significantly higher in our cohort in SSc patients with ILD compared with patients without ILD. These data support the use of the five serum biomarker candidates for the diagnosis of ILD in SSc patients. We also evaluated the association between each combination of the candidate biomarkers and ILD and observed that from all candidate biomarker combinations only the combination of KL-6 and SP-D was significantly associated with SSc-ILD. In contrast with previous studies, we did not find any association between the serum levels of KL-6, SP-D, CCL18 or MMP-7 and the extent of ILD in HRCT [4, 8, 11, 21]. Moreover, there was also no association between the circulating levels of YKL-40 and ILD extent in our cohort. Nevertheless, we found an association between YKL-40 and FVC (% predicted) as well as between SP-D and MMP-7 and DL_{CO} (% predicted).

Serum levels of SP-D and KL-6 were higher in our cohort in patients with previous progression of SSc-ILD; however, results did not reach statistical significance. These results are consistent with previous studies on SP-D, but inconsistent with previous reports on KL-6 [4, 8, 11]. One reason for this could be the lack of a standardized definition of progressive SSc-ILD. Indeed, our study uses the latest definition of progressive ILD, which captures smaller changes in lung function parameters [13]. An increase in the combination of KL-6 and SP-D was associated with $prog_{previous}$ ILD. Furthermore, higher KL-6 or SP-D baseline levels significantly increased the odds of future SSc-ILD progression. We observed an additional effect of the combination of KL-6 and SP-D over the

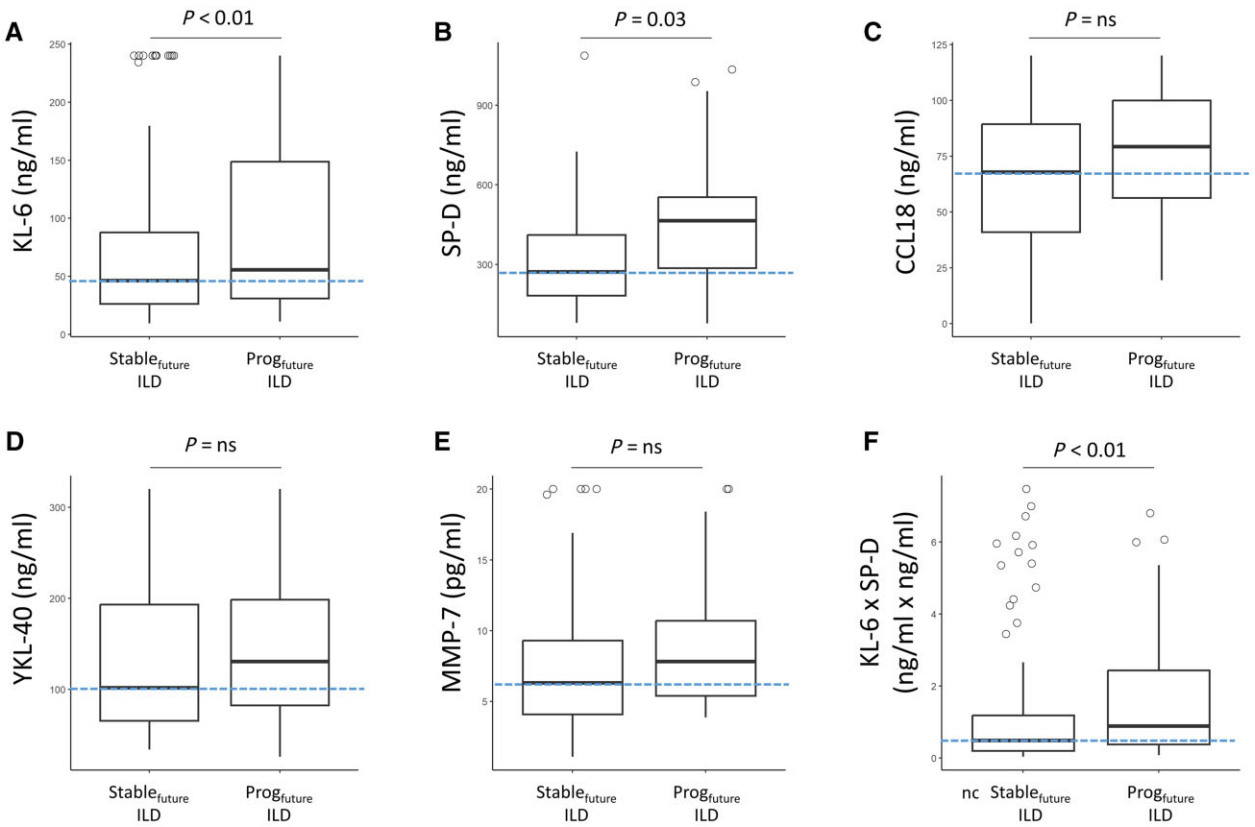

Figure 3. Serum levels of KL-6, SP-D, CCL18, YKL-40 and MMP-7 in SSc patients with future progression of ILD compared with patients with future stable ILD. Boxplots showing serum levels of KL-6 (ng/ml) (**A**), SP-D (ng/ml) (**B**), CCL18 (ng/ml) (**C**), YKL-40 (ng/ml) (**D**), MMP-7 (pg/ml) (**E**) and of the combination of KL-6 and SP-D (ng/ml × ng/ml) (**F**) in SSc patients with future progression of ILD ($prog_{future}$ ILD) compared with patients with future stable ILD ($stable_{future}$ ILD). (**A–E)** A logistic mixed model analysis adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, CRP, extensive disease and the presence of anti-Scl70 antibodies was performed. (**F**) A logistic mixed model adjusted for autocorrelation, age, sex, diffuse cutaneous involvement, CRP, extensive disease and anti-Scl70 antibodies was performed. Adjusted *P*-values are provided. CCL18: CC chemokine ligand 18; ILD: interstitial lung disease; KL-6: Krebs-von-den-Lungen-6; MMP-7: matrix metalloproteinase 7; SP-D: surfactant protein D; YKL-40: chitinase-3-like-protein 1

individual biomarker candidates; however, the effect was not statistically significant.

Our study has limitations. Although this is a large study cohort, all patients are from a single centre and validation in an external validation cohort would further strengthen our conclusions. Despite the inclusion of a relatively large number of patients in our study, further studies in larger multicentre cohorts are necessary to evaluate the predictive value of KL-6 and SP-D, individually or in combination, for future SSc-ILD progression. Moreover, despite the relative long follow-up period of our study of up to 3.2 years, further studies with longer follow-up periods are necessary to evaluate the long-term outcome of SSc-ILD.

However, our study also has several strengths. It provides data on five circulating biomarker candidates that reflect different underlying pathogenic processes of SSc-ILD in a well-characterised SSc cohort. Our study focuses not only on single markers, but also evaluates combinations of biomarker candidates that reflect different underlying pathophysiological processes. In contrast with other studies, we provide follow-up data on these five circulating biomarker candidates for most of the patients. Our study confirms previous results on the diagnostic value of KL-6, SP-D and CCL18 for SSc-ILD and provides evidence for the use of YKL-40 as a diagnostic biomarker for SSc-ILD. Our data also supports the use of MMP-7 for the diagnosis of SSc-ILD, as a less well-studied circulating biomarker candidate in SSc.

Moreover, our study is the first one to show an association between the combination of serum KL-6 and SP-D and previous and future progression of SSc-ILD, respectively. However, future studies are needed to investigate its predictive role for the progression of SSc-ILD.

We applied machine learning methods and employed a random forest model to classify ILD in patients with SSc and achieved a high accuracy. Random forest is a robust method that can even handle small or medium sample sizes and high data dimensionality, where linear mixed model can deal well with dependent data features. Moreover, random forest model provides feature importance measures, which helps to rank the different parameters by importance. Our ML-based framework is thus a suitable approach for studying biomarkers in orphan diseases with limited patient number such as SSc, thus paving the way for further validation studies.

In summary, our study demonstrates that KL-6, SP-D, CCL18, YKL-40 and MMP-7 can all serve as independent diagnostic biomarkers of SSc-ILD. KL-6 and SP-D can also be used for the stratification of SSc patients at risk for future SSc-ILD progression. We demonstrate that the combination of KL-6 and SP-D is superior to the individual biomarkers for the risk evaluation of future SSc-ILD progression. We thus provide evidence that the combination of two biomarkers reflective of lung epithelial cell damage can better evaluate the risk of SSc-ILD progression, compared with biomarkers of endothelial cell dysfunction, macrophage activation or

extracellular matrix remodelling. Further studies are required to evaluate the use of biomarker combinations for prediction of future SSc-ILD progression and outcome.

Supplementary material

Supplementary material is available at *Rheumatology* online.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files). Additional supporting information or raw data are available as de-identified participant datasets from the corresponding author upon request.

Author contributions

Conceptualization: A.-H.G., J.H.W.D. Methodology: A.-H.G., T.F., N.D., M.K., T.H., G.S., J.H.W.D. Investigation: A.-H.G., T.F., N.D., F.M., Y.-N.L., C.B., A.-E.M. Visualization: A.-H.G., T.F., N.D. Funding acquisition: A.-H.G., A.-E.M., M.K., G.S., J.H.W.D. Project administration: A.-E.M., T.H. Supervision: A.-H.G., M.K., J.H.W.D. Writing – original draft: A.-H.G., J.H.W.D. Writing – review and editing: all authors. All authors revised the article and read and approved the final version before submission.

Funding

This work was supported by the German Research Foundation (DFG) {DI 1537/14–1, DI 1537/17–1, DI 1537/20–1, DI 1537/22–1, DI 1537/23–1, SFB CRC1181 (project C01) and SFB TR221 [project number 324392634 (B04)] to J.H.W.D.}, the Wilhelm-Sander-Foundation (2013.056.1 to J.H.W.D.), the Else-Kröner-Fresenius-Foundation (EKFS) (2014_A47 and 2014_A184 to J.H.W.D.), the German Federal Ministry of Education and Research (BMBF) [MASCARA program, TP 2 (01EC1903A)] and a Career Support Award of Medicine of the Ernst Jung Foundation to J.H.W.D. Further support was provided by the BMBF (CompLS program, grant 031L0262C to N.D., M.K.), the German Research Foundation [SFB TR221, project number 324392634 to M.K., MA 9219/2–1 to A.-E.M., GRK 2740 (project B02) to M.K.], the Else-Kröner-Fresenius-Foundation (2021_EKEA.03 and 2022_EKMS.02 to A.-E.M.) and ELAN-Foundation Erlangen (19–12-06–1 to A.E.M. and 21–07-23–1 to A.-H.G.).

Disclosure statement: The authors declare following conflicts of interest: J.H.W.D. has consultancy relationships with AbbVie, Active Biotech, Anamar, ARXX, AstraZeneca, Bayer Pharma, Boehringer Ingelheim, Celgene, Galapagos, GSK, Inventiva, Janssen, Novartis, Pfizer, and UCB. J.H.W.D. has received research funding from Anamar, ARXX, BMS, Bayer Pharma, Boehringer Ingelheim, Cantargia, Celgene, CSL Behring, Galapagos, GSK, Inventiva, Kiniksa, Sanofi-Aventis, RedX, UCB. J.H.W.D. is stock owner of 4D Science and scientific lead of FibroCure. M.K. is founder and shareholder of BioInf4Life UG. A.-H.G. has received honoraria from Deutsches Zentrum für Lungenforschung (DZL)/BREATH. T.F., N.D., F.M., Y.-N.L., C.B., A.-E.M., T.H. and G.S. do not declare any potential conflicts of interest.

Acknowledgements

We thank Wolfgang Espach, Vladyslav Fedorchenko, Lena Summa, Christoph Liebel and Regina Kleinlein for excellent technical assistance.

References

1. Garen T, Lerang K, Hoffmann-Vold AM *et al.* Mortality and causes of death across the systemic connective tissue diseases and the primary systemic vasculitides. Rheumatology (Oxford) 2019;58:313–20.
2. Vonk MC, Walker UA, Volkmann ER *et al.* Natural variability in the disease course of SSc-ILD: implications for treatment. Eur Respir Rev 2021;30:200340.
3. Hoffmann-Vold AM, Fretheim H, Meier C, Maurer B. Circulating biomarkers of systemic sclerosis - interstitial lung disease. J Scleroderma Relat Disord 2020;5:41–7.
4. Elhai M, Hoffmann-Vold AM, Avouac J *et al.* Performance of candidate serum biomarkers for systemic sclerosis-associated interstitial lung disease. Arthritis Rheumatol 2019;71:972–82.
5. Salazar GA, Kuwana M, Wu M *et al.* KL-6 but not CCL-18 is a predictor of early progression in systemic sclerosis-related interstitial lung disease. J Rheumatol 2018;45:1153–8.
6. Lee CG, Da Silva CA, Dela Cruz CS *et al.* Role of chitin and chitinase/chitinase-like proteins in inflammation, tissue remodeling, and injury. Annu Rev Physiol 2011;73:479–501.
7. Elkington PT, Friedland JS. Matrix metalloproteinases in destructive pulmonary pathology. Thorax 2006;61:259–66.
8. Volkmann ER, Tashkin DP, Kuwana M *et al.* Progression of interstitial lung disease in systemic sclerosis: the importance of pneumoproteins Krebs von den Lungen 6 and CCL18. Arthritis Rheumatol 2019;71:2059–67.
9. Kennedy B, Branagan P, Moloney F *et al.* Biomarkers to identify ILD and predict lung function decline in scleroderma lung disease or idiopathic pulmonary fibrosis. Sarcoidosis Vasc Diffuse Lung Dis 2015;32:228–36.
10. Yamakawa H, Hagiwara E, Kitamura H *et al.* Serum KL-6 and surfactant protein-D as monitoring and predictive markers of interstitial lung disease in patients with systemic sclerosis and mixed connective tissue disease. J Thorac Dis 2017;9:362–71.
11. Elhaj M, Charles J, Pedroza C *et al.* Can serum surfactant protein D or CC-chemokine ligand 18 predict outcome of interstitial lung disease in patients with early systemic sclerosis? J Rheumatol 2013;40:1114–20.
12. Goh NS, Desai SR, Veeraraghavan S *et al.* Interstitial lung disease in systemic sclerosis: a simple staging system. Am J Respir Crit Care Med 2008;177:1248–54.
13. Raghu G, Remy-Jardin M, Richeldi L *et al.* Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline. Am J Respir Crit Care Med 2022;205:e18–e47.
14. RC Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2022. https://www.R-project.org/ (9 May 2023, date last accessed).
15. Pinheiro JB, R Core Team nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1–157. 2022. https://CRAN.R-project.org/package=nlme (9 May 2023, date last accessed).
16. Furukawa T, Matsui K, Kitano M *et al.* Relationship between YKL-40 and pulmonary arterial hypertension in systemic sclerosis. Mod Rheumatol 2019;29:476–83.
17. Distler O, Assassi S, Cottin V *et al.* Predictors of progression in systemic sclerosis patients with interstitial lung disease. Eur Respir J 2020;55:1902026.
18. Venables WN, Ripley BD. Modern applied statistics with S. 4th edn. New York: Springer, 2022.

19. Khanna D, Tashkin DP, Denton CP *et al.* Etiology, risk factors, and biomarkers in systemic sclerosis with interstitial lung disease. Am J Respir Crit Care Med 2020;201:650–60.
20. Karalilova R, Kazakova M, Sapundzhieva T *et al.* Serum YKL-40 and IL-6 levels correlate with ultrasound findings of articular and periarticular involvement in patients with systemic sclerosis. Rheumatol Int 2019;39:1841–8.
21. Moinzadeh P, Krieg T, Hellmich M *et al.* Elevated MMP-7 levels in patients with systemic sclerosis: correlation with pulmonary involvement. Exp Dermatol 2011;20:770–3.