Systemic sclerosis: can we identify patients at risk?

Leeuwen, N.M. van

Chapter 5

A new risk model is able to identify patients with a low risk of progression in Systemic Sclerosis

Nina M. van Leeuwen, Marc P. Maurits, Sophie I.E. Liem, Jacopo Ciaffi, Nina Ajmone Marsan, Maarten Ninaber, Cornelia F. Allaart, Henrike Gillet- van Dongen, Robbert Goekoop, Tom J. Huizinga, Rachel Knevel, Jeska K. de Vries-Bouwstra

Published RMD open

Objectives: To develop a prediction model to guide annual assessment of systemic sclerosis (SSc) patients tailored in accordance to disease activity.

Methods: A machine learning approach was used to develop a model that can identify patients without disease progression. SSc patients included in the prospective Leiden SSc cohort and fulfilling the ACR/EULAR 2013 criteria were included. Disease progression was defined as progression in ≥1 organ system, and/or start of immunosuppression or death. Using elastic-net-regularization, and including 90 independent clinical variables (100% complete), we trained the model on 75% and validated it on 25% of the patients, optimizing on negative predictive value (NPV) to minimize the likelihood of missing progression. Probability cutoffs were identified for low and high risk for disease progression by expert assessment.

Results: Of the 492 SSc patients (follow-up range: 2-10yrs), disease progression during follow-up was observed in 52% (median time 4.9yrs). Performance of the model in the test set showed an AUC-ROC of 0.66. Probability score cutoffs were defined: low risk for disease progression (<0.197, NPV:1.0; 29% of patients), intermediate risk (0.197-0.223, NPV:0.82; 27%) and high risk (>0.223, NPV:0.78; 44%). The relevant variables for the model were: previous use of cyclophosphamide or corticosteroids, start with immunosuppressive drugs, previous gastrointestinal progression, previous cardiovascular event, pulmonary arterial hypertension, modified Rodnan Skin Score, creatinine kinase, and diffusing capacity for carbon monoxide.

Conclusion: Our machine-learning-assisted model for progression enabled us to classify 29% of SSc patients as 'low risk'. In this group annual assessment programs could be less extensive than indicated by international guidelines.

INTRODUCTION

Systemic Sclerosis (SSc) is a heterogeneous disease. The spectrum of the disease ranges from rapidly progressive, with generalized fibrosis of the skin and the vital organs to a more indolent form developing over an extended period of time (1). The amount of patients with progression of SSc is substantial and progression occurs most often in early disease (2, 3). It is important to note that around 50% of patients will never show any signs of progression. To accurately assess the trajectory of the disease, several studies addressed identification of risk factors of future skin and organ progression in different SSc subpopulations (4, 5). Existing prediction models in SSc are often based on a subset of SSc patients, and do not capture the whole population (2, 6). Prediction of the disease course remains challenging in the individual patient which raises the questions whether personalized prediction in the heterogeneous SSc population is actually feasible.

For physicians in clinical practice, it is important to have clear guidance regarding intensity and frequency of follow-up, not only to identify disease progression timely but also to limit excessive diagnostics in mild SSc patients. Currently, no evidence based international guidelines for follow-up of SSc exist, except for the ESC/ERS guideline recommending annual echocardiography for detection of pulmonary arterial hypertension (PAH) (7). In 2019, an international standard for longitudinal follow-up describing points to address in annual assessment of patients with SSc was developed based on Delphi-expert consensus. Overall, 55 items were identified including clinical assessments, laboratory measurements, imaging and functional investigations (8). Whether the identified items are sufficient to identify disease progression timely in all patients is yet to be determined. Moreover, in some patients with mild disease, annual follow-up might even be more concise and assessing 55 tools on an annual basis might not be necessary. Of note, previous prediction studies concercing prevalent SSc cohorts might have underestimated progression in SSc by failing to capture the early rapid progressors (9). On the other hand, with the introduction of the ACR/EULAR 2013 criteria additional cases with less severe disease might be identified (10). Together, these observations provide the rationale for the design of data-driven recommendations that describe tailormade systematic assessments for individual SSc patients in line with their individual disease course.

Our prospective SSc cohort includes both mild and severe patients who undergo annual assessment, as the health care system in The Netherlands is characterized by high accessibility. Starting from 2009 all patients fulfilling Leroy criteria for early SSc have been included (11, 12). In the current study, we included detailed information on disease progression in our prospective cohort and we addressed an important limitation that is often encountered when searching for predictive factors in large datasets from SSc patient registries: the high incidence of missing data. Therefore, in our prediction model, we only included patients with complete data available on at least three visits.

With this study we aim to develop a tailormade model to guide annual assessment in individual SSc patients, with a special focus on patients with a low risk of disease progression in whom annual extensive investigations may be considered redundant. To address this we: 1) determined the proportion of patients without disease progression, 2) applied machine learning to build a prediction model in the patients with complete data available at ≥ 3 time points to predict lack of progression. Additionally, 3) we evaluated a second prediction model including the variables from the Delphi consensus guideline and compared the performance of this model to the Machine-Learning-Assisted model in order to assess which investigations are minimally needed to identify patients with a low risk of disease progression.

METHOD

Patient selection

In the Leiden University Medical Center (LUMC), all SSc patients, with a range of disease severity from mild to very severe, undergo annual extensive screening during a 1 to 2 day health care program (combined care in systemic sclerosis (CCISS)). This includes a detailed physical examination, modified Rodnan skin score (mRSS) assessment (13), laboratory testing (with autoantibody screening at baseline), electrocardiography (ECG), pulmonary function test, optional echocardiography (mandatory at baseline visit), optional 24-hour Holter ECG monitoring (mandatory at baseline visit), optional cardiopulmonary exercise tests (CPET) and optional high-resolution computed tomography (HRCT) (mandatory at baseline visit). Patients are requested to complete various questionnaires at every visit (14-22). Additionally, at every visit, blood and serum samples are collected and stored in the Leiden Scleroderma Biobank.

For the first part of the study (i.e. numbers of disease progressors), SSc patients who fulfilled the 1) ACR 1980 and/or LeRoy (from 2009-2013) criteria or the 2) ACR/EULAR 2013 and/or Leroy criteria and had at least two assessments were included (12, 23, 24). For the second part of the study, including the Machine-Learning-Assisted prediction model, we analysed SSc patients who fulfilled the ACR/EULAR 2013 criteria and had at least three complete visits (a third visit was necessary for our primary outcome). This ensured that included patients had complete data available of at least three time points. A complete visit consisted of at least a physical examination (including mRSS), laboratory testing, a pulmonary function test, ECG, HRCT and a transthoracic echocardiography. The strict inclusion criteria were necessary to limit the amount of missing data on important organ systems and only patients who underwent complete screening of organ systems were included in the model in order to minimalize the likelihood of missing any important organ involvement.

The cohort study was designed in accordance with the ethical principles of the Declaration of Helsinki. All patients gave written informed consent. Collection and analysis of data have been approved by the local ethics committee (Leiden CME number B16.037).

Outcomes

Disease progression was defined as progression in one or more organ systems; pulmonary, cardiac, gastro-intestinal, skin, renal, and/or myositis (supplementary table S1 for detailed explanation). For pulmonary, PAH, skin and renal crisis, progression was defined as described previously (10, 25, 26). Cardiac progression, gastro-intestinal progression and myositis were each defined using a combination of variables and based on consensus among authors. Use of immunomodulatory medication was recorded at every visit and included: cyclophosphamide, methotrexate, myocophenolate mofetil (MMF), azathioprine, corticosteroids, hydroxychloroquine and stem cell transplantation. Patients included in clinical trials (Resolve [lenabasum], Senscis [nintedanib], FocuSSced [Tocilizumab], ASTIS [autologous stemcell transplantion), RITIS [rituximab], ASTIS [stem cell transplantation] were also captured. Use of biologicals outside of trials was observed in <0.5% of the patients, therefore these were not depicted separately but were included in the primary outcome (27-31). The primary endpoint in the prediction model was defined as progression in ≥ 1 organ system, and/or start of immunosuppression (IS) or death between the two most recent visits.

Predictors

The included predictors in the Machine-Learning-Assisted model were selected based on the predictors identified by experts (8), additional predictors were selected based on clinical expertise and current literature. In order to prevent exclusion of too many patients due to missingness, we dropped four variables (out of 94) with a missingness percentage > 5% (nailfold videocapillaroscopy, of which annual collection started in 2013, and 3 variables derived from the UCLA GIT questionnaire, namely fecal soiling, diarrhea and distension/bloating, of which annual collection started in 2013). This resulted in 90 independent variables (100% complete in n=248 patients) to predict progression at the final assessment. The 90 variables included in our model -all 100% complete- are described in the supplementary Table S2. A timeline of the study can be found in the supplementary file (Figure S1).

Statistics

For the first part of the study we used descriptive statistics to evaluate the number of disease progressors in SSc during 10 years of follow-up. For the second part of the study, the development of the prediction model, we applied a machine learning approach known as “elastic net regularization”. The elastic net performs simultaneous regularization and variable selection in order to reduce variance with minimal risk of bias (32). Independent variables were all the variables collected during follow-up visits [predictors] until the prediction visit [event visit= primary outcome]. Disease progression on any organ system during follow-up was also included as independent variable in different manners: 1) progression between third **to** event visit and the prediction visit (dichotomous), 2) progression developed before the prediction visit (dichotomous), 3) the amount of times progression occurred between baseline visit and the prediction visit (quantitative), and 4) in how many organ systems progression occurred over time (quantitative). Including progression as an independent variable mimics the decision making of the physician in clinical care, where decisions regarding follow-up are made based on previous information (including progression occurring years before the current visit). Given the extensive amount of information from previous visits we examined whether these data could predict the development of progression at the final [event] visit. The dependent variable in the model was defined as progression in ≥ 1 organ system, and/or start of IS at the last recorded visit, or death after a complete baseline visit. All patients without any progression during follow-up were identified as 'nonprogressors'. In order to preserve the maximum number of patients, we filtered variables on 95% call rate prior to deleting incomplete cases, variables with more > 5% missingness were checked in the EPD and in case of true missingness deleted from the dataset. To develop the model using leave-one-out cross validation and independently validate the final model's performance, the included patients were randomly split in a training (75%) and a test (25%) set. The model was developed and optimized on the larger training set and subsequently applied to the test set with the lambda, alpha and coefficients set to the identified optima. The chosen predictive variables were entirely based on the training data. Risk probability scores and AUC-ROC were created (based on test data) to identify optimal cut-offs for the risk on disease progression.

Due to the detrimental impact of undertreatment in SSc, we opted to maximize negative predictive value (NPV) with a constraint on sensitivity, which minimizes the likelihood of missing progression. However, a benefit of the probabilistic nature of a prediction model is the flexibility of the case cut-off point; by sliding across the ROC curve one can choose to prioritize any preferred performance metric. Cut-offs for risk probability scores were defined based on the test characteristics (maximize NPV) and the distribution of probabilities plots in the test set. After cut-off selection, we ran a post-hoc analysis to assess the missed progressors.

Lastly, using logistic regression we built a prediction model including 51 (out of the 55) variables from the Delphi consensus guideline (8). We had to exclude leg edema, urine analysis, liver function test and New York Heart Association (NYHA) class due to missingness in > 5%. The performance, and the risk probability scores of the model including the Delphi variables as predictors were compared with the model derived from machine learning based on the AUC-ROC curve and the probability plots using descriptive analyses.

All analyses were performed in R version 3.5.0. The "glmnet" package was used for elastic net regularization and leave one out cross-validation was implemented through the "caret" package.

RESULTS

Patient population

For the first part of the study we included 492 SSc patients who completed at least two visits in our cohort (flowchart Figure 1). Seventy-nine percent was female (n= 389) with a mean age of 55 years (SD 14), a median disease duration since first non-Raynaud's symptom of 3.2 years years (IQR 1-10) and the median mRSS was 4 (IQR 0-6) (table 1).

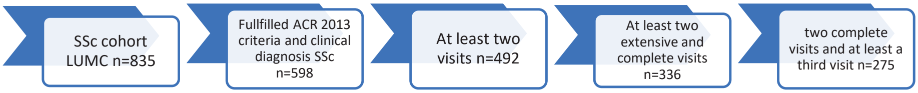
SSc cohort LUMC n=835
Fullfilled ACR 2013 criteria and clinical diagnosis SSc n=598
At least two visits n=492
At least two extensive and complete visits n=336
two complete visits and at least a third visit n=275

Figure 1. Flowcharts of inclusion process. Cut-off timepoint for inclusion: 01-July-2019. * Ninety-two patients had to be excluded due to missing data/incomplete data even though they had three or more visits. Of these, the majority did not show progression based on clinical and laboratory assessment, 6 minute walking distance and pulmonary function testing.

Progressors versus non-progressors in SSc cohort

In n= 492 SSc patients (2109 timepoints, range of follow-up 2-8 years), disease progression during follow-up was observed in 52% (n=2 57) after a median of 4.9 years (IQR 2-7) (Figure 2). Pulmonary (23%) and cardiac progression (29%) occurred most often, death (all-cause) occurred in 12% of the patients (n= 60). We confirm that patients with dcSSc, ILD and ATA at baseline were more likely to experience disease progression somewhere during the disease course (table 1). Forty-eight percent of the SSc patients (n= 235) did not show progression during follow-up (median 3.5 years (IQR 2-6).

Patient selection Machine-Learning-Assisted model

Of the 248 patients that could be included for development of the prediction model, 80% was female (n=220) with a mean age of 53 years (SD 14), a median disease duration since first non-Raynaud sign or symptom of 3.5 years (IQR 1-9) and median mRSS of 4 (IQR 1-6). The baseline characteristics of these patients are shown in table S3 in the supplementary file. Comparison of baseline demographic and clinical characteristics between these patients and the 492 patients with two assessments available showed that the patients included for the prediction model development were more often ATA positive and had higher prevalence of ILD. Other characteristics were not significantly different between the groups (supplementary file table S4).

Baseline characteristics	Total n=492	Non-Progressors N=235	Progressors N=257
Demographics			
Female, n (%)	389 (79)	193 (82)	196 (76)
Age, mean (SD)	55 (14)	55 (15)	55 (13)
Disease duration nonRP, median (IQR)	3.2 (0.9-10.3)	3.5 (0.8-10.5)	3.6 (1.1-9.3)
- lcSSc, median (IQR)	4.1 (1-11)	3.9 (1-11)	2.4 (1-11)
-DcSSc, median (IQR)	3.0 (1-8)	2.7 (1-7)	4.1 (0.5-9)
Organ involvement			
DcSSc, n (%)	118 (24)	**34 (15)**	**84 (33)**
mRSS, median (IQR)	4 (0-6)	**2 (0-5)**	**4 (1-7)**
DU, n (%)	62 (13)	29 (12)	33 (13)
DLCO% of pred, mean (SD)	66 (18)	69 (18)	64 (17)
FVC% of pred, mean (SD)	98 (23)	96 (24)	97 (21)
ILD on HRCT, n (%)	183 (37)	**66 (28)**	**117 (46)**
PAH, n (%)	26 (5)	10 (4)	16 (6)
GAVE, n (%)	9 (2)	4 (2)	5 (2)
Cardiac involvement, n (%)	28 (6)	14 (6)	14 (5)
Myositis, n (%)	8 (2)	6 (3)	2 (1)
Renal crisis, n (%)	14 (3)	6 (3)	8 (3)
Autoantibodies			
Anti-centromere, n (%)	194 (39)	**118 (50)**	**76 (30)**
Anti-topoisomerase, n (%)	116 (24)	**42 (18)**	**74 (29)**
Medication (current use)			
Corticosteroids, n (%)	42 (9)	16 (7)	26 (10)
Methotrexate, n (%)	68 (14)	34 (15)	34 (13)
Mycophenolate mofetil, n (%)	19 (4)	5 (2)	14 (5)
Hydroxychloroquine, n (%)	22 (5)	7 (3)	15 (6)
Cyclophosphamide, n (%)	11 (2)	4 (2)	7 (3)
Azathioprine, n (%)	14 (3)	2 (1)	12 (5)
ASCT, n (%)	4 (1)	2 (1)	2 (1)

Table 1. RP= Raynaud's phenomenon, dcSSc= diffuse cutaneous systemic sclerosis, mRSS= modified Rodnan skin score, DU= digital ulcers, DLCO= single-breath diffusing capacity of the lungs for carbon monoxide, FVC= forced vital capacity, ILD= interstitial lung disease, HRCT= high resolution computed tomography, PAH= pulmonary arterial hypertension, GAVE= gastric antral vascular ectasia, ASCT= autologous stem cell transplantation. Bold indicates significant differences p <0.05 between progressors versus non-progressors.

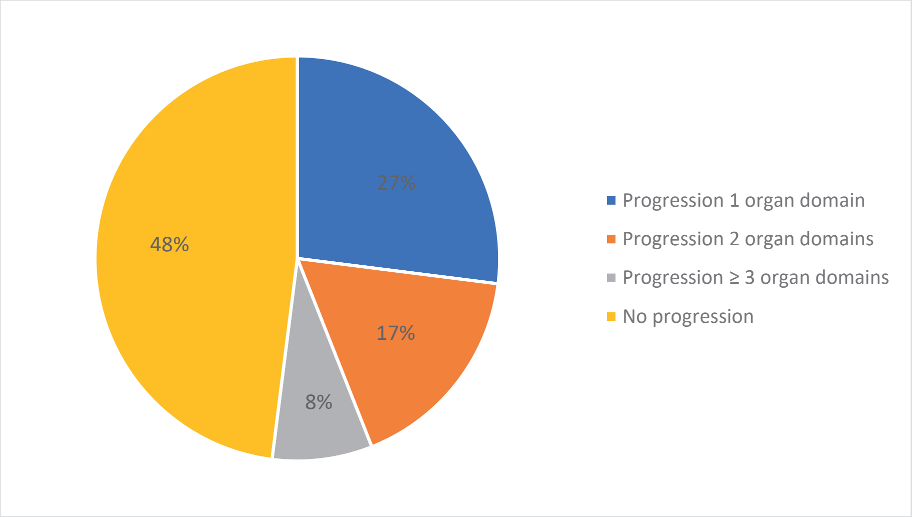

Figure 2. Organ progression in SSc cohort, progression was not always limited to one organ domain. Twenty-five % of the patients showed organ progression on more than one organ domain.

Machine-Learning-Assisted prediction model

After leave-one out cross validation, the final model consisted of 90 variables. The Machine-Learning-Assisted model identified 10 independent variables predictive for disease progression (supplementary table S5). The identified predictors were: previous use of cyclophosphamide (β 0.94) or corticosteroids (β 0.43), previous GI progression (β 0.34), a cardiac event in medical history (β 0.31), PAH (β 0.30), start of immunosuppressives (β 0.21), previous cardiac progression (β 0.08), mRSS (β 0.01), CK (β 0.0006), and DLCO (β -0.004).

The Machine-Learning-Assisted model had an AUC-ROC of 0.77 in the training set (n= 185). The mean (SD) probability score for risk of progression in non-progressors was 0.23 (0.05), and in progressors 0.31 (0.11) (supplementary figure S2). The AUC-ROC of the model in the validation set (n= 63) was 0.66 (figure 3 ROC curve and distribution of probability plot). In this set, the mean (SD) probability score for risk of progression in non-progressors was 0.24 (0.06) and in the progressors 0.29 (0.13). Based on expert opinion, the distribution of probabilities plot and the test characteristics of the validation set (maximize NPV), we identified two cut-offs to identify patients with low (< 0.197), intermediate (0.197-0.223) and high risk (> 0.223) of progression (table 2). A third threshold (> 0.627) corresponding to maximal specificity is also presented (table 2).

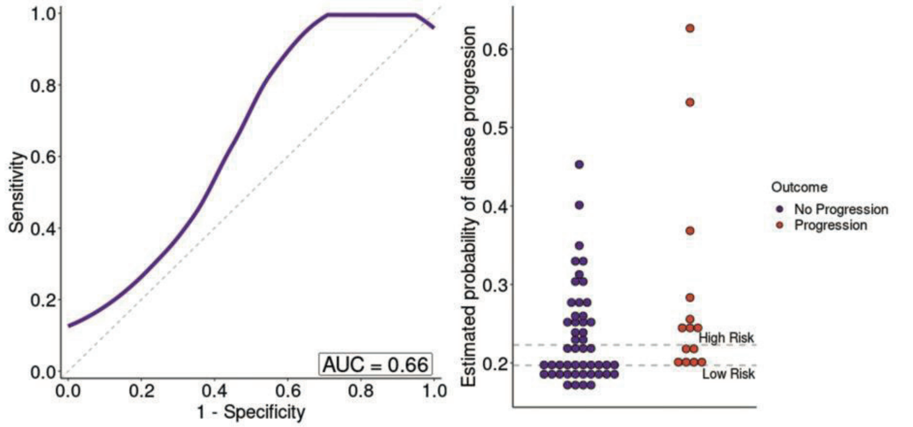

Figure 3. ROC curve and distribution of probability plot of the validation set in progressors and non-progressors.

Threshold	Sensitivity	Specificity	Accuracy	Positive predictive value	Negative predictive value
0.627	0	1	0.78	NaN	0.78
0.223	0.57	0.57	0.57	0.28	0.82
0.197	1	0.37	0.51	0.31	1

Table 2. Test characteristics of data driven prediction model. NaN= not a number.

Disease progression and probability scores in Machine-learning assisted model prediction model

Our primary outcome was progression at the event visit, which occurred in 60 out of 248 patients (24%). Progression was identified in all subdomains: disease subset (n= 3), skin (n= 4), lung (n= 14), cardiac (n= 28), GI (n= 15), renal (n= 2), PAH (n= 4), myositis (n= 6), start of IS therapy (n= 6), and all-cause death (n= 11; detailed overview of cause of death is shown in supplementary file S1). In the validation set (n= 63), 22% (n= 14) showed progression during the event visit, while 78% (n= 49) did not show progression. With guidance of the Machine-learning assisted model prediction model 28 patients were identified as high risk for progression which was correct in 32% (n= 9); 18 patients were identified as low risk which was correct in 100% (due to our strict cut-off), which means that 29% of the patients in the validation set (18 out of 63) had a low risk score and indeed did not show progression. Of the patients with intermediate risk according to our model (n= 17), five showed progression.

Progressors stratified for treatment initiation

To evaluate the clinical relevance of the probability scores we performed an additional analyses in the organ progressors group (validation set) by stratifying them for immunosuppressive treatment initiation after the data collection closure. We hand-searched the electronic patient files of the organ progressors to collect data on IS treatment initiation (started after data collection closure 01.07.2019). Our results showed that patients with organ progression at the most recent visit (primary outcome), for which medication was started, were more likely to score higher on the probability risk score. There was one patient with lung progression who also started treatment with a risk score just above the cut-off for low risk (figure 4). In the non-progressors with a low risk score (n= 18), we identified n= 8 patients who never had any IS treatment during their disease course, n= 9 did use IS medication somewhere during follow-up (HCQ: n= 3 due to arthralgia, polyarthritis, or synovitis, MTX: n= 6 due to limited skin involvement), and 1 patient is still on MMF treatment because of minimal skin (mRSS 2) and lung involvement (minimal interstitial changes on HRCT, without pulmonary function abnormalities).

Delphi score versus Machine-Learning-Assisted model

The most recently published guideline on follow-up in SSc is based on expert opinion by Delphi consensus which advises to yearly measure 55 variables. Based on these 55 independent variables, we built a prediction model [Delphi Model] in order to assess the ability of the Delphi items to identify patients at risk for progression, and compare the performance of the 'Delphi model' with the model derived from machine learning. The AUC-ROC of the Delphi model in the validation set was 0.65 (figure 5 ROC curve and distribution of probabilities plot). The mean (SD) probability score for the risk of progression in non-progressors was 0.14 (0.30), and in the progressors 0.42 (0.46). Of the 54 patients in this validation set, 13 developed progression, of whom 6 patients had a probability risk score below 0.007. The results of the Delphi model, shown in the distribution of probabilities of the validation set, made it difficult to identify cut-offs for low, intermediate and high risk patients on progression. Therefore a cut-off based on an NPV of 1.0, as we used in the Machine-Learning-Assisted model, is not feasible in this model.

The coefficients that were significant in the final model of this prediction set can be found in table S6 supplementary file. The included predictors with the largest coefficients were: previous use of corticosteroids (β 6.66), previous use of iloprost (β 15.1), previous use of bosentan (β -18.0), current use of MMF (β 5.98) or cardiac event in the past (β 5.39).

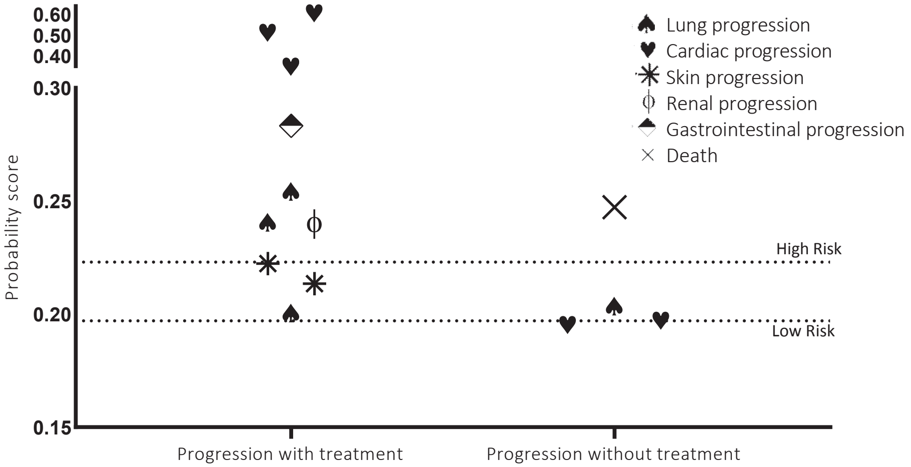

Figure 4. Probability risk scores of the progressors stratified for treatment initiation and organ domain. *Patients with cardiac progression and treatment (n= 3)*: ^{1} trifascicular block with pauses > 3 seconds for which pacemaker implantation, severe tricuspid insufficiency, ^{2} new right bundle branch block, decrease in LVEF < 50%, increase dyspnea,^{3} clinical cardiac involvement; supraventricular arrhythmias 2%, diastolic dysfunction grade 1, elevated troponin T and CK, progressive dyspnea). *Patients with cardiac progression without treatment (n= 2):* ^{1} LVEF < 54%, ^{2} supraventricular arrhythmias > 2% on 24h Holter ECG monitoring . *ILD progression with treatment (n= 3):* ^{1} mild fibrotic changes with a decrease in FVC (73% to 58%) and in DLCO (97% to 76%), ^{2} increase in fibrotic changes, decline FVC (52% to 42%) and decline in DLCO (48% to 28%), ^{3} progressive ILD and decline in FVC and DLCO (n=3). *ILD progression without treatment (n=1)*: ^{1} presence of ILD with bronchiectasis, honeycombing and an increase in reticular opacities, no clinical symptoms, with FVC decline (101% to 90%). *Skin progression with treatment (n= 2):* ^{1} mRSS increase from 10 to 17, ^{2} increase mRSS 10 to 23. *Gastrointestinal progression with treatment (n= 1)*: ^{1} weight loss > 10% in 1 year AND Hb decline. One patient developed renal crisis, one patient died due to lung carcinoma (also had supraventricular extrasystoles > 2 seconds and in increase in fibrotic changes on HRCT).

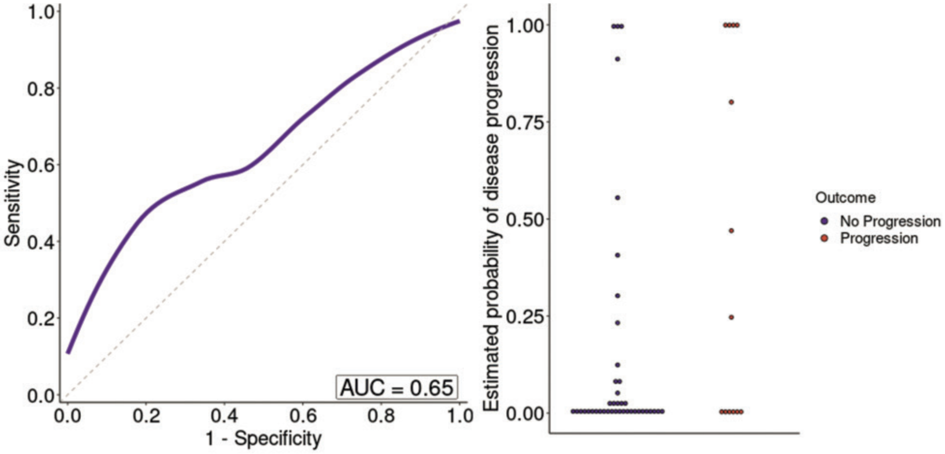

Figure 5. ROC curve and distribution of probabilities plot of the Delphi model stratified for progression.

DISCUSSION

Our newly developed prediction model was able to identify SSc patients with a low risk for disease progression in whom less extensive annual evaluation can be justified. We confirm that SSc is a severe and heterogeneous disease with overall progression occurring in 52% of the patients somewhere during follow-up. In total 235 patients did not experience disease progression during 3.5 years (IQR 2-6) of follow-up.

With the use of machine learning, we developed a prediction model and we managed to include 248 SSc patients with complete data on 90 variables on at least three visits. These patients had a median follow-up of 5.4 years (IQR 3.2-7.5). Although the overall accuracy of the model was moderate, it performed very well in identifying patients with a low risk for disease progression (29% NPV 1.0). For these patients we can adjust annual evaluation using a less extensive diagnostic program.

To identify patients at low risk, we calculated probability scores with the Machine-Learning-Assisted prediction model. The cut-off for low risk patients was very strict, since we did not want to miss any organ progression, with none of the progressors scoring under the low risk cut-off (NPV 1.0). Twenty-nine percent of the SSc patients were identified in the low risk group and extensive follow-up might not be necessary in this patient group. The Machine-Learning-Assisted model could therefore significantly reduce health-care costs without substantial risk to our patients. The assessments that are necessary to identify progression with our model are predominantly: use of IS medication in the past, presence of PAH, mRSS, DLCO and cardiac and GI involvement/progression. Based on this observation showing a diverse group of characteristics that identify risk of progression, we conclude that in all patients with a new diagnosis of SSc complete organ assessment is necessary to guide future follow-up.

To build the Machine-Learning-Assisted model, we used “elastic-net regularization”, a variable selection method that allows to address multicollinearity. It provides a more reproducible prediction than multiple regression, especially when predictors are highly correlated. Elastic net regularization has been shown to robustly maintain predictive accuracy even with a large number of predictors relative to the number of observations. We note that the variables in the final model are predictors, and can therefore not be interpreted as having a causal relationship with progression. Furthermore, since we used a regularization method, variables that play an important causal role could have been dropped from the model when other variables had a similar or stronger association.

Even though the CCISS care pathway is highly standardized and in accordance with international guidelines, we cannot rule out that factors related to the local health care situation have influenced the results. It is therefore important to validate this model in different health care systems. We did not calibrate the probabilities of the Machine-learning assisted model, whereby the probabilities are slightly different from the real risk. This was acceptable since we used a cut-off to identify patients at low risk for progression and not the full range of probabilities.

One of our secondary aims was to compare the Machine-Learning-Assisted model with a model based on the Delphi guideline including the selected tools, to evaluate if assessment of these 55 tools in every patient on a yearly basis might be redundant for a part of the SSc population. The prediction model based on the Delphi variables (including 51 expert opinion variables in the final model) had a similar AUC-ROC as the Machine-learning assisted model model (with only 10 out of 90 variables in the final model based on data driven selection). By using the identified variables for annual follow-up selected by experts to predict disease progression, the discrimination of probability scores between progressors and non-progressors improved but identification of low risk patients was more difficult, and physicians need to collect 51 variables. Forty-six percent of the patients that exhibited progression had a risk-probability close to zero (<0.007) according to the prediction model based on the Delphi model. The Machine-Learning-Assisted model was very well suited to identify patients at low risk as 29% had a probability below 0.197 and all these patients were non-progressors. The comparison between the two models demonstrates that the combination of all Delphi variables cannot directly be used to predict patients at (low) risk for progression. Clearly, the Machine-Learning-Assisted model as constructed in our study is useful to identify patients who are at low risk for disease progression and who therefore may not need intensive follow-up evaluations. Important sidenote, in both models, only patients who underwent complete evaluation for organ involvement and disease progression at least twice were included. Given the severe and heterogeneous nature of SSc, which is underlined by the fact that 52% of patients experienced disease progression during follow-up, in our opinion, annual extensive evaluation is justified in newly diagnosed SSc patients during the first two years. After two evaluations, our current data show that one could consider to apply the probability scores for risk on progression and identify patients in whom follow-up evaluations can be less extensive.

There are some limitations to be acknowledged. First, the clinical variables collected in this study ideally reflect disease activity, disease status and organ damage, to predict disease outcome. However, in SSc, uniform and validated definitions are lacking for some of the organ systems, which should be taken into account as a general limitation.

Secondly, evaluating progression of GAVE and/or PAH is difficult as in clinical practice RHC and endoscopies are not routinely applied as follow-up assessment. We chose to classify PAH patients as non-progressors based on stable pulmonary function testing (PFT), which might have missed some patients. With respect to GAVE we are reasonably convinced that patients with clinically relevant GAVE are correctly identified based on the fact we included hemoglobin in our dataset. Secondly, the follow-up duration might not be sufficient to capture all progressors. Although median follow-up duration is short (5.4 years), 54% of patients had a disease duration of >10 years since first non-RP at the end of the observation period. The follow-up period between progressors and non-progressors was different; however, the proportion of progressors is similar amongst groups stratified for follow-up duration (data not shown). importantly, we had to exclude 244 patients for our final model, as we defined 100% complete data on at least three visits as a prerequisite (for both independent and dependent variables) to predict as accurate as possible (flowchart figure 1). When building the model, we preferred to overestimate progression instead of missing important organ progression. With that in mind, we used these strict inclusion criteria, and, as a consequence, the possibility of selection bias must be considered. The patients included in the model were more often ATA positive and more often had ILD. Therefore, the population used to build the progression model probably had more severe disease, and as such the observed selection does not interferes with the primary aim. Of the 244 patients that had to be excluded for development of the prediction model, 81% was still in follow-up as part of the CCISS cohort, and the majority (89%) did not show progression based on clinical assessment, including PFT and 6 minute walking test (figure S3). The large time frame of the study might also be a limitation as IS treatment might have changed over the years. We evaluated the use of IS therapy for three time frames (2009-2012, 2013-2016, 2017-2019) and found a similar percentage of patients starting IS medication, which makes a large impact on primary outcome unlikely. Another limitation of the study is intrinsic to the heterogeneity of the disease. Included predictors might act as risk and protective factors at the same time, as our primary outcome was aggregated disease progression. We tried to build different models for every organ system, however the occurrence rate of progression was too low to create single reliable models. Finally, we did not have access to a prospective and independent validation sample with all 90 variables available to further test the model's validity. While a completely held-out test sample is statistically equivalent to a prospective sample from the same population, separate validation from a truly prospective sample could further examine the model's generalizability. Previous studies have looked at predictors individually, but the unbiased, data-driven approach, which is a major strength of our work, could contribute to tailor future directions for research and clinical practice. For instance, accurate prediction of patient outcome can be used to inform treatment planning decisions, where modeling the likelihood of disease progression is a critical outcome of interest to health care systems, providers, and stakeholders. Future development of these tools with larger training samples can improve prediction of patient outcome, even to the point where differential predictions of outcome for the personalization of monitoring of SSc might be possible.

A next step is to validate this model in clinical practice. Therefore, we have designed the trial 'From a pragmatic model to a pragmatic study: a non-inferiority randomized trial'. We aim to start inclusion in 2021. The aim of this trial is to evaluate whether annual assessment in patients who underwent extensive evaluation at least twice and are categorized as low risk patients based on our machine learning derived model, can be less extensive without jeopardizing health care utilization, quality of life, disease perceptions and disease course. In addition, an online tool will be developed to calculate risk scores for SSc patients (work in progress).

In the end, achieving equality of assessment worldwide will most likely increase the standard care for SSc. However, until now there were no existing evidence based guidelines for standardized follow-up of patients with SSc. This study showed that disease progression somewhere during follow-up occurs in 52% of the SSc patients, with a high variety between organ systems. Without the use of a prediction model these findings justify the annual complete organ assessments, at least for the first 5 years since first non-RP symptom. While identifying SSc patients at risk for progression remains difficult, our prediction model facilitates the stratification of low, intermediate and high risk patients. In conclusion, SSc patients with a low risk at progression can be identified with the use of the Machine-Learning-Assisted model and allows us to confidently identify a subset of patients who can safely reduce their visit frequency.

REFERENCES

1. Allanore Y, Simms R, Distler O, Trojanowska M, Pope J, Denton CP, et al. Systemic sclerosis. Nat Rev Dis Primers. 2015;1:15002.
2. Becker M, Graf N, Sauter R, Allanore Y, Curram J, Denton CP, et al. Predictors of disease worsening defined by progression of organ damage in diffuse systemic sclerosis: a European Scleroderma Trials and Research (EUSTAR) analysis. Ann Rheum Dis. 2019;78:1242-8.
3. Khanna D, Denton CP. Evidence-based management of rapidly progressing systemic sclerosis. Best Pract Res Clin Rheumatol. 2010;24:387-400.
4. Ledoult E, Launay D, Behal H, Mouthon L, Pugnet G, Lega JC, et al. Early trajectories of skin thickening are associated with severity and mortality in systemic sclerosis. Arthritis Res Ther. 2020;22:30.
5. Fasano S, Riccardi A, Messiniti V, Caramaschi P, Rosato E, Maurer B, et al. Revised European Scleroderma Trials and Research Group Activity Index is the best predictor of short-term severity accrual. Ann Rheum Dis. 2019;78:1681-5.
6. Wu W, Jordan S, Becker MO, Dobrota R, Maurer B, Fretheim H, et al. Prediction of progression of interstitial lung disease in patients with systemic sclerosis: the SPAR model. Ann Rheum Dis. 2018;77:1326-32.
7. Diab N, Hassoun PM. Pulmonary arterial hypertension: screening challenges in systemic sclerosis and future directions. Eur Respir J. 2017;49.
8. Hoffmann-Vold AM, Distler O, Murray B, Kowal-Bielecka O, Khanna D, Allanore Y. Setting the international standard for longitudinal follow-up of patients with systemic sclerosis: a Delphi-based expert consensus on core clinical features. RMD open. 2019;5:e000826.
9. Hao Y, Hudson M, Baron M, Carreira P, Stevens W, Rabusa C, et al. Early Mortality in a Multinational Systemic Sclerosis Inception Cohort. Arthritis Rheumatol. 2017;69:1067-77.
10. Boonstra M, Ninaber MK, Ajmone Marsan N, Huizinga TWJ, Scherer HU, de Vries-Bouwstra JK. Prognostic properties of anti-topoisomerase antibodies in patients identified by the ACR/EULAR 2013 systemic sclerosis criteria. Rheumatology. 2019;58:730-2.
11. Meijs J, Schouffoer AA, Ajmone Marsan N, Kroft LJ, Stijnen T, Ninaber MK, et al. Therapeutic and diagnostic outcomes of a standardised, comprehensive care pathway for patients with systemic sclerosis. RMD open. 2016;2:e000159.
12. LeRoy EC, Medsger TA, Jr. Criteria for the classification of early systemic sclerosis. J Rheumatol. 2001;28:1573-6.
13. Clements P, Lachenbruch P, Siebold J, White B, Weiner S, Martin R, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol. 1995;22:1281-5.
14. Clements PJ, Wong WK, Hurwitz EL, Furst DE, Mayes M, White B, et al. The Disability Index of the Health Assessment Questionnaire is a predictor and correlate of outcome in the high-dose versus low-dose penicillamine in systemic sclerosis trial. Arthritis Rheum. 2001;44:653-61.
15. Ware JE, Kosinski M, Keller S. SF-36 physical and mental health summary scales: a user's manual: Health Assessment Lab.; 1994.
16. Newnham EA, Harwood KE, Page AC. Evaluating the clinical significance of responses by psychiatric inpatients to the mental health subscales of the SF-36. J Affect Disord. 2007;98:91-7.
17. Mouthon L, Rannou F, Bérezné A, Pagnoux C, Arène J-P, Foïs E, et al. Development and validation of a scale for mouth handicap in systemic sclerosis: the Mouth Handicap in Systemic Sclerosis scale. Ann Rheum Dis. 2007;66:1651-5.

18. Schouffoer A, Strijbos E, Schuerwegh M, Mouthon L, Vlieland MV. Translation, cross-cultural adaptation, and validation of the Mouth Handicap in Systemic Sclerosis questionnaire (MHISS) into the Dutch language. Clin Rheumatol. 2013;32:1649.
19. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095-108.
20. Lamers L, Stalmeier P, McDonnell J, Krabbe P, Van Busschbach J. Kwaliteit van leven meten in economische evaluaties: het Nederlands EQ-5D-tarief. Ned Tijdschr Geneeskd. 2005;149:1574-8.
21. Meijs J, Pors D, Vliet VT, Huizinga T, Schouffoer A. Translation, cross-cultural adaptation, and validation of the UCLA Scleroderma Clinical Trial Consortium Gastrointestinal Tract Instrument (SCTC GIT) 2.0 into Dutch. Clin Exp Rheumatol. 2013;32:S-41-8.
22. Khanna D, Hays RD, Park GS, Braun-Moscovici Y, Mayes MD, McNearney TA, et al. Development of a preliminary scleroderma gastrointestinal tract 1.0 quality of life instrument. Arthritis Care Res (Hoboken). 2007;57:1280-6.
23. Lee MAOaP. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for scleroderma criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis rheum. 1980;23:581-90.
24. van den Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis rheum. 2013;65:2737-47.
25. Galie N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, et al. 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). Eur Respir J. 2015;46:903-75.
26. Butler EA, Baron M, Fogo AB, Frech T, Ghossein C, Hachulla E, et al. Generation of a Core Set of Items to Develop Classification Criteria for Scleroderma Renal Crisis Using Consensus Methodology. Arthritis Rheum . 2019;71:964-71.
27. Distler O, Highland KB, Gahlemann M, Azuma A, Fischer A, Mayes MD, et al. Nintedanib for Systemic Sclerosis-Associated Interstitial Lung Disease. NEJM. 2019;380:2518-28.
28. Boonstra M, Meijs J, Dorjée AL, Marsan NA, Schouffoer A, Ninaber MK, et al. Rituximab in early systemic sclerosis. RMD open. 2017;3:e000384.
29. van Laar JM, Farge D, Sont JK, Naraghi K, Marjanovic Z, Larghero J, et al. Autologous hematopoietic stem cell transplantation vs intravenous pulse cyclophosphamide in diffuse cutaneous systemic sclerosis: a randomized clinical trial. Jama. 2014;311:2490-8.
30. Khanna D, Lin CJF, Furst DE, Goldin J, Kim G, Kuwana M, et al. Tocilizumab in systemic sclerosis: a randomised, double-blind, placebo-controlled, phase 3 trial. The Lancet Resp med. 2020;8:963-74.
31. Spiera R, Hummers L, Chung L, Frech TM, Domsic R, Hsu V, et al. Safety and Efficacy of Lenabasum in a Phase II, Randomized, Placebo-Controlled Trial in Adults With Systemic Sclerosis. Arthritis Rheum. 2020;72:1350-60.
32. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R **STAT** SOC **B**. 2005;67:301-20.
33. van Leeuwen N MM, Liem S, Ciaffi J, Ajmone Marsan N, Ninaber M, Huizinga T, de Vries-Bouwstra J. Machine Learning Assisted Prediction of Progression in Systemic Sclerosis Patients: An Approach to Concise, Tailored Model Construction Using Outpatient Clinical Data [abstract]. Arthritis Rheumatol. 2020;72.