Development and initial validation of a composite response index for clinical trials in early diffuse cutaneous systemic sclerosis

Dinesh Khanna, MD, MS[1], Veronica J. Berrocal, PhD[2], Edward H. Giannini, MSc, DrPH[3], James R. Seibold, MD[4], Peter A. Merkel, MD, MS[5], Maureen D. Mayes, MD, MPH[6], Murray Baron, MD[7], Philip J. Clements, MD, MPH[8], Virginia Steen, MD[9], Shervin Assassi, MD, MS[6], Elena Schiopu, MD[1], Kristine Phillips, MD, PhD[1], Robert Simms, MD[10], Yannick Allanore, MD, PhD[11], Christopher P. Denton, MD, PhD[12], Oliver Distler, MD[13], Sindhu R. Johnson, MD, PhD[14], Marco Mattucci-Cerinic, MD, PhD[15], Janet Pope, MD[16], Susanna Proudman, MD[17], Jeffrey Siegel, MD[18], Weng Kee Wong, PhD[8], Athol U. Wells, MD[19], and Daniel E. Furst, MD[8]

[1]University of Michigan Scleroderma Program, Ann Arbor, MI, USA; [2]University of Michigan, School of Public Health, Ann Arbor, MI, USA, [3]Cincinnati Children's Hospital, Cincinnati, OH, USA [4]Scleroderma Research Consultants, Litchfield, CT, USA, [5]University of Pennsylvania, Philadelphia, PA, USA, [6]University of Texas Health Science Center Houston, TX, USA, [7]Jewish General Hospital, McGill University, Montreal, Quebec, Canada, [8]UCLA, Los Angeles, CA, USA [9]Georgetown University, Washington, DC, USA [10]Boston University, Boston, MA, USA [11]Paris Descartes University, Cochin Hospital, Paris, France, [12]Centre for Rheumatology, Royal Free and University College London Medical School,

London, UK, [13]Division of Rheumatology, University Hospital Zurich, Zurich, Switzerland, [1414]Toronto Scleroderma Program, Toronto Western Hospital, University of Toronto, Toronto, Ontario, Canada, [15]University of Florence, Firenze, Italy, [16]Schulich School of Medicine, Western University, London and St Joseph's Health Care, London, ON, [17]Royal Adelaide Hospital, North Terrace, Adelaide, SA, Australia, [18]Genentech/Roche, San Francisco, CA, USA, [19]Royal Brompton Hospital, London, UK

Corresponding author
Dinesh Khanna, MD, MSc
Director, University of Michigan Scleroderma Program
Division of Rheumatology/Dept. of Internal Medicine Suite 7C27
300 North Ingalls Street, SPC 5422
Ann Arbor, MI 48109
Email: khannad@med.umich.edu
Phone: 734.647.8173
Fax: 734.763.5761

CONFLICTS

Dr. Seibold has consultancies relevant to the present work with Bayer, Boehringer-Ingelheim, EMD Serono, FibroGen, Novartis, Sanofi-Aventis, Celgene, DART, InterMune and Sigma Tau.
Dr. Giannini has no conflicts related to this project.
Dr. Johnson is supported by the Canadian Institutes of Health Research Clinician Scientist Award.
Dr Furst has research grants or consultancies for Actelion, Gilead, Novartis, Amgen, Genentech, Roche, BMS
Dr. Denton has/had consultancy relationship and/or has received research funding from Actelion, Roche/Genetech, Pfizer, GlaxoSmithKline, BMS, CSL Behring, Novartis, Sanofi-Aventis, Inventiva, Biogen-Idec.
Dr. Distler has/had consultancy relationship and/or has received research funding in the area of SSc and related conditions from Actelion, Pfizer, Ergonex, BMS, Sanofi-Aventis, United BioSource Corporation, Roche/Genentech, Medac, Biovitrium, Boehringer Ingelheim Pharma, Novartis, 4 D Science, Active Biotec, Bayer-Schering, Sinoxa/ Serodapharm, EpiPharm, Biogen Idec, Inventiva, GSK and Pharmacyclics.

Introduction: Early diffuse systemic sclerosis (dcSSc) is a multisystem disease characterized by rapid changes of skin and internal organs. Our objective was to develop a composite response index in dcSSc for use in randomized controlled trials (RCT).

Methods: We followed well-established consensus and data-driven approaches and subsequently developed paper patient profiles (n =150) using 2 cohorts of dcSSc. Scleroderma experts were invited to rate 20 patient profiles each and assess if each patient had improved or not over a period of 1 year. Using profiles where consensus was reached, we fit logistic regression models where the binary outcome referred to whether patient was improved or not, and the change in the core items from baseline to follow-up were entered as covariates. For each model, sensitivity and specificity were computed. We tested the final index in a previously completed RCT.

Results: Sixteen of 31 core variables were chosen to be included as part of the patient profiles after consensus meeting and review of test characteristics of patient-level data. Forty experts rated profiles and the logistic regression model (including changes in the modified Rodnan skin score, forced vital capacity % predicted, patient and physician global assessments, and HAQ-DI over 1 year) had sensitivity of 0.982 (95%CI 0.98-0.983) and specificity of 0.931 (95%CI 0.929-0.932) and highest face validity. In addition, consensus was achieved for subjects deemed as non-improved based on a significant decline in renal or cardiopulmonary involvement. The index was able to differentiate methotrexate from placebo in a RCT ($p< 0.05$).

Conclusion: We have developed a composite response index applicable to study of dcSSc.

Systemic sclerosis (Scleroderma, SSc) is one of the most fatal rheumatic diseases[1], and is associated with substantial morbidity and many detrimental effects on health-related quality of life[2]. Recent years have seen progress in the development and validation of outcome measures and refinement of trial methodology in SSc [3-6]. This is paralleled by an increased understanding of the pathogenesis of SSc [7] and development of targeted therapies [8]. Modified Rodnan Skin Score, a measure of skin thickness [5], has been used as the primary outcome measure in clinical trials of diffuse cutaneous SSc (dcSSc). However, the complexity and heterogeneity of the disease mandate a composite response measure that will capture different organ involvements and patient-reported outcomes. Validated combined response indices are more likely to be responsive to change than individual measures [9-11], will facilitate drug development and improve assessment of efficacy of therapeutic intervention.

A useful composite index in dcSSc would provide a measure that may improve the ability to measure efficacy facilitate comparison of responses across trials and provide an improved assessment of efficacy of therapeutic agents. Regulatory and funding agencies would then have greater confidence in proposals for interventions, and medical professionals and patients could obtain new evidence on the efficacy of various interventions in the short and long term. This would significantly improve the potential to manage dcSSc. In addition, a composite index would facilitate the standardization, conduct, reporting, and interpretation of clinical trials and could also aid in comparing therapies from different trials.

Therefore, our objective was to develop a composite response index in dcSSc (CRISS) for use in clinical trials.

Patients and Methods:

This iterative process included well-accepted expert consensus [12] and data driven approaches (Figure 1).

Consensus meeting: Members of the Scleroderma Clinical Trials Consortium (SCTC) participated in a Delphi exercise followed by face-to-face nominal group technique (NGT) and this approach has been published elsewhere [4]. Domains and instruments were selected for subsequent data collection.

Data-driven exercise:

A. **Longitudinal observational cohort:** Due to a lack of positive trials in dcSSc and as consequence of the fact that previous trials did not include some of the core set items chosen in the consensus exercise [13], we launched a longitudinal observational cohort of patients with early dcSSc (< 5 years from 1st non-Raynaud's phenomenon sign or symptom) at 4 US Scleroderma Centers with funding from the NIH [14]. The observational cohort recruited 200 patients over a period of 1-year with dcSSc defined as skin thickening proximal, as well as distal, to the elbows or knees, with or without involvement of the face and neck. Exclusion criteria included life expectancy of less than 1 year and non-proficiency of the English language. All 31 core items emerged from the consensus meeting were included to enable an assessment of their psychometric properties (feasibility, reliability, and validity [including sensitivity to change]). Feasibility was defined as completion of the core set measure by > 50% of participants at two time points, redundancy was defined as either a Spearman or Pearson correlation coefficient of at least 0.80, while sensitivity to change was calculated over the 1-year period. Appropriate patient and physician anchors and transition questions were included to assess psychometric properties of core items. For example, modified Likert scale (transition health question) was employed by physicians and patients at 1-year follow up to determine the change in overall condition during the past year on a scale from 1 (“much better”) to 5 (“much worse”). Responses of 1 or 2 were considered an overall improvement, ratings of 4 or 5 were considered a decline in health, while a rating of 3 meant that there was no appreciable change in overall health. Effect size (ES) was calculated using the transition questions as anchors and Cohen’s “rule-of-thumb” for interpreting ES: values of 0.20-0.49 represent a small change, values between 0.50-0.79 a medium change, and ≥0.80 a large change[15]. Core set items that were significant at $p< 0.20$ (for dichotomous measures) or had an effect size ≥ 0.20 in the “Improved” group (with respect to either patient or physician assessments) were further assessed using the modified content validity index matrix [16]. Seven Steering Committee members scored each cell on an ordinal scale (1-4). Each cell was scored according to the following scale: a score of 4 (highest score) was assigned when the cell refereed to a value or an attribute well established in the literature or through systematically obtained information; a score of 3 indicated a value or an attribute somewhat known and accepted, but may need minor alteration or modification; a score of 2 indicated that the rater was unable to assess the attribute without additional information or research; and, finally a score of 1 (lowest score) meant that the attribute should definitely not be used as a core variable. Expert could also assign “not applicable” if they were unfamiliar with an item or different aspects of feasibility, reliability, and validity for the item. Cells scored as 3 or 4 were considered to be supportive of an individual item. Based on results from psychometrics analysis and expert input, a modified nominal group exercise was conducted via webinar by EHG where consensus was defined a priori as ≥75% agreement on each cell of the matrix and overall inclusion/ exclusion of the item as a core set item.

B. **Development and rating of paper patient profiles**. We then developed paper patient profiles using actual data from two cohorts in part due to missing data in the NIH cohort. The Canadian Scleroderma Research Group (CSRG) data was included for patients with dcSSc and disease duration of < 5 years and completeness of data at baseline and follow-up on 15 core variables (except “patient skin interference”) were selected. Since the core variable “patient skin interference last month” was not measured in the Canadian cohort, we imputed its values. Using the NIH cohort, we determined which of the other 15 core variables were useful predictors of patient “skin interference last month” by fitting a linear regression to the NIH cohort data with patient skin interference as outcome and the remaining 15 core variables as covariates. We fitted the linear regression model to the baseline data and the follow-up data separately and imputed the data for “patient skin interference” in the Canadian cohort.

Since patient interviews were not performed as part of the Delphi and NGT, literature was searched to assess the most prevalent/ bothersome issues faced by patients with SSc [17, 18]. Based on this, pain and fatigue (as assessed by SF-36 vitality scale) were included as part of the patient profiles.

Fifty-four international scleroderma experts in clinical care and trial design were subsequently invited to participate in a web-based evaluation of 20 patient profiles each. The experts were randomized based on their location (North America (29) vs. Rest of the World (25)) and years of experience (>10 years [N=38] vs. ≤ 10 years of scleroderma experience [N=16]). For each patient profile, the rater was asked three questions: 1. Do you think the patient has improved, stabilized, or worsened (or unable to tell) over 1-year; 2. If the patient was rated improved or worsened, by how much: considerably, somewhat, or a little; and 3. Please rank the three most important variables that influenced your decision regarding change or stability. Here, the physician raters could choose all the core items from a pull-down menu. Consensus was called if a proportion of at least 75% among those who rated the same patient profile agreed that the patient was improved, stable or worsening. When there was lack of consensus, the Steering Committee members were asked to rate the profiles that were not assigned to them before, followed by a web-based nominal group exercise to discuss each profile in detail. These patient profile ratings were then included with the previous voting and percentage consensus was recalculated. If the proportion of agreement on a patient profile was ≥ 75%, the patient was deemed as having reached consensus. Finally, we sought consensus among SSc experts on which level of change in internal organ involvement would deem a patient as not improved.

To determine whether there was a clear distinction among the 16 core variables in their helpfulness to guide raters in determining whether a patient was improved or not, we conducted a cluster analysis. We used the responses from the raters to the question “Please rank the most important variables that influenced your decision regarding change or stability”, and we clustered the 16 core variables based on the number of times a variable was ranked the most useful, the second most useful and the third most useful. Specifically, we applied the K-means algorithm to the 16x3 data matrix appropriately normalized and rescaled. Since the K-means algorithm requests that the number of clusters in which to group the data be specified a priori, we determined the number of clusters by running the algorithm with K=1,2,..15 clusters. For each number of clusters K, we computed the within-clusters sum of squares, which provides an indication of the degree of similarity within clusters. A lower within-sum of squares is better as it indicates that the clusters are rather homogeneous within themselves but they are different from one another. To determine the number of clusters K, we evaluated for which K there was the largest drop in the within-cluster sum of squares compared to the previous value (corresponding to K-1 clusters), but after which (for K+1 clusters) there was not a considerable change.

Development of response definitions

Using only profiles where consensus was reached, we fitted logistic regression models to the binary outcome representing whether a patient had been rated by experts as being improved (=1) vs. not (=0) and with change in the core items as covariates. For each model, we calculated sensitivity, specificity and area under the curve (AUC). Additionally, using the estimates of the logistic regression coefficients, we derived, for each patient profile, the predicted log-odds (and thus, the predicted probability) that the patient would be rated as improved. We then compared the predicted probability to the raters' consensus opinion on the patient. Accuracy of the predictions could be evaluated in different ways. Using the predicted probabilities in their continuous form, accuracy in the predictions can be quantified via the Brier score [19], a scoring rule that can be interpreted as the equivalent of the Mean Squared Error of the predicted probabilities compared to the binary (yes-improved=1, no-not improved=0) truth.

If y_i represents the raters' consensus opinion on patient i with y_i =1 if the patient has been rated as improved and y_i =0 if the patient has been rated as not improved, and p_i is the predicted probability that the patient is improved, obtained from the logistic regression model, the Brier score is defined as:

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

The Brier score, ranges from 0 to 1, can be used for model selection with the model having the lowest Brier Score having the best predictive performance.

We also tested whether the distribution of the predicted probabilities have a different distribution for the patient profiles who were rated improved by the experts and for those who were rated not improved by performing the non-parametric Mann-Whitney test. Alternatively, the predicted probabilities could be transformed into binary classifications by choosing a threshold and defining “improved” all the patients for which the predicted probability is above the chosen threshold and “not improved” all the patients for which the predicted probability is below the threshold. To identify which threshold (e.g. cut point) to use, we considered different possible cut points from 0.1 to 1.0. For each of the thresholds considered, we derived the corresponding sensitivity and specificity. We made a plot of the sensitivity and specificity as a function of the threshold and determined which threshold had the highest sensitivity and specificity. The data-driven definitions were discussed with the Steering Committee regarding content and face validity.

To evaluate the contribution of each core component to the final CRISS index, we computed the generalized coefficient of determination R^2 for logistic regression [20].

Validation in an independent cohort

The index was tested in a randomized controlled trial of methotrexate vs. placebo in early dcSSc [21]. This trial was chosen as individual patient data were and all final variables were available in this database. We applied the CRISS index to the subjects with complete data and, for each subject, derived the predicted probability that a subject was improved using the predicted probability equation (see Results section). We transformed the continuous predicted probabilities ranging from 0 to 1 into a binary classification, by defining of each subject “improved” or “not improved” depending on whether the predicted probability was above 0.6 or not. We then tested whether the probability of being improved was independent of being on methotrexate (e.g. whether the probability of being improved was the same in the groups of subjects) by performing a chi-square test. We also assessed whether the distributions of the predicted probabilities for the subjects on methotrexate and subjects on placebo were different using the Mann-Whitney test.

Results

Structured Consensus Exercise

Eleven domains and 31 items were identified as the core set meeting OMERACT filters. The 11 domains included: skin, musculoskeletal, cardiac, pulmonary, gastrointestinal, renal, Raynaud’s phenomenon, digital ulcers, health-related quality of life and function, global health, and biomarkers. OMERACT input was obtained during the consensus exercise [3, 22].

NIH observational registry

Two hundred patients with early dcSSc were recruited at baseline and 150 had complete baseline and 1-year data. In these 150 patients, mean (SD) age was 50.4 (11.7), years, 74.7% were female, 78% were Caucasian and 10.7 % were

Latino with mean disease duration (dated from 1st non-Raynaud's sign or symptom) of 2.3 (1.5) years, mean modified Rodnan skin score (mRSS) of 21.4 (10.1) units, mean FVC% predicted of 82.3% (18.5) and mean HAQ-DI of 1.0 (0.8; Table 1).

Measures that lacked feasibility due to low completion rate (<50%) at 1 year included durometer (a measure of skin hardness [23]), right heart catheterization, Borg dyspnea index, 6-minute walk test, and Raynaud's Condition Score [24] that required patient diary records.

Using patient global assessment anchor of improved vs. not, 57% were rated as 'improved' and 43% were rated as "non-improved". Using physician global assessment anchor of improved vs. not, 58% were rated as 'improved' and 42% were rated as "non-improved". Using these anchors, 5 items were found to be not responsive to change or insufficiently common: tender joint count, presence of renal crisis, estimated GFR, body mass index, presence of digital ulcers, and ESR. EHG led a modified nominal group review wherein consensus was achieved on domains/ items that should be used for development of paper patients (Figure 2). It was decided to keep renal crisis and presence/absence of digital ulcers as core set items due to their impact of prognosis in early dcSSc. No redundancy was noted in the 16 core measures at baseline and changed scores as assessed by the correlation coefficients (Appendix Tables 1-2).

Patient Profiles

Patient profiles (examples shown in the Appendix Tables 3-5) were rated by 40 experts (74% completion). In response to the question, “Please rank the most important variables that influenced your decision regarding change or stability”, experts ranked MRSS as most important 44% of the time, followed by FVC% predicted (14.5%), patient global assessment (11.0%), physician global assessment (9.1%), and HAQ-DI (8.0%; Table 2). All other core measures were ranked as most influential in the decision making less than 2% of the time. Examination of the within-cluster sum of squares seem to indicate that K=2 is a good choice. We then clustered the 16x3 matrix using 2 clusters and obtained that MRSS, FVC% predicted, patient global assessment, physician global assessment, and HAQ-DI clustered together and were separated statistically from the remaining core variables (Table 2).

Consensus was achieved in 107 (71.3%) of patient profiles. The Steering Committee reviewed and discussed those profiles on which consensus was not reached and rescored them as improved, worsened or stable (if not done previously by the individual) using nominal group techniques. Following this, final consensus was achieved in 118 (78.7%) profiles that were used for developing the response definitions.

Logistic regression models

Using data from the 118 profiles where consensus was reached, we fit logistic regression models with binary outcome whether a patient had been rated by experts as being improved vs. not and as covariates the change from baseline to follow-up in the 16 core variables. We examined various models, increasing at each step the number of predictors included in the logistic regression model. In 1-variable models (models where only one covariate was included), AUC ranged from 0.47 (for change in presence/absence of new digital ulcers) to 0.92 (for change in MRSS; Appendix Table 6). In a 2-variable model, change in MRSS and change in FVC% predicted yielded the highest AUC (0.96; Appendix Table 7) but was deemed not to have content validity. Different definitions of response and their corresponding AUC, sensitivity and specificity were discussed by the Steering Committee (data available from the corresponding author). The 5-variable model including change in MRSS, FVC% predicted, physician global assessment, patient global assessment, and HAQ-DI was voted as having the greatest face validity (Table 3). This model had a sensitivity of 0.982 (95% CI 0.982, 0.983), specificity of 0.931 (95% CI 0.930, 0.933), and AUC of 0.986. The Brier score was 0.038 (lower score has better predictive performance). As the data was non-normally distributed, non-parametric test indicated that the distributions of the predicted probability of improving were different for the subjects who improved and those who did not (p-value < 0.0001; Figure 1a). Using depiction of sensitivity vs. specificity of improved vs. not improved group, a threshold of 0.6 had the best combination of specificity and sensitivity values (Figure 1b).

Defining a patient who is non-improved irrespective of improvement in other core measures

The Steering Committee considered circumstances where a patient may improve in a particular outcome measure (such as MRSS or FVC% predicted) but have clinically significant worsening or end organ damage to another organ (e.g., development of renal crisis or PAH). There was consensus that such patients should not be considered as improved in a clinical trial. The Steering Committee voted on new onset of renal crisis, new onset or worsening lung fibrosis, new onset PAH, or new onset of left ventricular failure (Table 4). The international experts subsequently endorsed these definitions as well.

Application in trial

CRISS is a 2-step process. In step 1, subjects who develop new onset of renal crisis, new onset or worsening lung fibrosis, new onset PAH, or new onset of left ventricular failure (Table 4) during the trial are considered as non-improved and assigned a probability of 0.0. For the remaining subjects with complete data, Step 2 involves assigning the predicted probability of improving for each subject using the following equation (equation to derive predicted probabilities from a logistic regression model):

$$\frac{exp\left[-5.54 - 0.81 * \Delta_{MRSS} + 0.21 * \Delta_{FVC\%} - 0.40 * \Delta_{Pt-glob} - 0.44 * \Delta_{MD-glob} - 3.41 * \Delta_{HAQ-DI}\right]}{1 + exp\left[-5.54 - 0.81 * \Delta_{MRSS} + 0.21 * \Delta_{FVC\%} - 0.40 * \Delta_{Pt-glob} - 0.44 * \Delta_{MD-glob} - 3.41 * \Delta_{HAQ-DI}\right]}$$

where Δ_{MRSS} indicates the change in MRSS from baseline to follow-up, Δ_{FVC} denotes the change in FVC% predicted from baseline to follow-up, $\Delta_{Pt\text{-}glob}$ indicates the change in patient global assessment, $\Delta_{MD\text{-}glob}$ denotes the change in physician global assessment, and $\Delta_{HAQ\text{-}DI}$ is the change in HAQ-DI. All changes are absolute change ($Time_2$ –$Time_{baseline}$). Subjects for which the predicted probability is greater or equal to 0.60 are considered improved, while subjects for which the predicted probability is below 0.60 are considered non-improved. The 2 groups (drug vs. placebo) can then be compared in a 2x2 table and using appropriate significance tests. The predicted probabilities obtained using the CRISS can also be assessed as a continuous variable and the distributions of the probability of improving for patients on drug vs. placebo can be compared using non-parametric tests.

Contribution of 5 core components to the CRISS

We computed the R^2 for the logistic regression models that had each of the 5 core components of the CRISS as the single predictors. MRSS explained 66.3% of the variation, FVC% predicted explained 36.1% of the variation, physician global assessment explained 24.5% of variation, patient global assessment explained 23.7% variation, and 28.5% was explained by HAQ-DI.

To assess how changes in the core variables are related to the predicted probabilities of improving on each patient profile, Appendix Figure 1(a)-(e) presents scatterplot of the change in MRSS, change in FVC% predicted, change in the patient global, change in physician global and change in HAQ-DI versus the predicted probabilities for the 118 patient profiles. A change in MRSS, FVC% predicted and HAQ-DI are strong indicators of whether a patient is likely to be improved or not. In each scenario, a decrease of MRSS or HAQ-DI from baseline to follow-up and an increase in FVC% predicted corresponds to very high probabilities of improving. For patient global and physician global, the association between probability of improving and change in these two core components is less evident.

Validation in a clinical trial

We used the individual patient data from the methotrexate vs. placebo trial to assess our definition of response. Data for change in MRSS, FVC% predicted, patient global assessment, physician global assessment, and HAQ-DI was available for 35 of 71 patients at 1 year. Using the CRISS, we derived the predicted probability of improving for each of the 35 patients and classified them into improved and not improved using a probability cutoff of 0.6. With this criterion, 11 of 19 subjects who were on methotrexate were rated as improved whereas 3 of 16 subjects in placebo were rated as improved (p=0.04; Appendix Figure 2). When the data was assessed as continuous measure, the distribution of the predicted probability for improvement were statistically different (p= 0.02).

Discussion

We have developed a composite index for trials in early dcSSc using well-established consensus and data-driven approaches. The index includes measures that assess change in two common and prominent manifestations of early dcSSc (skin and ILD), functional disability (as assessed by HAQ-DI), and patient and physician global assessments. In addition, the index captures clinically meaningful decline in internal organ involvement that deems the patient has not improved during the clinical trial. We subsequently validated this in a clinical trial and showed that CRISS index can differentiate methotrexate from placebo in early dcSSc.

Traditionally, trials in early dcSSc have focused on skin or lung involvement [25, 26]. MRSS has been used as the primary outcome measure for the trials of skin fibrosis [5]. MRSS meets the OMERACT criteria as a fully validated measure of outcome [27], but is also a surrogate of internal organ involvement and mortality in early dcSSc [28, 29]. However, the trials to date have largely been negative and MRSS has been questioned as primary outcome measure where post-hoc analysis of negative trials has shown stability/ improvement in MRSS over time [30, 31]. The CRISS index is the first step is capturing the multisystem involvement of dcSSc and includes patient perspective and impact on functional disability.

The CRISS index is calculated as a 2-step process. The first step evaluates clinically significant decline in renal or cardiopulmonary involvement and if present, the patient is adjudicated as non-improved. The second step assesses remaining patients and calculates the predicted probability of improvement. Here, the Steering Committee discussed different response definitions and decided to go with a data-driven definition rather than using a percent improvement definition, that includes expert consensus, as suggested by the ACR subcommittee [32]. In addition, data-driven definitions (e.g., disease activity index for rheumatoid arthritis [33]) have been successfully used for regulatory approval in other rheumatic diseases.

The goal of CRISS is assess new pharmacologic agents that target the underlying pathogenesis and have impact on overall disease activity/ severity. Our hope is that CRISS use in patients with dcSSc will greatly facilitate the interpretation of results from clinical trials and form the basis for drug approval. Rather than using numerous outcomes that vary from trial to trial, the core set of measures used in CRISS will produce a single efficacy measure. This process will lessen the ambiguity associated with the presentation of multiple test statistics, some of which may be significant and others not and facilitate meta-analysis. It will likely also allow a decrease in the patients necessary for appropriately powered clinical trials, as it has been true for the use of combined indices in rheumatoid arthritis. It should also be noted that the use of CRISS does not preclude the addition of other measures in a trial; it simply provides one standardized outcome that can be easily compared and understood. If the goal of the trial is to focus on a particular organ (e.g., use of vasodilators for underlying digital ulcers), then CRISS index is can be used as a secondary/ exploratory measure.

The initial panel of domains (11) and items (31) offered a comprehensive view of the marked heterogeneity of SSc and at first was seen to potentially mimic the comprehensive structure of BILAG and SLEDAI [34]. However, many items were discarded based on lack of sensitivity to change in our actual data gathering exercise and others were demonstrated to lack feasibility. It is the data-driven basis for our approach to development of the CRISS, which supports our relatively simple and accessible panel of items.

Our research has many strengths. It is first concerted effort by the scleroderma community to address lack of a robust composite index for a multisystem disease. We used well-accepted expert consensus and data-driven methodologies and successful derived and validated the index in early dcSSc. Second, the index captures organ involvement in early dcSSc, patient assessment of their overall disease, functional disability, and physician global assessment.

Our study is not without limitations. First, the CRISS index is developed for early dcSSc and may not be valid for late dcSSc or lcSSc. A similar exercise in late lcSSc might focus on vascular complications such as digital ulcers or pulmonary arterial hypertension but would not include MRSS. The majority of past and ongoing trials are focused on early dcSSc due to dynamic changes in skin and internal organ involvement that may responsive to pharmacologic intervention. Second, we did not get patient input during development of the index. We acknowledged this limitation and searched the literature [17, 18] that led to inclusion of fatigue and pain during the development of patient profile but neither measure survived the nominal group exercises. Nonetheless; two of the constituent measures of the CRISS index include patient global assessment and patient-reported functional assessment.

In conclusion, we have developed and provide initial validation of a novel composite index for clinical trials in early dcSSc.

Acknowledgement

We thank following colleagues for participating in rating of patient profiles

Avouac, Jerome; Carreira, Patricia ; Chung, Lorinda; Csuka, Mary Ellen; Czirjak, Laszlo; Frech, Tracy; Herrick, Ariane; Hinchcliff, Monique; Hsu, Vivian; Inanc, Murat; Jimenez, Sergio; Kahaleh, Bashar; Kowal-Bielecka, Otylia; Medsger Jr., Thomas A; Müller-Ladner, Ulf; Nikpour, Mandana; Shah, Ami; Stevens, Wendy; Valentini, Gabriele ; van Laar, Jacob M; Varga, John; Vonk, Madelon; Walker, Ulrich A

References

1. Ioannidis JP, Vlachoyiannopoulos PG, Haidich AB, Medsger TA, Jr., Lucas M, Michet CJ, Kuwana M, Yasuoka H, van den HF, Te BL *et al*: **Mortality in systemic sclerosis: an international meta-analysis of individual patient data**. *AmJ Med* 2005, **118**(1):2-10.
2. Khanna D, Kowal-Bielecka O, Khanna PP, Lapinska A, Asch SM, Wenger N, Brown KK, Clements PJ, Getzug T, Mayes MD *et al*: **Quality indicator set for systemic sclerosis**. *ClinExpRheumatol* 2011, **29**(2 Suppl 65):S33-S39.
3. Khanna D, Distler O, Avouac J, Behrens F, Clements PJ, Denton C, Foeldvari I, Giannini E, Huscher D, Kowal-Bielecka O *et al*: **Measures of response in clinical trials of systemic sclerosis: the combined response index for systemic sclerosis (CRISS) and Outcome Measures in Pulmonary Arterial Hypertension related to Systemic Sclerosis (EPOSS)**. *J Rheumatol* 2009, **36**(10):2356-2361.
4. Khanna D, Lovell DJ, Giannini E, Clements PJ, Merkel PA, Seibold JR, Matucci-Cerinic M, Denton CP, Mayes MD, Steen VD *et al*: **Development of a provisional core set of response measures for clinical trials of systemic sclerosis**. *AnnRheum Dis* 2008, **67**(5):703-709.
5. Khanna D, Merkel PA: **Outcome measures in systemic sclerosis: an update on instruments and current research**. *CurrRheumatol Rep* 2007, **9**(2):151-157.
6. Chung L, Denton CP, Distler O, Furst DE, Khanna D, Merkel PA: **Clinical trial design in scleroderma: where are we and where do we go next?** *Clin ExpRheumatol* 2012, **30**(2 Suppl 71):S97-102.
7. Abraham DJ, Varga J: **Scleroderma: from cell and molecular mechanisms to disease models**. *Trends Immunol* 2005, **26**(11):587-595.
8. Nagaraja V, Denton CP, Khanna D: **Old medications and new targeted therapies in systemic sclerosis**. *Rheumatology (Oxford)* 2014.
9. van der Heijde DM, t Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA, van Rijswijk MH, van de Putte LB: **Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score**. *Ann Rheum Dis* 1990, **49**(11):916-920.
10. Paulus HE, Egger MJ, Ward JR, Williams HJ: **Analysis of improvement in individual rheumatoid arthritis patients treated with disease-modifying antirheumatic drugs, based on the findings in patients treated with placebo. The Cooperative Systematic Studies of Rheumatic Diseases Group**. *Arthritis Rheum* 1990, **33**(4):477-484.
11. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, Katz LM, Lightfoot R, Jr., Paulus H, Strand V: **American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis**. *Arthritis Rheum* 1995, **38**(6):727-735.
12. Nair R, Aggarwal R, Khanna D: **Methods of Formal Consensus in Classification/Diagnostic Criteria and Guideline Development**. *SeminArthritis Rheum* 2011.

13. Merkel PA, Silliman NP, Clements P, Denton CP, Furst DE, Mayes M, Pope J, Polisson RP, Streisand JB, Seibold JR: **Patterns and Predictors of Change in Outcome Measures in Clinical Trials in Scleroderma** *Arthritis Rheum* 2005, **52:S282-283**

14. Wiese AB, Berrocal VJ, Furst DE, Seibold JR, Merkel PA, Mayes MD, Khanna D: **Correlates and responsiveness to change of measures of skin and musculoskeletal disease in early diffuse systemic sclerosis**. *Arthritis Care Res (Hoboken)* 2014.

15. Cohen J: **Statistical power analysis for the behavioral sciences**, 2nd edn. Hillsdale, NJ: Erlbaum; 1988.

16. Davies EH, Surtees R, DeVile C, Schoon I, Vellodi A: **A severity scoring tool to assess the neurological features of neuronopathic Gaucher disease**. *Journal of inherited metabolic disease* 2007, **30**(5):768-782.

17. Bassel M, Hudson M, Taillefer SS, Schieir O, Baron M, Thombs BD: **Frequency and impact of symptoms experienced by patients with systemic sclerosis: results from a Canadian National Survey**. *Rheumatology (Oxford)* 2011, **50**(4):762-767.

18. Suarez-Almazor ME, Kallen MA, Roundtree AK, Mayes M: **Disease and Symptom Burden in Systemic Sclerosis: A Patient Perspective**. *JRheumatol* 2007.

19. Gneiting T, Raftery A: **Strictly proper scoring rules**. *Journal of the American Statistical Association* 2007, **102**.

20. Nagelkerke NGD: **A note on a general definition of the coefficient of Determination**. *Biometrika* 1991.

21. Pope JE, Bellamy N, Seibold JR, Baron M, Ellman M, Carette S, Smith CD, Chalmers IM, Hong P, O'Hanlon D *et al*: **A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma**. *Arthritis Rheum* 2001, **44**(6):1351-1358.

22. Furst D, Khanna D, Matucci-Cerinic M, Clements P, Steen V, Pope J, Merkel P, Foeldvari I, Seibold J, Pittrow D *et al*: **Systemic sclerosis - continuing progress in developing clinical measures of response**. *J Rheumatol* 2007, **34**(5):1194-1200.

23. Merkel PA, Silliman NP, Denton CP, Furst DE, Khanna D, Emery P, Hsu VM, Streisand JB, Polisson RP, Akesson A *et al*: **Validity, reliability, and feasibility of durometer measurements of scleroderma skin disease in a multicenter treatment trial**. *Arthritis Rheum* 2008, **59**(5):699-705.

24. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, Korn JH, Simms RW, Csuka ME, Medsger TA, Jr. *et al*: **Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon**. *Arthritis Rheum* 2002, **46**(9):2410-2420.

25. Khanna D, Clements PJ, Furst DE, Korn JH, Ellman M, Rothfield N, Wigley FM, Moreland LW, Silver R, Kim YH *et al*: **Recombinant human relaxin in the treatment of systemic sclerosis with diffuse cutaneous involvement: A randomized, double-blind, placebo-controlled trial**. *Arthritis Rheum* 2009, **60**(4):1102-1111.

26. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, Arriola E, Silver R, Strange C, Bolster M *et al*: **Cyclophosphamide versus placebo in scleroderma lung disease**. *N Engl J Med* 2006, **354**(25):2655-2666.
27. Merkel PA, Clements PJ, Reveille JD, Suarez-Almazor ME, Valentini G, Furst DE: **Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6**. *J Rheumatol* 2003, **30**(7):1630-1647.
28. Clements PJ, Hurwitz EL, Wong WK, Seibold JR, Mayes M, White B, Wigley F, Weisman M, Barr W, Moreland L *et al*: **Skin thickness score as a predictor and correlate of outcome in systemic sclerosis: high-dose versus low-dose penicillamine trial**. *Arthritis Rheum* 2000, **43**(11):2445-2454.
29. Steen VD, Medsger TA, Jr.: **Severe organ involvement in systemic sclerosis with diffuse scleroderma**. *Arthritis Rheum* 2000, **43**(11):2437-2444.
30. Merkel PA, Silliman NP, Clements PJ, Denton CP, Furst DE, Mayes MD, Pope JE, Polisson RP, Streisand JB, Seibold JR *et al*: **Patterns and predictors of change in outcome measures in clinical trials in scleroderma: an individual patient meta-analysis of 629 subjects with diffuse cutaneous systemic sclerosis**. *Arthritis Rheum* 2012, **64**(10):3420-3429.
31. Amjadi S, Maranian P, Furst DE, Clements PJ, Wong WK, Postlethwaite AE, Khanna PP, Khanna D: **Course of the modified Rodnan skin thickness score in systemic sclerosis clinical trials: Analysis of three large multicenter, double-blind, randomized controlled trials**. *Arthritis Rheum* 2009, **60**(8):2490-2498.
32. Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, Paulus H, Wallace C: **Development of classification and response criteria for rheumatic diseases**. *Arthritis Rheum* 2006, **55**(3):348-352.
33. van der Heijde DM, van't Hof MA, van Riel PL, van Leeuwen MA, van Rijswijk MH, van de Putte LB: **Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis**. *Ann Rheum Dis* 1992, **51**(2):177-181.
34. Yee CS, Isenberg DA, Prabu A, Sokoll K, Teh LS, Rahman A, Bruce IN, Griffiths B, Akil M, McHugh N *et al*: **BILAG-2004 index captures systemic lupus erythematosus disease activity better than SLEDAI-2000**. *Ann Rheum Dis* 2008, **67**(6):873-876.

Table 1: Baseline demographics of patients who participated in the NIH observational study with baseline and 1 year data

	Baseline N	
Age, mean (SD)	150	50.4 (11.7)
Female, N (%)		112 (75%)
Race, N (%) Caucasian African American Asian	150	117 (78%) 13 (9%) 11 (7%)
Ethnicity, N (%) Hispanic Non-Hispanic	150	 16 (11%) 134 (89%)
Disease duration (yrs), mean (SD)	144	1.59 (1.34)
Years since first Raynaud symptom, mean (SD)	128	2.87 (2.49)
Years since first non-Raynaud symptom, mean (SD)	129	2.32 (1.5)
BMI, mean (SD)	96	26.02 (7.1)
MRSS, mean (SD)	150	21.4 (10.1)
Durometer, mean (SD)	113	272.4 (64.51)
Forced vital capacity % predicted, mean (SD)	140	82.32 (18.50)
Total lung capacity% predicted, mean (SD)	109	87.83 (20.38)
Diffusion capacity of carbon monoxide % predicted, mean (SD)	140	65.05 (20.86)
HRCT consistent with ILD, N (%)	99	79 (80%)
6-minute walking distance, mean (SD)	50	421.6 (139.25)
Borg dyspnea (0-10 scale), mean (SD)	46	1.92 (1.51)
Tendon Friction rubs, N (%)	140	40 (29)
Small joint contractures, N (%)	133	78 (29)
Large joint contractures, N (%)	133	39 (59)
Digital tip ulcers, N (%)	150	15 (10%)
HAQ-DI, mean (SD)	150	1.02 (0.79)
Digital ulcers VAS (0-150), mean (SD)	134	20.93 (40.91)
Raynaud's VAS (0-150), mean (SD)	135	32.70 (40.81)
Breathing VAS (0-150), mean (SD)	138	23.07 (36.72)
GI VAS (0-150), mean (SD)	136	22.60 (34.44)
Disease severity VAS (0-150), mean (SD)	138	56.40 (42.88)
Pain VAS (0-10), mean (SD)	140	4.0 (2.8)
SF-36 PCS, mean (SD)	138	37.56 (12.95)
SF-36 MCS, mean (SD)	138	44.23 (6.00)
Physician global assessment (0-10), mean	143	4.44 (2.19)

(SD)		
Patient global assessment (0-10), mean (SD)	140	4.07 (4.0)
Antinuclear antibody, N (%)	116	94 (81.0%)
Anti-SCl-70 antibody, N (%)	115	34 (30%)
Serum creatinine phosphokinase, mean (SD)	127	143.90 (184.5)
Serum Platelets, mean (SD)	143	315.2 (102.5)
Serum brain natriuretic peptide, mean (SD)	105	161.3 (824.0)
Serum erythrocyte sedimentation rate, mean (SD)	121	23.38 (22.64)
Serum C-reactive protein, mean (SD)	116	2.08 (4.94)

VAS=visual analog scale; PCS=Physical component scale; MCS=Mental component scale

Table 2. Ranking of the core variables by scleroderma experts and cluster analysis

Variable	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)	Cluster
MRSS	374 (44.1%)	131 (15.5%)	75 (8.9%)	1
FVC% predicted	123 (14.5%)	148 (17.5%)	72 (8.5%)	1
Physician global assessment	77 (9.1%)	116 (13.7%)	88 (10.4%)	1
Patient global assessment	93 (11%)	69 (8.2%)	115 (13.6%)	1
HAQ-DI	68 (8%)	112 (13.2%)	99 (11.7%)	1
Vitality SF-36	12 (1.4%)	37 (4.4%)	101 (11.9%)	2
GI VAS	25 (2.9%)	44 (5.2%)	43 (5.1%)	2
Pain	11 (1.3%)	38 (4.5%)	82 (9.7%)	2
Tendon friction rubs	11 (1.3%)	33 (3.9%)	23 (2.7%)	2
Breathing VAS	13 (1.5%)	25 (3%)	32 (3.8%)	2
Digital ulcers VAS	7 (0.8%)	38 (4.5%)	17 (2%)	2
Raynaud’s VAS	11 (1.3%)	18 (2.1%)	43 (5.1%)	2
Patient interference skin last month	2 (0.2%)	21 (2.5%)	22 (2.6%)	2
No. digital ulcers	9 (1.1%)	11 (1.3%)	17 (2%)	2
Renal crisis	11 (1.3%)	3 (0.4%)	2 (0.2%)	2
Body mass index	1 (0.1%)	3 (0.4%)	15 (1.8%)	2

Table 3. Final CRISS model consisting of 5-variables with highest face validity

Variables (calculated as change from baseline to 1 year)	Area under the curve (AUC)	Sensitivity (95% CI)	Specificity (95% CI)	Unadjusted Beta coefficients
MRSS	0.9935	0.9464	0.9310	-1.06
FVC predicted				0.30
HAQ-DI				-0.67
Patient global assessment				-0.90
Physician global assessment				-5.61

Table 4 Expert consensus on definition of a patient who is not-improved during a trial

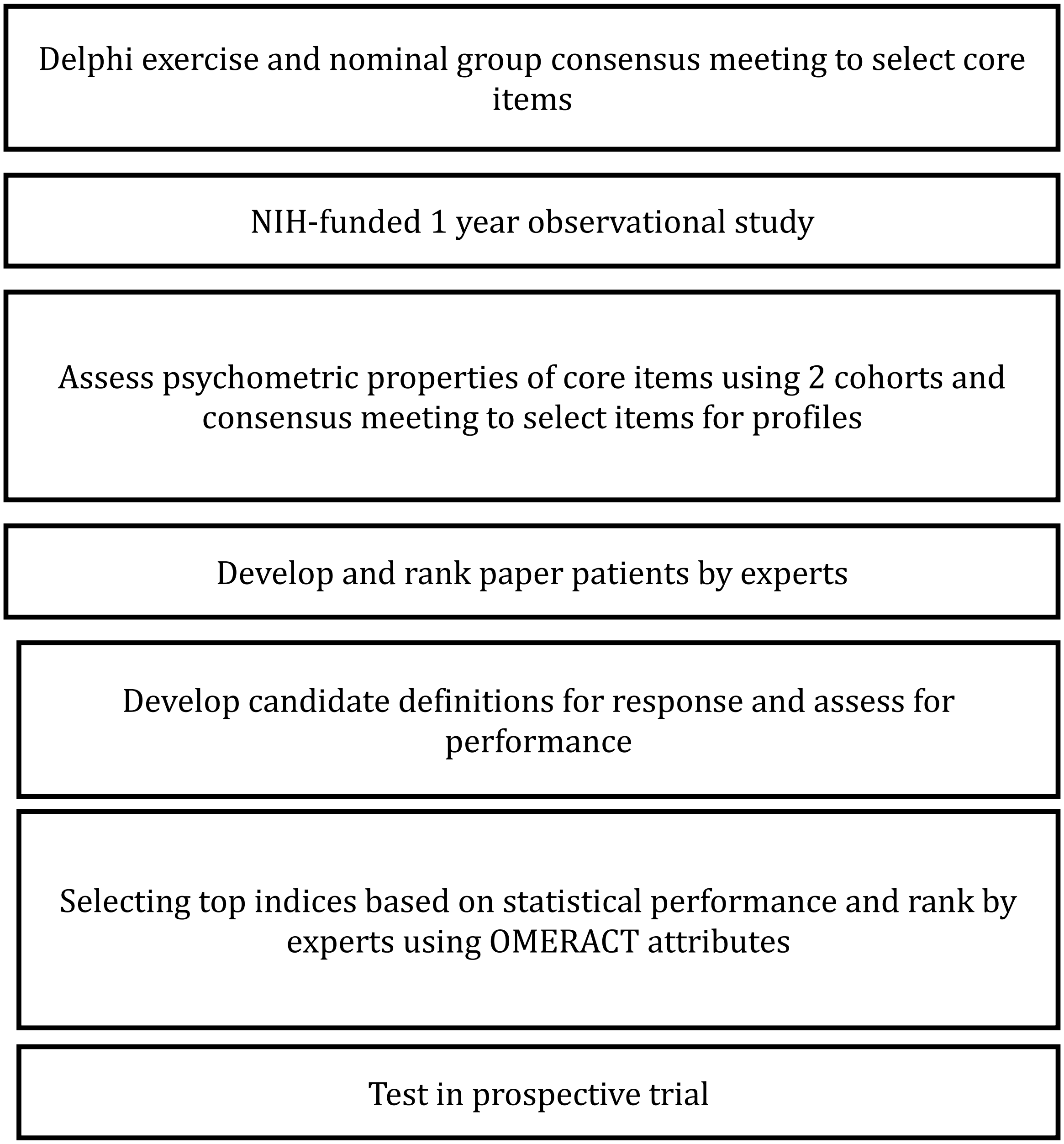
Patient is considered not improved* if he/she develops

- **New scleroderma renal crisis**
- **Decline in FVC% predicted≥ 15% (relative), confirmed by another FVC% within a month, HRCT to confirm ILD (if previous HRCT did not show ILD) and FVC% predicted below 80% predicted****
- **New onset of left ventricular failure (defined as ejection fraction ≤45%) or new onset of pulmonary arterial hypertension requiring treatment****

***Irrespective of improvement in other core items**

**** Attributable to SSc**

Figure 1: Expert consensus and data-driven approaches using to develop CRISS index.



Delphi exercise and nominal group consensus meeting to select core items

NIH-funded 1 year observational study

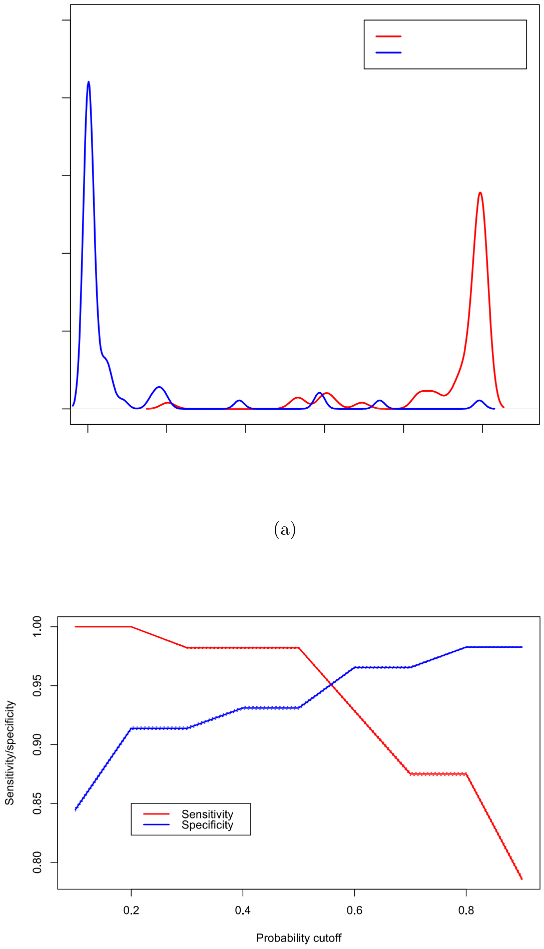
Assess psychometric properties of core items using 2 cohorts and consensus meeting to select items for profiles

Develop and rank paper patients by experts

Develop candidate definitions for response and assess for performance

Selecting top indices based on statistical performance and rank by experts using OMERACT attributes

Test in prospective trial



(a)

(b)

Figure 2. (a) Distribution of the predicted probability of improving for patients rated improved by the experts (red curve) and patients rated not improved by experts (blue curve). (b) Sensitivity (red line) and specificity (blue line) of the predicted classification of patients into “improved” and “not improved” as a function of the predicted probability cutoff. The cutoff considered are 0.1, 0.2, 0.3, … 0.9 and the predicted classification are derived as follow: if the predicted probability for a subject is greater than the probability cutoff, the subject is rated as “improved”, otherwise it is not.

Appendix Table 1. Correlation between the continuous variables among the 16 core variables at baseline.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
V1	1.0	-0.26	0.43	0.60	0.33	0.49	0.31	0.04	0.16	0.09	0.09	0.04	0.03	0.17
V2		1.0	-0.22	-0.33	-0.23	-0.20	-0.18	0.02	-0.03	-0.17	-0.003	-0.11	-0.27	-0.16
V3			1.0	0.46	0.57	0.66	0.56	0.23	0.26	0.17	0.02	-0.06	0.28	0.25
V4				1.0	0.45	0.54	0.33	0.17	0.18	0.11	0.04	0.08	0.13	0.10
V5					1.0	0.55	0.57	0.35	0.35	0.19	-0.02	0.01	0.41	0.30
V6						1.0	0.60	0.19	0.44	0.26	0.11	0.06	0.30	0.22
V7							1.0	0.17	0.47	0.41	0.11	0.09	0.34	0.33
V8								1.0	0.15	0.06	-0.05	0.06	0.26	0.07
V9									1.0	0.35	0.20	0.15	0.39	0.45
V10										1.0	0.16	0.11	0.20	0.23
V11											1.0	-0.04	-0.02	0.02
V12												1.0	0.19	0.07
V13													1.0	0.36
V14														1.0

V1=MRSS, V2=FVC predicted, V3=HAQ-DI, V4=MD global, V5=Patient global, V6=Patient skin interference, V7=Pain, V8=Vitality, V9=Raynaud VAS, V10=Digital Ulcers VAS, V11=Number of digital ulcers, V12=BMI, V13=Breathing VAS, V14=GI VAS
*renal crisis and tendon friction rubs not included

Appendix Table 2. Correlation between the change scores in the 16 core continuous variables.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
V1	1.0	-0.30	0.22	0.26	0.16	0.32	0.21	0.12	0.17	0.17	-0.10	0.07	0.08	0.17
V2		1.0	-0.39	-0.31	-0.27	-0.29	-0.33	0.03	-0.06	-0.17	0.10	0.002	-0.30	-0.10
V3			1.0	0.17	0.27	0.31	0.23	-0.005	0.08	-0.05	-0.009	-0.18	0.30	0.05
V4				1.0	0.25	0.46	0.19	-0.09	0.18	0.03	-0.08	0.04	0.33	0.26
V5					1.0	0.13	0.25	-0.007	0.002	0.05	-0.14	-0.10	0.16	0.25
V6						1.0	0.28	-0.08	0.15	-0.07	-0.02	0.22	0.30	0.02
V7							1.0	0.07	0.27	0.10	0.22	0.11	0.33	0.23
V8								1.0	0.001	-0.12	-0.03	0.01	-0.12	-0.14
V9									1.0	0.20	0.35	0.20	0.23	0.47
V10										1.0	-0.13	0.11	0.05	0.36
V11											1.0	0.008	0.06	0.05
V12												1.0	0.16	-0.07
V13													1.0	0.28
V14														1.0

V1=MRSS, V2=FVC predicted, V3=HAQ-DI, V4=MD global, V5=Patient global, V6=Patient skin interference, V7=Pain, V8=Vitality, V9=Raynaud VAS, V10=Finger Ulcers VAS, V11=Number of digital ulcers, V12=BMI, V13=Breathing VAS, V14=GI VAS
*renal crisis and tendon friction rubs not included

Appendix Table 3. Example of a patient rated as “improved” by the experts. Predicted probability of improving is 0.99 according to CRISS index.

	Baseline	Follow-up	Absolute change
Age	51.6 yrs		
Disease duration (months)	12.98		
Global assessments			
Patient global assessment (0-10)	3	1	-2
MD global assessment (0-10)	3	3	0
Muscoloskeletal			
HAQ-DI (0-3)	0.625	0	-0.625
Tendon friction rubs	No	No	No change
Skin			
MRSS (0-51)	13	3	-10
Patient interference skin last month	2	0	-2
Lung			
FVC% predicted	62	75	13
Breathing VAS (0-10)	2	0	-2
Renal			
Renal crisis	No	No	No change
Gastrointestinal			
GI VAS (0-10)	3	3	0
Body Mass Index (BMI)	25.40	26.58	1.18
Raynaud’s			
Raynaud’s VAS (0-10)	2	1	-1
Digital ulcers			
Digital ulcers VAS (0-10)	0	0	0
Number of digital ulcers	0	0	0
HRQOL			
Pain VAS (0-10)	3	1	-2
Fatigue (SF-36 Vitality scale) (0-100)	42.31	35.12	-7.19

Appendix Table 4. Example of a patient rated improved by the experts. Predicted probability of improving is 0.596 according to CRISS index.

	Baseline	Follow-up	Absolute change
Age	64.65 yrs		
Disease duration (months)	30.74		
Global assessments			
Patient global assessment (0-10)	1	0	-1
MD global assessment (0-10)	7	4	-3
Muscoloskeletal			
HAQ-DI (0-3)	0.375	0.250	-0.125
Tendon friction rubs	No	No	No change
Skin			
MRSS (0-51)	21	15	-6
Patient interference skin last month	8	5	-3
Lung			
FVC% predicted	86	81	-5
Breathing VAS (0-10)	0	0	0
Renal			
Renal crisis	Yes	Yes	No change
Gastrointestinal			
GI VAS (0-10)	0	0	0
Body Mass Index (BMI)	25.12	24.82	-0.3
Raynaud’s			
Raynaud’s VAS (0-10)	3	4	1
Digital ulcers			
Digital ulcers VAS (0-10)	0	8	8
Number of digital ulcers	0	0	0
HRQOL			
Pain VAS (0-10)	0	2	2
Fatigue (SF-36 Vitality scale) (0-100)	35.12	35.12	0.0

Appendix Table 5. Example of a patient rated “worsened by the experts”. Predicted probability of improving is 0.002 according to the CRISS index.

	Baseline	Follow-up	Absolute Change
Age at baseline	53.6 yrs		
Disease duration at baseline (months)	43.3		
Global assessments			
Patient global assessment (0-10)	1	2	1
MD global assessment (0-10)	1	2	1
Muscoloskeletal			
HAQ-DI (0-3)	0	0	0
Tendon friction rubs	No	Yes	Change to worsen
Skin			
MRSS (0-51)	7	5	-2
Patient interference skin last month	3	2	-1
Lung			
FVC% predicted	87	80	-7
Breathing VAS (0-10)	0	1	1
Renal			
Renal crisis	No	No	No change
Gastrointestinal			
GI VAS (0-10)	0	1	1
Body Mass Index (BMI)	24.68	24.68	0
Raynaud’s			
Raynaud’s VAS (0-10)	0	3	3
Digital ulcers			
Digital ulcers VAS (0-10)	0	0	0
Number of digital ulcers	0	0	0
HRQOL			
Pain VAS (0-10)	1	1	0
Fatigue (SF-36 Vitality scale) (0-100)	37.52	35.10	-2.42

Appendix Table 6. One variable logistic model using expert consensus definition of improved vs. not

Variable	Area under the curve (AUC)	Sensitivity	Specificity	Brier Score
MRSS	0.9231	0.8392	0.8793	0.108
FVC predicted	0.7906	0.6429	0.7586	0.184
MD global	0.7743	0.7143	0.7241	0.197
Patient global	0.7448	0.7143	0.6207	0.204
HAQ-DI	0.7107	0.6429	0.6897	0.200
Pain	0.6857	0.6071	0.7586	0.218
Vitality	0.6856	0.4643	0.7414	0.225
VAS Breathing	0.6670	0.375	0.8103	0.219
GI VAS	0.6667	0.7857	0.4483	0.220
Patient interference skin	0.6601	0.5179	0.7586	0.226
Raynaud's VAS	0.6190	0.4286	0.7241	0.238
Tendon friction rubs	0.5640	0.2321	0.8966	0.245
Digital ulcers VAS	0.5503	0.2857	0.7931	0.247
BMI	0.4946	0.1786	0.8276	0.250
Number of digital ulcers	0.4764	0.0179	0.931	0.249

Appendix Table 7. Two variable logistic model using expert consensus definition of improved vs. not

Variable	Area under the curve (AUC)	Sensitivity	Specificity	Brier Score
MRSS, FVC predicted	0.9632	0.8929	0.9138	0.068
MRSS, HAQ-DI	0.9615	0.9107	0.8793	0.076
MRSS, Patient global	0.9560	0.875	0.8966	0.081
MRSS, MD global	0.9450	0.875	0.9310	0.094
FVC predicted, HAQ-DI	0.8519	0.7679	0.8448	0.158
FVC predicted, Patient global	0.8548	0.7679	0.8448	0.152
FVC predicted, physician global	0.8544	0.750	0.8103	0.158
HAQ-DI, Patient global	0.7982	0.7143	0.7241	0.184
HAQ-DI, physician global	0.8094	0.6607	0.7931	0.181
Patient global, physician global	0.8265	0.7321	0.7759	0.170