| manuscript.: | |
|---|---|
| | |

# The American College of Rheumatology provisional 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, MPH[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 W. 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 Matucci-Cerinic, MD, PhD[15], Janet Pope, MD[16], Susanna M. Proudman, MBBS[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, [14]Toronto Scleroderma Program, Toronto Western Hospital, University of Toronto, Toronto, Ontario, Canada, [15] Department of Experimental & Clinical Medicine, Division of Rheumatology AOUC, 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 and Discipline of Medicine, University of Adelaide, , Adelaide, SA, Australia, [18]Genentech/Roche, San Francisco, CA, USA, [19]Royal Brompton Hospital, London, UK

**Corresponding author**
Dinesh Khanna, MD, MSc
Professor of Medicine
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. Khanna has/had consultancy relationship with and/or has received research funding from Bayer, Biogen Idec, Bristol-Myers Squibb, Celgene, Cytori, EMD Serono, Forward, Genentech/ Roche, Gilead, Glaxo SmithKline, Lycera, Medac, Sanofi-Aventis/Genzyme, and Seattle Genetics.

ACR Provisional Composite Index for Scleroderma Clinical Trials

Drs. Assassi, Baron, Berrocal, Clements, Giannini, Mayes, Schiopu, Phillips, Pope, Wong and Wells have no conflicts related to this project.

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. Merkel has/had consultancy relationships with Actelion, ChemoCentryx, Glaxo-Smith-Kline, and Sanofi and has received research funding from Actelion, Bristol-Myers Squibb, Celgene, GlaxoSmithKline, and Genentech/ Roche.

Dr. Steen has/had consultancy relationship with and/or has received research funding from Actelion, Bayer, Bristol-Myers Squibb, Celgene, CSL Behring, Cytori, Genentech/ Roche, Gilead, InterMune, Sanofi-Aventis/Genzyme, and United Therapeutics.

Dr. Simms is on the Gilead Speakers bureau, has/had consultancies with Actelion, Cytori. Has/had grant support from Actelion, Gilead, Medimmune, and InterMune.

Dr. Allanore has/had consultancy relationship and/or has received research funding with Actelion, Bayer, Behring, Biogen Idec, Genentech/ Roche, Inventiva, Pfizer, Sanofi/Genzyme, Servier, and UCB.

Dr. Denton has/had consultancy relationship and/or has received research funding from Actelion, Genentech/ Roche, Pfizer, GlaxoSmithKline, BMS, CSL Behring, Novartis, Sanofi-Aventis, Inventiva, and 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, Bristol -Myers Squibb, Sanofi-Aventis, United BioSource Corporation, Genentech/ Roche, Medac, Biovitrium, Boehringer Ingelheim Pharma, Novartis, 4 D Science, Active Biotec, Bayer-Schering, Sinoxa/ Serodapharm, EpiPharm, Biogen Idec, Inventiva, GSK and Pharmacyclics.

Dr. Johnson is supported by the Canadian Institutes of Health Research Clinician Scientist Award.

Dr. Matucci Cerinic has/had consultancy relationship and/or has received research funding with Actelion, Bayer, Behring, Bristol-Myers Squibb, MSD Pfizer, and UCB.

Dr. Proudman has research grants and consultancies for Actelion, Bayer and Glaxo Smith Kline.

Dr. Siegel is an employee of Genentech.

Dr. Furst has/had research grants with AbbVie, Actelion, Amgen, Bristol-Myers Squibb, Gilead, GSK, NIH, Novartis, Pfizer, Genentech/ Roche, UCB and has/had consultancies

with AbbVie, Actelion, Amgen, Bristol-Myers Squibb, Cytori, Janssen, Gilead, GSK, NIH, Novartis, Pfizer, Genentech/ Roche, UCB and does CME programs with AbbVie, Actelion, and UCB.

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Abstract**

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

**Methods**: We developed 150 paper patient profiles with standardized clinical outcome elements (core set items) using patients with dcSSc. Forty scleroderma experts rated 20 patient profiles each and assessed whether each patient had improved or not over a period of 1 year. Using profiles where raters reached a consensus on whether the patients were improved vs. not (79% of profiles examined), we fit logistic regression models where the binary outcome referred to whether the patient was improved or not, and the change in the core set items from baseline to follow-up were entered as covariates. We tested the final index in a previously completed RCT.

**Results**: Sixteen of 31 core items were included in the patient profiles after a consensus meeting and review of test characteristics of patient-level data. The logistic regression model that included the following core set items: changes in the Rodnan skin score, forced vital capacity (FVC)% predicted, patient and physician global assessments, and HAQ-DI over 1 year had sensitivity of 0.982 (95%CI: 0.981-0.983), specificity of 0.931 (95% CI: 0.930-0.932), and had the highest face validity. Subjects with a significant decline in renal or cardiopulmonary involvement were classified as not improved, regardless of improvements in other core items. The index was able to differentiate the effect of methotrexate from placebo in a 1-year RCT ($p< 0.05$).

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Conclusion**: We have developed CRISS that is appropriate for use as an outcome assessment in RCT of early dcSSc.

ACR Provisional Composite Index for Scleroderma Clinical Trials

## Background

Systemic sclerosis (scleroderma, SSc) is one of the most life-threatening rheumatic diseases (1, 2), and is associated with substantial morbidity and many detrimental effects on health-related quality of life (3). In recent years, progress has been made in the development and validation of outcome measures and refinement of trial methodology in SSc (4-7). These advances were paralleled by an increased understanding of the pathogenesis of SSc (8) and development of potential targeted therapies (9). The Modified Rodnan Skin Score, a measure of skin thickness (6), 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 captures multiple organ involvement and patient-reported outcomes.

An accepted, validated, composite response index in dcSSc could substantially facilitate drug development and clinical research. Compared to individual outcome measures, a composite index has the potential to be more responsive to change (10-12), improve assessment of therapeutic interventions, and facilitate the comparison of responses across trials. Regulatory and funding agencies would then have greater confidence in proposals for interventions.

Our objective was to develop a Composite Response Index in Systemic Sclerosis (abbreviated CRISS) for use in clinical trials.

## Patients and Methods

The index was developed using well-accepted expert consensus (13) and data-driven approaches (Figure 1), including the American College of Rheumatology standards for the development of response criteria (14). Details are included in the Supplementary material. The basic process was as follows: i) We conducted a consensus exercise to select domains and outcome measures (core items) for potential inclusion in the composite response index. ii) We then tested the psychometric properties of the core items in a longitudinal cohort of patients followed over 1 year to assess the items' feasibility, reliability, validity, and sensitivity to change. iii) We developed a set of 150 patient profiles based on the data generated from the cohort study (and using the core items). Forty scleroderma experts were invited to classify each patient profile as improved or not improved. iv) We performed statistical reduction of the data to a minimum number of domains and core items, which retained the maximally responsive index and was acceptable to the experts (face validity). v) We then tested the ability of the composite response index to discriminate among therapies using results from a previously published randomized controlled trial (RCT). The following paragraphs describe each step in greater detail.

**(i). Structured consensus exercise to develop domains and core items**: We conducted a structured, 3-round Delphi exercise to reach consensus on core items for clinical trials of SSc the details of which have been published elsewhere (5). Briefly, an initial list of potential domains and items was composed by a steering committee and then the members of the Scleroderma Clinical Trials Consortium (SCTC). Round 1 asked the SCTC members to list items in 11 pre-defined domains and Round 2 asked

respondents to rate the importance of the chosen items on a 1-9 ordinal scale. This was followed by a face-to-face meeting where, under expert facilitators, consensus was reached using the Nominal Group Technique (13) about the domains and core items to test in a database (5). During this exercise, the Steering Committee discussed the feasibility, reliability, redundancy, and validity of the items.

**(ii). Data collection and evaluation of psychometric properties in a longitudinal observational cohort:** Due to a lack of positive trials in dcSSc and as a consequence of the fact that previous trials did not include some of the core items chosen in the consensus exercise (15), we launched a longitudinal observational cohort (the CRISS Cohort) of patients with early dcSSc (< 5 years from 1$^{st}$ non-Raynaud's phenomenon sign or symptom) at 4 US Scleroderma Centers (16). The observational cohort, recruited over 1 year, included 200 patients 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. Patients were followed for 12 months and outcomes were collected at baseline and 12 months. Exclusion criteria included life expectancy of less than 1 year and non-proficiency in English. All core items that emerged from the consensus meeting were included to enable an assessment of their psychometric properties (e.g., feasibility, reliability, and face, content, and construct validity [including sensitivity to change]) (17). Feasibility was defined as completion of the core set item by > 50% of subjects at two time points, redundancy was defined as either a Spearman or Pearson correlation coefficient of at least 0.80 at baseline or during follow up. Sensitivity to change was calculated over the 1-year period using appropriate patient and physician

anchor and transition questions. For example, a modified Likert scale (transition health question) was employed by physicians and patients at the 1-year follow-up visit to determine the change in overall condition during the prior year on a scale from 1 (“much better”) to 5 (“much worse”). Responses of 1 or 2 were considered an improvement in health, ratings of 4 or 5 were considered a decline in health, and a rating of 3 was considered to mean that there was no appreciable change in overall health. For this analysis, those who answered “1” or “2” were categorized as “improved” on both transition questions and those who scored “3”, “4” or “5” were categorized as “not improved”. 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 (18). Core items that were significant at predefined $p< 0.20$ (for dichotomous measures) or had an effect size ≥ 0.20 in the “Improved” group (with respect to either patient or physician assessments) were included in the next stage.

Eight Steering Committee members (see Acknowledgement section) reviewed the data and scored each core item on an ordinal scale (1-4) for feasibility, reliability, and face, content, and construct validity [including sensitivity to change] using the modified content validity index matrix (19): a score of 4 (highest score) was assigned when the item referred 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 that 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 a score of 1 (lowest score) meant that the attribute should

definitely not be used as a core item. Experts could also assign “not applicable” if they were unfamiliar with an item or different aspects of feasibility, reliability, and validity for the item. Items scored as 3 or 4 were considered supportive of an individual item.

Based on results from psychometrics analysis and expert input, a modified Nominal Group Technique exercise was conducted via webinar by E. Giannini where consensus was defined a priori as ≥75% agreement on each item of the matrix and overall inclusion/ exclusion of the item as a core item. During the NGT webinar, summary statistics were provided for each core set item and the moderator encouraged to discuss each item by each committee member and then as a group. This process ensured all participants had an opportunity to contribute. Subsequently, each item was rescored (if the committee member felt that it should be changed) and summary statistics were generated. Items that were found to lack feasibility, reliability, and validity (<75% raters assigning score of 3 or better) were excluded from the next step.

**(iii). Development and ratings of representative patient profiles:** In this step, we developed 150 paper patient profiles using actual data from the CRISS Cohort. To have sufficient data for the representative patients, we also obtained data from early dcSSc (defined as the disease duration < 5 years) in the Canadian Scleroderma Research Group (CSRG) database (20), a large observational Canadian scleroderma cohort. Since patient interviews were not performed as part of the consensus meeting (Step i), the medical literature was searched to assess the most prevalent/ bothersome issues faced by patients with SSc (21-23). Based on this, pain and fatigue (assessed by the SF-36 vitality scale), were included as part of the patient profiles.

ACR Provisional Composite Index for Scleroderma Clinical Trials

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 profiles were randomly assigned to experts based on their location (North America [N=29] vs. Europe [N=21] vs. Australia [N=4]) and years of experience with management of SSc (>10 years [N=38] vs. ≤ 10 years of scleroderma experience [N=16]) to prevent systematic bias in rating due to practice patterns. 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 as improved or worsened, by how much did the patient's condition change?: considerably, somewhat, or a little.*

*3. How would you rank the three most important core items that influenced your decision regarding change or stability?*

Consensus was met if at least 75% among those who rated the same patient profile agreed that the patient had improved, stabilized, or worsened. 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 Technique exercise to discuss each profile in detail . These patient profile ratings were then added to the previous voting and percentage consensus was recalculated. If the proportion of agreement on a patient profile was ≥ 75%, the case was deemed as having reached consensus. This process produced a final list of 16 core items. Finally, we sought consensus among SSc experts on the level of change in internal organ involvement that would classify a patient as not improved.

(iv)**. Development of response definitions**

Using only profiles where consensus was reached, we fit logistic regression models to the binary outcome, i.e., whether a patient had been rated by experts as being improved (=1) vs. not improved (=0), Not improved included scenarios rated as either no change or worsened. . We examined various models, increasing at each step the number of predictors (core set items) included in the logistic regression model. For each model, we calculated sensitivity, specificity, and area under the curve (AUC). Additionally, using the estimates of the logistic regression beta 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 was evaluated in several ways. Using the predicted probabilities in their continuous form, accuracy in the predictions was quantified by the Brier score (24); the model with the lowest Brier Score is interpreted to have the best predictive performance.

We also tested whether the predicted probabilities had a different distribution for the patient profiles which were rated improved by the experts and for those that were rated not improved. We assessed the difference in the two distributions via the non-parametric Mann-Whitney test. We examined whether the predicted probabilities could be transformed into binary classifications by choosing a threshold and defining "improved" for all patients for which the predicted probability is above the chosen threshold and "not improved" for all patients for which the predicted probability is below the threshold. To identify which threshold (i.e., 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 of the predicted binary classification of patients into improved (=1) or not improved (=0). We made a plot of the sensitivity and specificity as a function of each 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 determine whether there was a clear distinction among the 16 core items in their helpfulness to guide raters in determining whether a patient was improved or not, we conducted a cluster analysis. To evaluate the contribution of each core component to the final CRISS, we computed the generalized coefficient of determination or pseudo $R^2$ for logistic regression (25).

**(v). Preliminary evaluation in an independent cohort**

The composite index was tested in a randomized controlled trial of methotrexate vs. placebo in early dcSSc (26). This trial was chosen as individual patient data were recorded and all final core items were available in this database. We applied the CRISS 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 each subject “improved” or “not improved” depending on whether the predicted probability was above the threshold with the highest sensitivity and specificity (identified in Step # iv). 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 two groups of subjects – placebo and methotrexate) 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

**(i). Structured Consensus Exercise to develop domains and core items**

. A total of 50 SCTC investigators participated in Round 1, providing 212 unique items for the 11 domains, and rated 177 items in Round 2. The ratings of 177 items were reviewed by the Steering Committee, and 11 domains and 31 items were identified as the core items that met the Outcome Measures in Rheumatology (OMERACT ) filters of truth, feasibility, and discrimination. 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. Attendees of OMERACT conference in 2008 provided input during the consensus exercise (4, 27).

**(ii). Data collection and evaluation of psychometric properties in a longitudinal observational cohort**

CRISS Cohort

ACR Provisional Composite Index for Scleroderma Clinical Trials

Two hundred patients with early dcSSc were recruited at baseline and 150 had both 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 Hispanic 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).

Core items that lacked feasibility due to low completion rate (< 50%) at 1 year included durometer (a device to measure the skin hardness (28)), right heart catheterization, Borg dyspnea index, 6-minute walk test, and Raynaud's Condition Score (29) (required daily patient diary records).

Using the patient global assessment as the metric to classify patients as improved vs. not, 57% of patients were rated as "improved" and 43% were rated as "not improved". Using physician global assessment, 58% of patients were rated as "improved" and 42% were rated as "not improved". The Spearman correlation among the definitions was 0.46, supporting use of 2 global transition questions. Using these transition questions, 5 items were found to be not responsive to change or occurred in less than 10% of the cohort: tender joint count, presence of renal crisis, estimated GFR, body mass index, presence of digital ulcers, and erythrocyte sedimentation rate. A modified Nominal Group review was performed wherein consensus was achieved on 16 core items that should be used for the development of paper patients. It was decided to keep renal crisis and presence/absence of digital ulcers as core items due to their impact on

prognosis in early dcSSc. No redundancy was noted in the core items at baseline and change scores as assessed by the correlation coefficients (Appendix Tables 1-2).



**(iii). Development and ratings of representative patient profiles**

A total of 150 patient profiles were rated by 40 of 54 invited experts (74% completion) (20 profiles rated by each expert; examples shown in the Appendix Tables 3-5). The median number of experts that rated a profile was 6, and the range was 4-13. In response to the instruction, “Please rank the most important core items that influenced your decision regarding change or stability”, experts ranked MRSS as the “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 items were ranked as most influential in the decision making less than 2% of the time.

Initially, consensus was achieved for 107 (71.3%) of the patient profiles. The Steering Committee then rescored the remaining 43 profiles as improved, worsened, or stable, and final consensus was achieved in 118 (78.7%) profiles. These profiles were then used for developing the response definitions.

**(iv). Development of response definitions**

Logistic regression models

There were 118 profiles for which consensus was reached; these profiles were used in the statistical models that examined response definitions regarding improvement based

on change in the 16 core items. In 1-core item models (models where only one covariate was included), AUC ranged from 0.47 (for the model including as single covariate the change in presence/absence of new digital ulcers) to 0.92 (for the model including as single covariate the change in MRSS; Appendix Table 6). In a 2-core item 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 as it did not include either the patient or physician perspective. 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-core item 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 2). The clustering algorithm supported 5-core item model with the first cluster contained the following 5 items—MRSS, FVC% predicted, patient global assessment, physician global assessment, and HAQ-DI and the second cluster included all the remaining core items (Table 3). This model had a sensitivity of 0.9821 (95% CI: [0.9816, 0.9827]), specificity of 0.9310 (95% CI: [0.9300, 0.9321]), and AUC of 0.9861. The Brier score was 0.038 (lower score indicates a better predictive performance). As the data were not normally distributed, non-parametric tests were used to assess whether 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 2a). Using depiction of sensitivity vs. specificity for improved vs. not improved group, a threshold of 0.6 had the best combination of specificity and sensitivity values (Figure 2b). The 5-core item logistic regression model can be used not only to derive predicted probabilities of improving on

a 0-1 scale, but also to derive the log-odds of improving for each subject. The latter can take any value: a log-odds of 0 means that a subject has equal odds to improve as to not improve (i.e. predicted probability of 0.5 or 50%) while a positive (negative) log-odds means that a subject has greater (lower) odds of improving.

Contribution of 5 core components to the CRISS

We computed the pseudo $R^2$ for the logistic regression models that included all the 5 core items of the CRISS as well as the pseudo $R^2$ for logistic regression models including each single predictor. Combined, the 5 core items explained 89.3% of the variability in the data. Individually, when used in a single-core item logistic regression model, MRSS explained 66.3% of the variation, FVC% predicted explained 36.1% of the variation, physician global assessment explained 24.5% of the variation, patient global assessment explained 23.7% of the variation, and HAQ-DI explained 28.5% of the variation.

To assess how changes in the core items are related to the predicted probabilities of improving on each patient profile, Appendix Figure 1(a)-(e) presents a 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, all calculated from baseline to 12 months. 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.

Defining a patient who is not improved irrespective of improvement in other core items

The Steering Committee considered circumstances in which 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 pulmonary arterial hypertension). There was consensus that such patients should be defined as not improved in a clinical trial. The Steering Committee voted and determined that the following items met this definition: new onset of renal crisis, new-onset or worsening lung fibrosis, new onset of pulmonary arterial hypertension, or new onset of left ventricular failure (Table 4). The international experts subsequently endorsed these definitions as well.

**(v). Preliminary evaluation in a randomized controlled clinical trial**

We used the individual patient data from a clinical trial comparing treatment of dcSSc with methotrexate vs. placebo to assess our definition of response (26). 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 with complete baseline and 1-year data and classified them into improved and not improved using a probability cutoff of 0.6 (decided analytically in Step #iv). With this criterion, 11 of 19 subjects who received methotrexate were rated as improved whereas 3 of 16 subjects in the placebo

group were rated as improved (p=0.04; Appendix Figure 2). When the data were assessed as a continuous measure, the distribution of the predicted probability for improvement was statistically different between the placebo and the methotrexate groups (p= 0.02).

**Application in a clinical trial**

The CRISS was developed with a goal to summarize the changes in the clinical and patient-reported outcomes in a single composite score that conveys the likelihood (or probability) that the patient has improved. If there is an effective agent for treatment of dcSSc, the assumption is that the patient treated with the agent will have a higher probability of improvement as summarized by CRISS vs. placebo or an ineffective agent. CRISS is a 2-step process for use in a clinical trial and is described in Table 4. In Step 1, subjects who develop new onset of renal crisis, new-onset or worsening lung fibrosis, new onset of pulmonary arterial hypertension, or new onset of left ventricular failure during the trial are considered as not improved and assigned a probability of improving equal to 0.0. For the remaining subjects with complete data, Step 2 involves computing the predicted probability of improving for each subject using the equation in Table 4. Subjects for whom the predicted probability is greater or equal to 0.60 are considered improved, while subjects for whom the predicted probability is below 0.60 are considered not improved. The 2 groups (drug vs. placebo or an active comparator) can then be compared in a 2x2 table 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. For trials that incorporate components of CRISS at multiple time points, the CRISS was developed using data at 12 month. Therefore, there is lack of data to support its performance at earlier time periods. We recommend using 12-month data as primary/ secondary outcome measure and using others such as baseline to 3, 6, and/or 9 months as exploratory outcomes. We recommend capturing the data at each patient visit using specific case report forms for organ involvement. We also encourage developing an adjudication committee that can help with validating that cardio-pulmonary-renal involvement occurred. If case report forms are not developed and included in the trial, then these should be captured as part of adverse events [all of them should be classified as serious adverse events]. Specifically, non-availability of this data [if no specific case report forms are developed upfront] should not be taken as missing data as these should be captured as adverse events/ serious adverse events. If there is missing data for the components of Step 2, we recommend considering the reason for missingness and using appropriate statistical methods. Missing data for the 5 components in Step 2 should be imputed till Month 12 before calculating the score.

## Discussion

We have developed a composite response index for trials (CRISS) in early dcSSc using well-established consensus and data-driven approaches. The CRISS includes core items that assess change in two common and prominent manifestations of early dcSSc (skin and interstitial lung disease), functional disability (as assessed by the HAQ-DI), and patient and physician global assessments. In addition, the CRISS captures clinically meaningful declines in internal organ involvement requiring treatment that classify the patient as having not improved (regardless of changes in other parameters) during the clinical trial. We subsequently tested CRISS using data from a clinical trial and showed that the CRISS identified different probabilities of improvement for early dcSSc subjects in the placebo and methotrexate groups, suggesting that methotrexate has the potential to improve the overall health condition in the dcSSc subjects after 1 year.

Traditionally, trials in early dcSSc have focused on skin or lung involvement (30, 31). MRSS has been used as the primary outcome measure for the trials of skin fibrosis (6). MRSS meets the OMERACT criteria as a fully validated measure of outcome (32), but is also a surrogate of internal organ involvement and mortality in early dcSSc (33, 34). However, clinical trials in dcSSc to date have largely been “negative” and MRSS has been questioned as a primary outcome measure where post-hoc analysis of negative trials has shown stability/improvement in MRSS over time (35, 36). The CRISS incorporates multisystem involvement in dcSSc and includes the patient perspective and the impact of the disease on functional disability. CRISS was developed with a goal to summarize the changes in the clinical and patient-reported outcomes in a single composite score that conveys the likelihood (or probability) that the patient has

improved. For an effective treatment for dcSSc, the assumption is that patients treated with the agent will have a higher probability of improvement as summarized by CRISS vs. placebo or an ineffective agent.

The CRISS is calculated as a 2-step process (Table 4). The first step evaluates clinically significant decline in renal or cardiopulmonary involvement that requires treatment; if present, the patient is classified as not improved. The definitions chosen for internal organ involvement were based on published data and expert opinion that was felt to be clinically significant and would trigger pharmacologic management. The second step assesses remaining patients and calculates the predicted probability of improvement. Here, the Steering Committee discussed different response definitions and decided on using a data-driven definition as suggested by the ACR Criteria subcommittee (37). In addition, data-driven definitions of disease activity have been successfully used for regulatory approval in other rheumatic diseases (38, 39).

The goal of CRISS is to assess if new pharmacologic agents have an impact on overall disease activity/severity. Our hope is that the use of CRISS in clinical trials on dcSSc will greatly facilitate the interpretation of results and form the basis for drug approvals. Rather than using numerous outcomes that vary from trial to trial, the core set of items 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-analyses. It will likely also allow a decrease in the number of patients necessary for appropriately powered clinical trials, as has been the case for other composite indices in rheumatoid arthritis. It should also be noted that the use of CRISS does not preclude the addition of other items in a trial; it

simply provides one standardized outcome that can be easily compared and understood across trials. The individual components of CRISS would each likely be important secondary outcomes to assess in any trial. If the goal of a trial is to focus on a particular organ (e.g., use of vasodilators for underlying digital ulcers), then the CRISS can be used as a secondary measure.

The initial panel of domains (N=11) and items (N=31) offered a comprehensive view of the marked heterogeneity of SSc and at first was modeled on the comprehensive structure of the BILAG and SLEDAI measures used in trials of systemic lupus erythematous (40). However, many items were discarded based on lack of sensitivity to change in our actual data gathering exercise and others were shown to lack feasibility. As an example, the CRISS does not include items for worsening gastrointestinal disease or digital ulcers but it is anticipated that patient and physician global assessments will capture these. The data-driven approach used in the development of the CRISS strongly supports the relatively simple and accessible panel of items.

There are other indices that have been developed in SSc. The European Scleroderma Study Group (41) has proposed a composite index to assess SSc-related disease activity in routine clinical care but it has not been validated as an outcome measure in clinical trials. A severity index (42), a measure that encompasses disease activity and damage has been proposed and can be used in trials to complement CRISS.

This study has several strengths. It is the first concerted effort by the scleroderma research community to address the lack of a robust composite index for this multisystem disease. We used well-accepted expert consensus and data-driven

methodologies and successfully derived the index in early dcSSc. The index addresses several domains of illness by capturing single-organ involvement in early dcSSc, patient assessment of overall disease, functional disability, and physician global assessment. We were only able to test the index in a single, small RCT that had loss to follow-up; CRISS therefore requires further validation in a prospective RCT of adequate size.

Our study is not without limitations. The CRISS was developed for early dcSSc and may not be valid for late dcSSc or limited cutaneous SSc (lcSSc). A similar exercise in late lcSSc might focus on vascular complications such as digital ulcers, calcinosis, or pulmonary arterial hypertension but might not include MRSS. The majority of past and ongoing therapeutic clinical trials are focused on early dcSSc due to dynamic changes in skin and internal organ involvement that may be responsive to pharmacologic intervention. We did not obtain patient input during the development of the index. We acknowledge this limitation and searched the literature for patient input regarding scleroderma (21, 22); this led to inclusion of fatigue and pain during the development of patient profiles but neither measure remained in the final core set of items following the Nominal Group exercises. Nonetheless, two of the constituent core items of the CRISS include patient global assessment and patient-reported functional assessment. We also note that CRISS should be considered as a preliminary index. Although the index was tested in a RCT, missing data in the trial (>50%) precludes definitive conclusion and the CRISS may need to be revised as more data becomes available from future trials. We had 118 paper patient profiles where there was expert consensus and these profiles were used to develop different response definitions. Although this is standard

methodology, this may be suboptimal for testing 16 core set items. This may also explain high AUC of 0.968 for the index.

Lastly, as our goal was to develop a response index for change, baseline scores are not included in the algorithm. Other indices such as ACR 20 for rheumatoid arthritis or ACR 30 for juvenile arthritis also employ only changes in core items and not baseline values. Although the baseline scores can influence the changed scores, randomization should provide a balanced cohort.

In conclusion, we have developed a novel composite index for use in clinical trials in early dcSSc. The index should be considered provisional and needs to be validated in RCTs of dcSSc.

**Acknowledgement**

Research reported in this publication was supported by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number UO1 AR055057. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Drs. Khanna and Berrocal were also supported by NIH/NIAMS K24 AR063120. Dr. Johnson is supported by the Canadian Institutes of Health Research Clinician Scientist Award.

The Steering committee included: Murray Baron, MD, Philip J. Clements, MD, MPH, Daniel E. Furst, MD, Dinesh Khanna, MD, MS, Maureen D. Mayes, MD, MPH, Peter A. Merkel, MD, MPH, James R. Seibold, MD, and Virginia Steen, MD

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, et al. Mortality in systemic sclerosis: an international meta-analysis of individual patient data. Am J Med 2005;118(1):2-10.
2. Elhai M, Meune C, Avouac J, Kahan A, Allanore Y. Trends in mortality in patients with systemic sclerosis over 40 years: a systematic review and meta-analysis of cohort studies. Rheumatology (Oxford) 2012;51(6):1017-26.
3. Khanna D, Kowal-Bielecka O, Khanna PP, Lapinska A, Asch SM, Wenger N, et al. Quality indicator set for systemic sclerosis. Clin Exp Rheumatol 2011;29(2 Suppl 65):33-9.
4. Khanna D, Distler O, Avouac J, Behrens F, Clements PJ, Denton C, 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):2356-61.
5. Khanna D, Lovell DJ, Giannini E, Clements PJ, Merkel PA, Seibold JR, et al. Development of a provisional core set of response measures for clinical trials of systemic sclerosis. Ann Rheum Dis 2008;67(5):703-9.
6. Khanna D, Merkel PA. Outcome measures in systemic sclerosis: an update on instruments and current research. Curr Rheumatol Rep 2007;9(2):151-7.
7. 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 Exp Rheumatol 2012;30(2 Suppl 71):97-102.
8. Abraham DJ, Varga J. Scleroderma: from cell and molecular mechanisms to disease models. Trends Immunol. 2005;26(11):587-95.
9. Nagaraja V, Denton CP, Khanna D. Old medications and new targeted therapies in systemic sclerosis. Rheumatology (Oxford) 2014.
10. van der Heijde DM, t Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA, et al. 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):916-20.
11. 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):477-84.
12. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38(6):727-35.
13. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011;41(2):95-105.
14. Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, et al. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55(3):348-52.

15. Merkel PA, Silliman NP, Clements P, Denton CP, Furst DE, Mayes M, et al. Patterns and Predictors of Change in Outcome Measures in Clinical Trials in Scleroderma Arthritis Rheum 2005;52:282-3.
16. Wiese AB, Berrocal VJ, Furst DE, Seibold JR, Merkel PA, Mayes MD, et al. Correlates and responsiveness to change of measures of skin and musculoskeletal disease in early diffuse systemic sclerosis. Arthritis Care Res (Hoboken) 2014:1731-9.
17. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1(1):73-5.
18. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988. p.273-406.
19. Davies EH, Surtees R, DeVile C, Schoon I, Vellodi A. A severity scoring tool to assess the neurological features of neuronopathic Gaucher disease. J Inherit Metab Dis 2007;30(5):768-82.
20. Fan X, Pope J, Baron M. What is the relationship between disease activity, severity and damage in a large Canadian systemic sclerosis cohort? Results from the Canadian Scleroderma Research Group (CSRG). Rheumatol Int 2009:1205-10.
21. 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):762-67.
22. Suarez-Almazor ME, Kallen MA, Roundtree AK, Mayes M. Disease and Symptom Burden in Systemic Sclerosis: A Patient Perspective. J Rheumatol 2007;34(8):1718-26.
23. Stamm TA, Mattsson M, Mihai C, Stocker J, Binder A, Bauernfeind B, et al. Concepts of functioning and health important to people with systemic sclerosis: a qualitative study in four European countries. Ann Rheum Dis 2011;70(6):1074-9.
24. Gneiting T, Raftery A. Strictly proper scoring rules. J Am Stat Assoc 2007;102:359-78.
25. Nagelkerke NGD. A note on a general definition of the coefficient of Determination. Biometrika 1991;78(3):691-2.
26. Pope JE, Bellamy N, Seibold JR, Baron M, Ellman M, Carette S, et al. A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma. Arthritis Rheum 2001;44(6):1351-8.
27. Furst D, Khanna D, Matucci-Cerinic M, Clements P, Steen V, Pope J, et al. Systemic sclerosis - continuing progress in developing clinical measures of response. J Rheumatol 2007;34(5):1194-1200.
28. Merkel PA, Silliman NP, Denton CP, Furst DE, Khanna D, Emery P, 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.
29. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46(9):2410-20.
30. Khanna D, Clements PJ, Furst DE, Korn JH, Ellman M, Rothfield N, 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):1102-11.

31. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med 2006;354(25):2655-66.
32. 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):1630-47.
33. Clements PJ, Hurwitz EL, Wong WK, Seibold JR, Mayes M, White B, 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):2445-54.
34. Steen VD, Medsger TA, Jr. Severe organ involvement in systemic sclerosis with diffuse scleroderma. Arthritis Rheum 2000;43(11):2437-44.
35. Merkel PA, Silliman NP, Clements PJ, Denton CP, Furst DE, Mayes MD, 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):3420-9.
36. Amjadi S, Maranian P, Furst DE, Clements PJ, Wong WK, Postlethwaite AE, et al. 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):2490-8.
37. Singh JA, Solomon DH, Dougados M, Felson D, Hawker G, Katz P, et al. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55(3):348-52.
38. 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):177-81.
39. Luijten KM, Tekstra J, Bijlsma JW, Bijl M. The Systemic Lupus Erythematosus Responder Index (SRI); a new SLE disease activity assessment. Autoimmun Rev 2012;11(5):326-9.
40. Yee CS, Isenberg DA, Prabu A, Sokoll K, Teh LS, Rahman A, et al. BILAG-2004 index captures systemic lupus erythematosus disease activity better than SLEDAI-2000. Ann Rheum Dis 2008;67(6):873-6.
41. Valentini G, D'Angelo S, Della RA, Bencivelli W, Bombardieri S. European Scleroderma Study Group to define disease activity criteria for systemic sclerosis. IV. Assessment of skin thickening by modified Rodnan skin score. Ann Rheum Dis 2003;62(9):904-5.
42. Medsger TA, Jr., Silman AJ, Steen VD, Black CM, Akesson A, Bacon PA, et al. A disease severity scale for systemic sclerosis: development and testing. J Rheumatol 1999;26(10):2159-67.
43. Steen VD, Mayes MD, Merkel PA. Assessment of kidney involvement. Clin Exp Rheumatol 2003;21(3 Suppl 29):29-31.
44. Hoeper MM, Bogaard HJ, Condliffe R, Frantz R, Khanna D, Kurzyna M, et al. Definitions and diagnosis of pulmonary hypertension. J Am Coll Cardiol 2013;62(25 Suppl):42-50.

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Table 1: Baseline demographics of patients who participated in the CRISS Cohort with baseline and 1 year data**

| | Baseline N | |
|---|---|---|
| Age, mean (SD) | 150 | 50.4 (11.7) |
| Female, N (%) | | 112 (75%) |
| Race, N (%)<br>Caucasian<br>African American<br>Asian<br>Other or not provided | 150 | <br>117 (78%)<br>13 (9%)<br>11 (7%)<br>9 (6%) |
| Ethnicity, N (%)<br>Hispanic<br>Non-Hispanic | 150 | <br>16 (11%)<br>134 (89%) |
| Disease duration from first non-Raynaud symptom (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) |
| Body mass index, mean (SD) | 96 | 26.02 (7.1) |
| Modified Rodnan skin score, mean (SD) | 150 | 21.4 (10.1) |
| Durometer, mean (SD) | 113 | 272.4 (64.5) |
| Forced vital capacity % predicted, mean (SD) | 140 | 82.32 (18.5) |
| Total lung capacity % predicted, mean (SD) | 109 | 87.83 (20.4) |
| Diffusion capacity of carbon monoxide % predicted, mean (SD) | 140 | 65.05 (20.9) |
| High-resolution computer tomography consistent with interstitial lung disease, N (%) | 99 | 79 (80) |
| 6-minute walking distance, mean (SD) | 50 | 421.6 (139.2) |
| 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 (59) |
| Large joint contractures, N (%) | 133 | 39 (29) |
| Digital ulcers, N (%) | 150 | 15 (10) |
| Health assessment questionnaire-disability index, mean (SD) | 150 | 1.0 (0.8) |
| Digital ulcers VAS (0-150), mean (SD) | 134 | 20.9 (40.9) |
| Raynaud's VAS (0-150), mean (SD) | 135 | 32.7 (40.8) |
| Breathing VAS (0-150), mean (SD) | 138 | 23.1 (36.7) |
| GI VAS (0-150), mean (SD) | 136 | 22.6 (34.4) |
| Disease severity VAS (0-150), mean (SD) | 138 | 56.4 (42.9) |
| Pain VAS (0-10), mean (SD) | 140 | 4.0 (2.8) |
| SF-36 PCS, mean (SD) | 138 | 37.6 (12.9) |

ACR Provisional Composite Index for Scleroderma Clinical Trials

| | | |
|---|---|---|
| SF-36 MCS, mean (SD) | 138 | 44.2 (6.0) |
| Physician global assessment VAS (0-10 cm), mean (SD) | 143 | 4.4 (2.2) |
| Patient global assessment VAS (0-10 cm), mean (SD) | 140 | 4.1 (4.0) |
| Antinuclear antibody, N (%) | 116 | 94 (81) |
| Anti-SCL-70 antibody, N (%) | 115 | 34 (30) |
| Serum creatine phosphokinase (IU/L) , mean (SD) | 127 | 143.9 (184.5) |
| Serum platelets (k/uL), mean (SD) | 143 | 315.2 (102.5) |
| Serum brain natriuretic peptide (pg/ml), mean (SD) | 105 | 161.3 (824.0) |
| Serum erythrocyte sedimentation rate (mm/hr), mean (SD) | 121 | 23.4 (22.6) |
| Serum C-reactive protein (mg/dL), mean (SD) | 116 | 2.1 (4.9) |

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

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Table 2. Final CRISS model consisting of 5 core items with highest face validity**

| Core items (calculated as changed from baseline to 1 year) | Area under the curve (AUC) | Sensitivity (95% CI) | Specificity (95% CI) | Unadjusted Beta coefficients | Standard errors |
|---|---|---|---|---|---|
| MRSS<br>FVC predicted<br>HAQ-DI<br>Patient global assessment<br>Physician global assessment | 0.9861 | 0.9821 (0.9816, 0.9827) | 0.9310 (0.9300, 0.9321) | -0.81<br>0.21<br>-0.40<br>-0.44<br><br>-3.41 | 0.21<br>0.08<br>0.24<br>0.26<br><br>1.75 |

MRSS= modified Rodnan skin score, FVC= Forced vital capacity, HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Table 3. The table describes ranking of the 16 core items by scleroderma experts and results of the cluster analysis**

| Core item | 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 skin interference last month | 2 (0.2%) | 21 (2.5%) | 22 (2.6%) | 2 |
| Number of digital ulcers | 9 (1.1%) | 11 (1.3%) | 17 (2%) | 2 |
| Presence of 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 |

MRSS= modified Rodnan skin score, FVC= Forced vital capacity, HAQ-DI= health assessment questionnaire-disability index, GI= gastrointestinal, VAS= visual analog scale, MRSS= modified Rodnan skin score

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Table 4. Application of CRISS in a clinical trial**

CRISS is a 2-step process.

**Step 1:** Subjects who develop new or worsening of cardiopulmonary and/or renal involvement due to systemic sclerosis are considered as not improved (irrespective of improvement in other core items) and assigned a probability of improving equal to 0.0. Specifically if a subject develops any of the following

- **New scleroderma renal crisis (43)**
- **Decline in forced vital capacity (FVC)% predicted ≥15% (relative), confirmed by another FVC% within a month, high resolution computer tomography (HRCT) to confirm interstitial lung disease (ILD; if previous high resolution computer tomography of chest did not show ILD) and FVC% predicted below 80% predicted***
- **New onset of left ventricular failure (defined as left ventricular ejection fraction ≤45%) requiring treatment***
- **New onset of pulmonary arterial hypertension (PAH) on right heart catheterization (44) requiring treatment*.** PAH is defined as mean pulmonary artery pressure ≥ 25 mm Hg at rest and an end-expiratory pulmonary artery wedge pressure ≤ 15 mm Hg and a pulmonary vascular resistance >3 Wood units

***= Attributable to systemic sclerosis**

**Step 2:** For the remaining subjects, Step 2 involves computing 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 $\Delta_{MRSS}$ indicates the change in MRSS from baseline to follow-up, $\Delta_{FVC}$ 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 ($Time_2 - Time_{baseline}$).

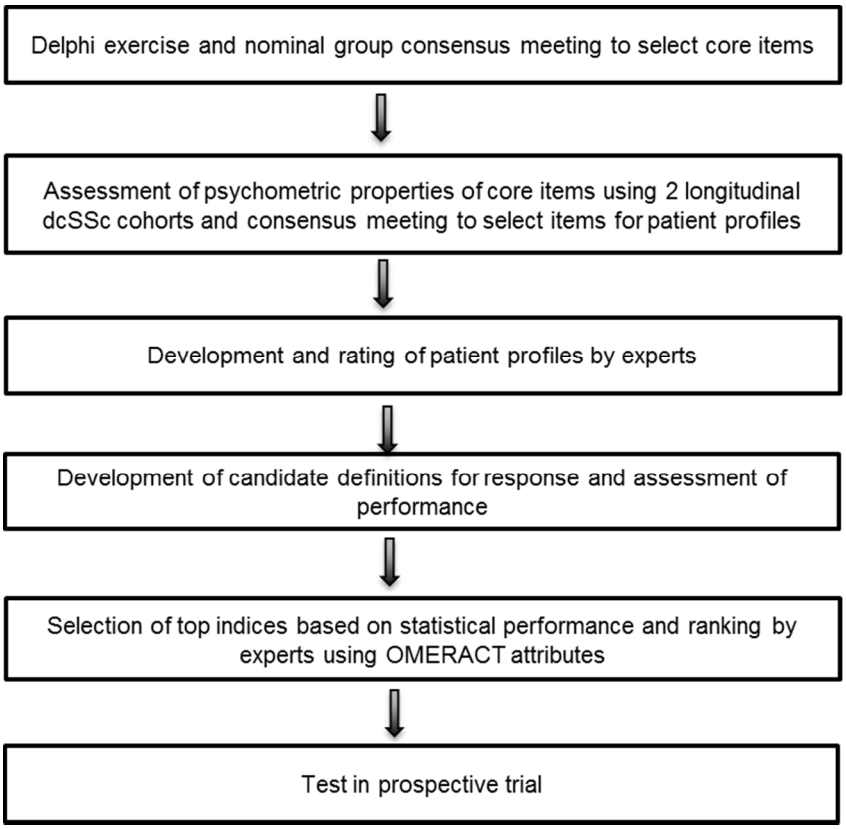
ACR Provisional Composite Index for Scleroderma Clinical Trials

Definition of scleroderma renal crisis [adapted from **(43)]**

| |
|---|
| *A. Hypertensive SRC (fulfills both A1 and A2)* |
| 1. New onset hypertension, defined as any of the following: |
| a) Systolic blood pressure ≥ 140 mgHg<br>b) Diastolic blood pressure ≥ 90 mgHg<br>c) Rise in systolic blood pressure ≥ 30 mmHg<br>d) Rise in diastolic blood pressure ≥ 20 mmHg |
| AND |
| 2. One (1) of the following five (5) features: |
| a) Increase in serum creatinine by 50+% over baseline OR serum creatinine ≥120% of upper limit of normal for local laboratory<br>b) Proteinuria ≥2+ by dipstick<br>c) Hematuria ≥2+ by dipstick or ≥10 RBCs/HPF<br>d) Thrombocytopenia: <100,000 platelets/$mm^3$<br>e) Hemolysis defined as anemia not due to other causes and either of the following:<br>(1) Schistocytes or other RBC fragments seen on blood smear<br>(2) increased reticulocyte count |
| *B. Normotensive SRC (fulfills both B1 and B2)* |
| 1. Increase in serum creatinine >50% over baseline OR serum creatinine ≥120% of upper limit of normal for local laboratory |
| AND |
| 2. One (1) of the following five (5) features: |
| a) Proteinuria ≥2+ by dipstick<br>b) Hematuria ≥2+ by dipstick or ≥10 RBCs/HPF<br>c) Thrombocytopenia: <100,000/$mm^3$<br>d) Hemolysis defined as anemia not due to other causes and either of the following:<br>(1) Schistocytes or other RBC fragments seen on blood smear<br>(2) Increased reticulocyte count<br>e) Renal biopsy findings consistent with scleroderma renal crisis (microangiopathy) |

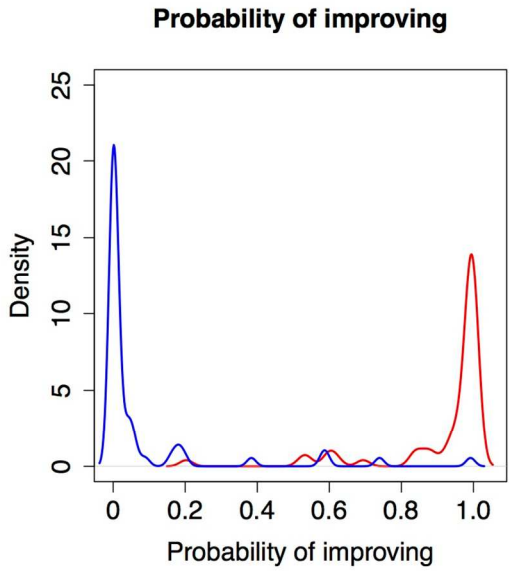
ACR Provisional Composite Index for Scleroderma Clinical Trials

**Figure 1: Expert consensus and data-driven approaches used to develop CRISS**

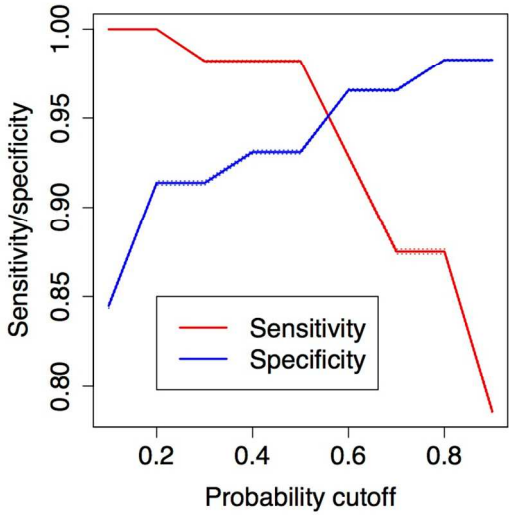


ACR Provisional Composite Index for Scleroderma Clinical Trials






(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 cutoffs considered are 0.1, 0.2, 0.3, … 0.9 and the predicted classifications are derived as follow: if the predicted probability for a subject is greater than the probability cutoff, the subject is rated as “improved”, otherwise subject is not.

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Appendix Table 1. Correlation between the continuous core items among the 14 core items 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=Physician 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

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Appendix Table 2. Correlation between the change scores in the 14 core continuous core items.***

| | 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=Physician 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

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Appendix Table 3. Example of a patient rated "improved" by the experts. Predicted probability of improving is 0.99 according to CRISS.**

| | Baseline | Follow-up | Absolute change |
|---|---|---|---|
| Age | 51.6 years | | |
| Disease duration (months) | 12.98 | | |
| **Global assessments** | | | |
| *Patient global assessment (0-10)** | 3 | 1 | -2 |
| *Physician global assessment (0-10)** | 3 | 3 | 0 |
| **Musculoskeletal** | | | |
| *HAQ-DI (0-3)** | 0.625 | 0 | -0.625 |
| Tendon friction rubs* | No | No | No change |
| **Skin** | | | |
| *MRSS (0-51)** | 13 | 3 | -10 |
| Patient skin interference 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 |

*included in Step 2; ** included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Appendix Table 4. Example of a patient rated "improved" by the experts. Predicted probability of improving is 0.60 according to CRISS.**

| | Baseline | Follow-up | Absolute change |
|---|---|---|---|
| Age | 64.65 years | | |
| Disease duration (months) | 30.74 | | |
| **Global assessments** | | | |
| *Patient global assessment (0-10)** | 1 | 0 | -1 |
| *Physician global assessment (0-10)** | 7 | 4 | -3 |
| **Musculoskeletal** | | | |
| *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 skin interference 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 |

*included in Step 2; ** included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

ACR Provisional Composite Index for Scleroderma Clinical Trials

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

| | Baseline | Follow-up | Absolute Change |
|---|---|---|---|
| Age | 53.6 years | | |
| Disease duration (months) | 43.3 | | |
| **Global assessments** | | | |
| *Patient global assessment (0-10)** | 1 | 2 | 1 |
| *Physician global assessment (0-10)** | 1 | 2 | 1 |
| **Musculoskeletal** | | | |
| *HAQ-DI (0-3)** | 0 | 0 | 0 |
| Tendon friction rubs* | No | Yes | Change to worsen |
| **Skin** | | | |
| *MRSS (0-51)** | 7 | 5 | -2 |
| Patient skin interference 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 |

*included in Step 2; ** included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

ACR Provisional Composite Index for Scleroderma Clinical Trials

**Appendix Table 6. One core item logistic model using expert consensus definition of improved vs. not**

| Core item | 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 |
| Physician 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 skin interference last month | 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 |
| Body mass index | 0.4946 | 0.1786 | 0.8276 | 0.250 |
| Number of digital ulcers | 0.4764 | 0.0179 | 0.931 | 0.249 |

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

ACR Provisional Composite Index for Scleroderma Clinical Trials

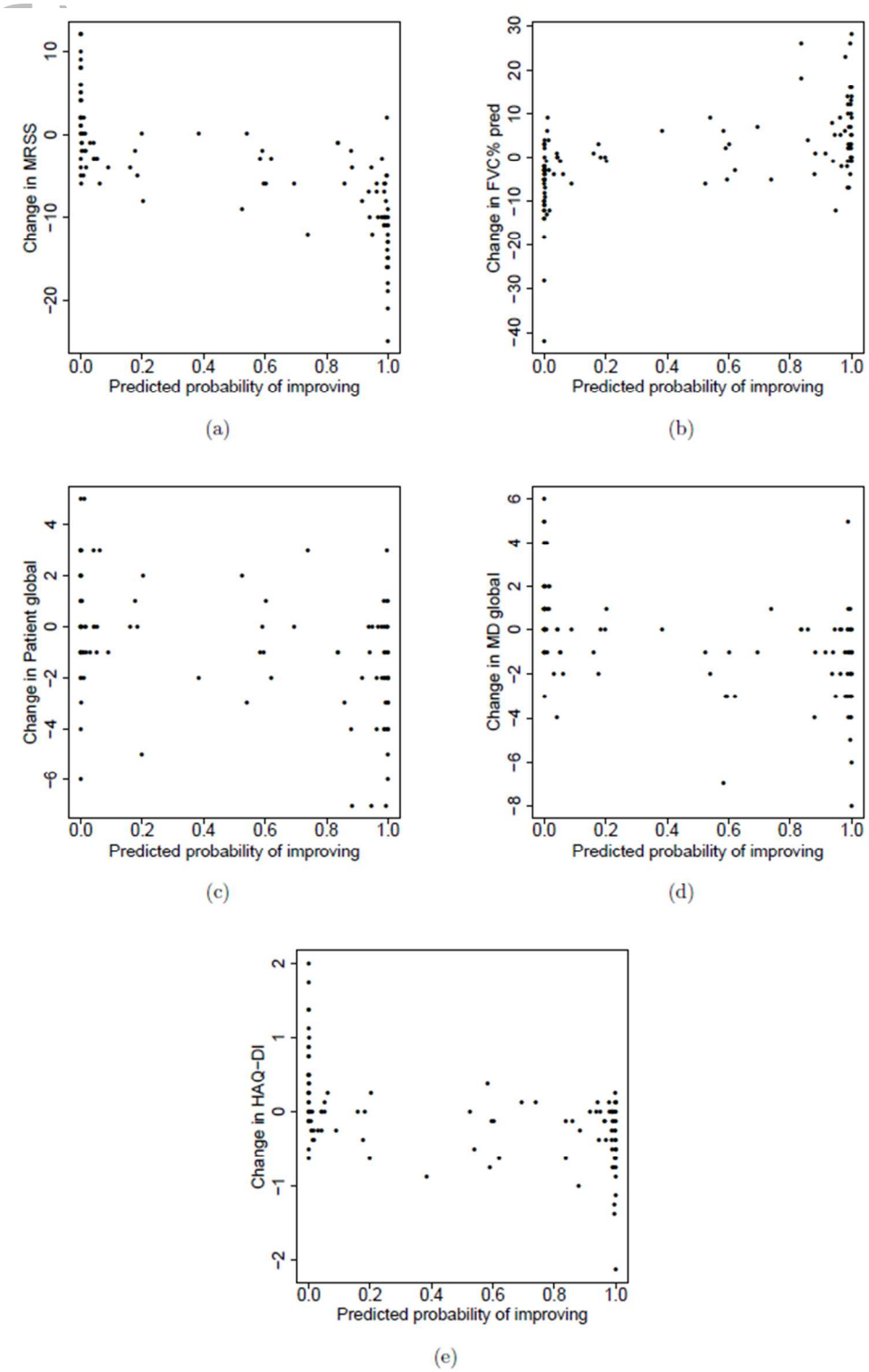
Appendix Table 7. Two core item logistic model using expert consensus definition of improved vs. not

| Core item | 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, physician 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 |

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity

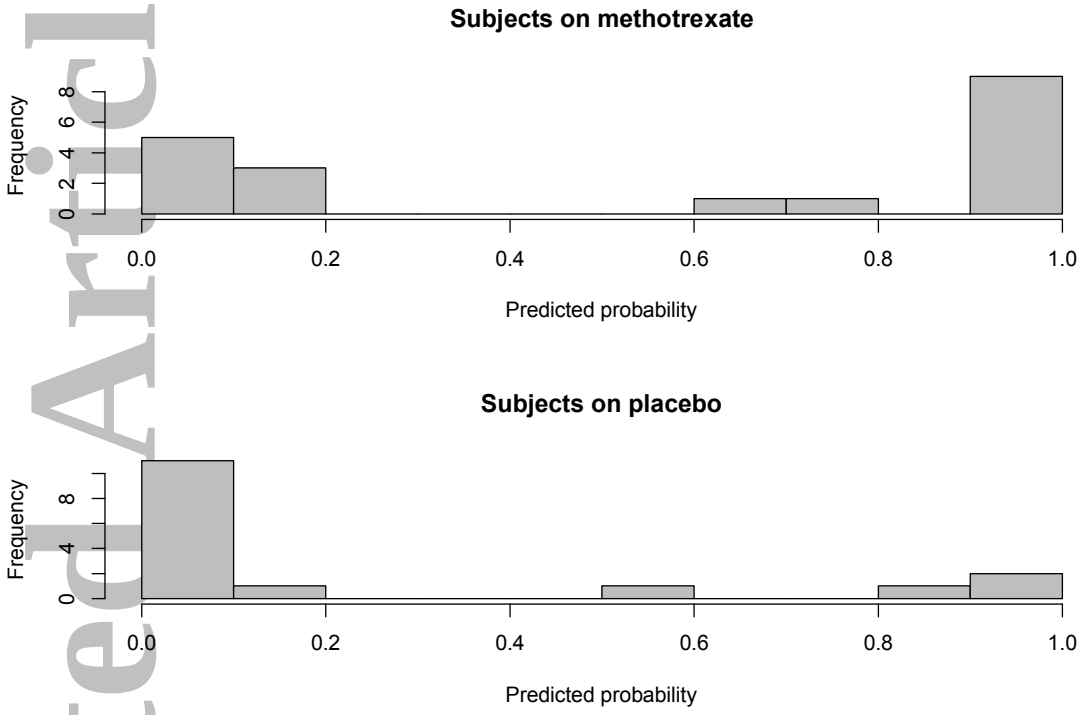
ACR Provisional Composite Index for Scleroderma Clinical Trials

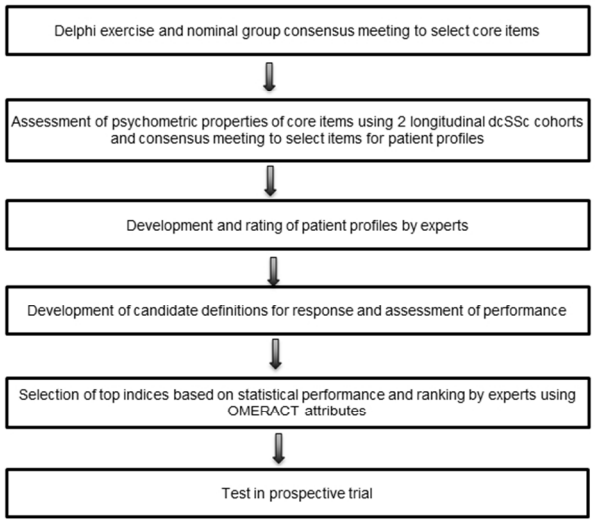
Appendix Figure 1. (a) Change in MRSS, (b) Change in FVC% predicted, (c) Change in patient global assessment, (d) Change in physician global assessment, and (e) Change in HAQ-DI versus the predicted probability of improving yielded by CRISS.



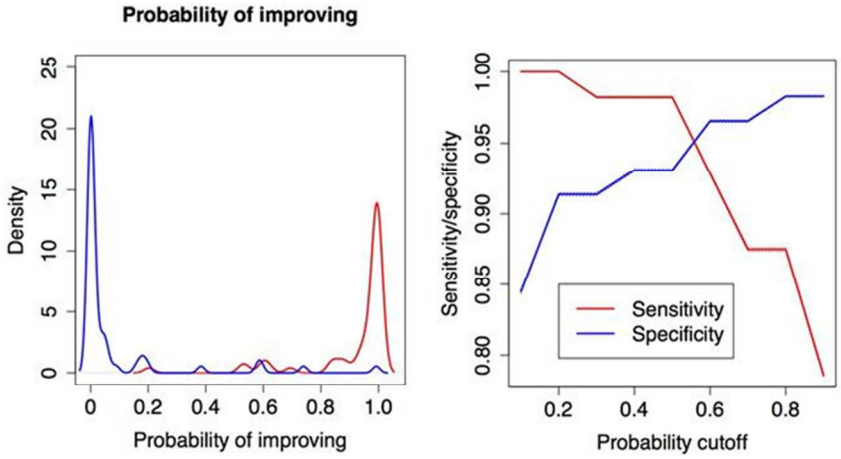
ACR Provisional Composite Index for Scleroderma Clinical Trials

Appendix Figure 2. Histogram of the predicted probabilities of improving in subjects in the RCT study of methotrexate vs. placebo.






254x190mm (96 x 96 DPI)




254x190mm (96 x 96 DPI)

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

## Supplementary material CRISS paper

This Supplementary material is meant to provide more details on the "Patients and Methods" section of the manuscript.

### (i) Structured consensus exercise to develop domains and core set items

This details have been published in the Khanna D, et al. Annals of the Rheumatic diseases 2008; 67(5):703-9 . The original domains and core set items included were

| | |
|---|---|
| | **SKIN** |
| 1 | Modified Rodnan Skin Score (Range 0-51) |
| 2 | Visual analog scale (VAS) or /Likert of **patient** global assessment for skin activity |
| 3 | VAS or /Likert of **physician** global assessment for skin activity |
| 4 | Durometer |
| | **MUSCULOSKELETAL** |
| 1 | Tender Joint Count |
| 2 | Tendon friction rubs assessed by the **physician** |
| | **CARDIAC** |
| 1 | [1]Cardiac Echocardiogram with doppler |
| 2 | [2]Right heart catheterization* |
| 3 | 6-minute walk test* |
| 4 | Borg Dyspnea Instrument * |
| | **PULMONARY** |
| 1 | [3]Pulmonary Function Testing |
| 2 | [4]Validated measure of Dyspnea |
| 3 | Breathing VAS from the Scleroderma Health Assessment Questionnaire (S-HAQ)[42] |
| 4 | High resolution computer tomography (HRCT): quantifiable scale* |
| | **RENAL** |
| 1 | Calculated creatinine clearance based on serum creatinine(Cockroft-Gault or MDRD formula) |
| 2 | Pre-defined renal crisis (Presence or absence) |
| | **GASTROINTESTINAL** |
| 1 | Body mass index (BMI) |
| 4 | Validated gastrointestinal tract VAS scale (part of S-HAQ) [42] or other SSc-Validated GI questionnaire |
| | **HEALTH-RELATED QUALITY OF LIFE AND FUNCTION** |
| 1 | Health Assessment Questionnaire-Disability Index (HAQ-DI)[43] |
| 2 | VAS-pain scale from the HAQ-DI[43] |
| 3 | SF-36 version 2[44] |
| | **GLOBAL HEALTH** |
| 1 | VAS/Likert **patient** global severity |
| 2 | VAS/Likert **physician** global severity |

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

| | |
|---|---|
| 3 | [5]Scleroderma-related health transition by patient |
| 4 | [6]Scleroderma-related health transition by physician |
| **RAYNAUD'S** | |
| 1 | Raynaud's Condition Score[45] |
| 2 | VAS Raynaud's (part of S-HAQ) [42] |
| **DIGITAL ULCERS** | |
| 1 | Active digital tip ulcer count on the volar surface |
| 2 | VAS digital ulcer (part of S-HAQ) [42] |
| **BIOMARKERS** | |
| 1 | Acute phase reactant(s)—Erythrocyte sedimentation rate and/or C-reactive protein |

Standardized central reading mechanism strongly encouraged, * if relevant to the study

Parameters to be measured for the items in the core set

[1]Echocardiogram with Doppler— Measure pulmonary artery systolic pressure, left and right ventricular systolic and diastolic parameters, pericardial effusion, and chamber size
[2]Right heart catheterization— pulmonary artery pressure, pulmonary vascular resistance, wedge pressure, cardiac output/cardiac index
[3]Pulmonary Function Test— Forced Vital Capacity (FVC), Diffusion Capacity (DLCO), Total Lung Capacity (TLC)
[4]Validated dyspnea questionnaires e.g., University of California San Diego (UCSD) Dyspnea Questionnaire, St. George Respiratory Questionnaire, Mahler's Dyspnea Index
[5]Scleroderma-related health transition by patient—Compared to one year ago, how do you rate your overall scleroderma— much better, a little better, no change, a little worse, much worse)
[6]Scleroderma-related health transition by physician (Compared to one year ago, how do you rate your patient's overall scleroderma— much better, a little better, no change, a little worse, much worse)

## (ii) Data collection and evaluation of psychometric properties in a longitudinal observational cohort.

Thirty-one core set items, based on a structured consensus exercise (identified in (i)) were prospectively collected in 200 patients with early diffuse cutaneous systemic sclerosis (dcSSc) at 4 U.S. Scleroderma centers. Core set items included continuous, binary, and categorical variables.
We assessed the feasibility, reliability, face and content validity, redundancy and sensitivity to change of each core set item using the following criteria:

- *Feasibility*: a core set item was defined to be feasible is more than 50% of the subjects had complete data at both baseline and 12-month follow-up.
- *Reliability*: Internal consistency reliability for the SF-36, the St. George's questionnaire and the HAQ-DI questionnaires was evaluated using Cronbach's alpha and found to be acceptable ≥ 0.70.
- *Face and content validity*: This is based on the consensus exercise in step (i).

- *Redundancy*: To determine redundancy in the core set items we computed the correlation coefficients among the 31 core set items at baseline, at the 12-months follow up and in the change from baseline to follow up. Correlations were computed using Spearman or Pearson correlation coefficient depending on the core set item: Spearman correlation was used when one or both core set items were discrete (e.g. either binary or categorical); Pearson correlation was used when core set items were continuous. A correlation coefficient, either Spearman or Pearson, of ≥ 0.80 at either baseline or 12-month follow-up indicated core set items that were redundant.
- *Sensitivity to change*: We determined sensitivity to change for each core set item using two transition questions using a patient and a physician question. At the 1-year follow-up, each subject and clinician was asked to rate the overall scleroderma condition compared to the previous year. Both transition questions used a Likert 1-5 scale with 1 representing "much better" and 5 indicating "much worse". Those who answered "1" or "2" were categorized as "improved" on both transition questions. Analogously, those who scored "3", "4" or "5" were categorized as "not improved" on the transition questions. We subsequently assessed the sensitivity to change for each core set item using two different procedures depending on whether the core set item was a binary vs. a continuous variable. For binary variables, we tested whether there was a statistically different distribution (at predefined p<=0.20) among the patients who changed status from baseline to follow-up (e.g. no organ involvement at baseline to new involvement at follow-up, and vice versa) between the improved and non-improved patients with respect to either transition question using Fisher's exact test. Core set items with significantly different distribution between the improved and non-improved groups with respect to either transition question were classified as sensitive to change.

  For continuous variables, we assessed sensitivity to change using the effect size in the improved and non-improved groups. The effect size was defined as the change in the core set item from baseline to follow-up in each corresponding subgroup (e.g. in the improved and in the non-improved group) divided by the standard deviation at baseline for that subgroup. Continuous core set items for which the effect size was at least 0.20 (defined as small effect size) in the improved group with respect to either transition question was deemed as sensitive to change.

  For both continuous and discrete core set items, sensitivity to change was determined using only the available data. Missing data was not imputed.

  After all these attributes for each core set item were obtained, eight members of the steering committee underwent a modified Nominal Group Technique to determine which of the 31 core set items should be retained for the next step in the algorithm. Ed Giannini moderated this internet-based session using webinar. Specifically, steering committee members were

provided with the above-mentioned information (e.g. correlation values among the core set items, sensitivity to change results for each core set item, etc.) along with summary statistics for each of the 31 core set items. The Steering Committee members reviewed the data and scored each core set item on an ordinal scale (1-4) for the following attributes: feasibility, reliability, and face, content, and construct validity [including sensitivity to change] using the modified content validity index matrix. A score of 4 (highest score) was assigned when the item referred 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 that 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 a score of 1 (lowest score) meant that the attribute should definitely not be used as a core set item. Experts could also assign "not applicable" if they were unfamiliar with an item or different aspects of feasibility, reliability, and validity for the item. Based on each attribute, the experts were asked if the core set item should be considered in the combined response index; score of 3 or 4 with ≥75% consensus was considered supportive of an individual core set item. Other core set items were removed from next step.

As patients were not involved in any step of the CRISS development, pain and vitality score from the SF-36 (to assess fatigue) was added to list of core set items in representation of the patient's pain and fatigue.

At the end of this consensus exercise, 16 core set items were retained for the next step in the algorithm development.

**(iii) Development and ratings of representative patient profiles.**

We developed representative patient profiles by sampling patients from the CRISS cohort and from the Canadian Scleroderma Research Group (CSRG) database. Since the CRISS cohort included only patients with dcSSc, in selecting patients from the CSRG database, we limited ourselves to patients who had dcSSc, disease duration of less than 5 years (from 1$^{st}$ non-Raynaud's sign or symptom) and complete data at baseline and follow-up on all the 15 core set items, excluding "patient skin interference" which was not measured in the CSRG cohort. This identified N=94 patients in the CSRG database.

To assess whether the CSRG database and the CRISS cohort could be considered as two samples from the same population, we tested whether there were significant differences among the CRISS cohort and the subset of patients in the CSRG database that satisfied the CRISS cohort enrollment criteria. We compared the 15 selected core set items in the two populations at baseline and

follow-up and found no significant differences between the two groups of patients.

To impute patient skin interference for the CSRG patients, we used the CRISS cohort and fit a regression model, separately, for baseline and follow-up data, with patient skin interference as the outcome variable and the remaining 15 core set items as the predictors. Using the estimated regression coefficients obtained from the two linear regression models, we predicted patient skin interference for all the CSRG patients at baseline and follow-up, respectively. At baseline, being male and HAQ-DI at baseline were the only significant ($p < 0.05$) predictors of skin interference. At follow-up, the modified Rodnan Skin Score (MRSS), Breathing visual analog score (VAS), GI VAS, Digital ulcers VAS, and HAD-DI were significant predictors of skin interference.

Patient profiles developed using the CRISS cohort were obtained by selecting those patients in the CRISS cohort that had no missing data at baseline and follow-up (N=43). To these 43 patients, we added an additional 13 subjects for at total of 56 patient profiles developed using the CRISS cohort. Of these additional 13 subjects in the CRISS cohort, 5 had no missing data at baseline but had missing data for one core set item at the 12-months follow-up, 3 subjects had no missing data at baseline but missing data for two core set items at the 12-months follow-up, and 5 subjects had no missing data at follow-up but missing data on two core set items at baseline. We chose to use those 5 subjects who had no missing data at follow-up but missing data at baseline on two core set items rather than subjects who had complete data at baseline but only missing one core set item at follow-up (there were 14 subjects with these characteristics in the CRISS cohort) based on the type of core set items for which the data was missing (e.g. discrete vs. continuous, more easily predictable based on the other available data or not). To impute the missing data for the 13 CRISS cohort patients, we used two different strategies depending on the nature of the core set item.

If the core set item for which we had missing data at a given time point (baseline or follow-up) was binary, we imputed the binary missing value by fitting a logistic regression to the binary variable of interest with the remaining core set items at the given time point as predictors. We then predicted the missing values for the subjects with missing data either at baseline or follow-up using the formula for the probability of success in logistic regressions. In cases where the prediction was ambiguous, e.g. the predicted probability of a success for the missing binary core set item was close to 50%, we set the missing value equal to 0 or 1 based on what was the most prevalent outcome in a subset of subjects that (i) had similar demographic information as the patient(s) in consideration, and (ii) had similar values, in the core set items that were found to be significant predictors in the logistic regression, to those of the patient(s) in consideration.

For continuous core set items, we used a similar strategy replacing the logistic regression with a linear regression model. In other words, for each continuous core set item with missing data either at baseline or follow-up in the 13 additional CRISS cohort subjects, we fit a linear regression model separately for baseline and follow-up data where the outcome measure was the continuous core set item with missing data and the covariates were the remaining core set items. Missing values were imputed according to the linear regression model.

The 56 patient profiles created using the CRISS cohort and the 94 patient profiles developed using the CSRG database were randomly assigned for ratings to a panel of 54 scleroderma experts, located in North America (N=29), Europe (N=21) and Australia (N=4), and with different years of experience on SSc management, over 10 years (N=38) versus at most 10 years (N=16). Each expert was asked to rate 20 patient profiles and to answer the following two questions:

*1. Do you think the patient has improved, stabilized, or worsened (or unable to tell) over 1-year?;*

*2. How would you rank the three most important core set items that influenced your decision regarding change or stability?*

Forty experts rated the profiles and 14 did not provide any rating.

We examined the raters' answers in several steps. In the first stage, we examined responses to the first question and we claimed that consensus was reached when at least 75% of the experts provided the same answer for a patient profile (N=107). Patient profiles on which consensus was not reached but where there was a 60-74% agreement among the experts (N=43), were rated again by the Steering Committee, and discussed in detail using the Nominal Group Technique. After discussion, the Steering Committee casted its votes on these patient profiles (N=43) and the initial ratings were tabulated to derive the percentage of agreement among raters. If there was at least 75% agreement, we stated that consensus was reached. Eleven additional patient profiles out of the 43 reevaluated by the Steering Committee, got added to the initial 107 patients where consensus was reached, yielding a total of 118 patient profiles.

Using data relative to the 118 patient profiles, we examined the experts' reply to the following question:
*How would you rank the three most important core set items that influenced your decision regarding change or stability?*

We tabulated how many times each of the core set item was deemed to be the most important, the second most important and the third most important core set item in helping a scleroderma expert determining whether a patient was

improved, worsened or stabilized. We created a 16 x 3 frequency table that reported for each core set item these counts, resulting in sixteen 3-dimensional vectors. To determine whether we could identify groups of core set items that had similar counts for the number of times they were deemed to be the most useful, the second most useful and the third most useful item in influencing a rater decision, we performed a cluster analysis on the sixteen 3-dimensional vectors. Specifically, we assessed whether clusters of core set items existed by running a well-established clustering algorithm, the K-means algorithm, on the 16 x 3 frequency table. Since the K-means algorithm requests that the number of groups in which to cluster the 16 core set items be specified a priori, we determined the number of clusters by running the algorithm with K=1,2,…15 clusters. We selected the appropriate number of clusters in which to group the 16 core set items by looking at the degree of similarity within each cluster, which can be quantified using the within-clusters sum of squares. A lower within-clusters sum of squares is preferred as it indicates that the clusters are rather homogeneous within themselves but they are different from one another. The appropriate number K of clusters was chosen as the smallest number K such that there was not an appreciable difference in the level of homogeneity within clusters if the core set items were grouped in to K clusters vs K+1. The within-clusters sum of squares criteria indicated that the 16 core set item could be appropriately grouped into 2 clusters, each including core set item with a similar degree of helpfulness in rating patient profiles. The first cluster included MRSS, FVC% predicted, patient global assessment, physician global assessment, and HAQ-DI and the second cluster included all the remaining core set items.

**(iv) Development of response definition.**

We developed the response definitions using data on the 118 patient profiles where consensus was reached. We used the raters' consensus decision as the binary outcome variable, with 1 indicating that a patient had been rated as improved at the 1-year follow-up (N=56) and 0 indicating that a patient had not been rated improved (these included N=29 that were considered to be stable and N=33 who were deemed worsened). We used the change in the 16 core set items from baseline to follow-up as the predictors in a suite of logistic regression models, starting from simple logistic regression models with only one predictor to a model including all 16 core set items.

The different logistic regression models were compared based on the area under the curve (AUC), sensitivity and specificity where the gold standard was the consensus-based improved/not improved rating. Additionally, we compared the various logistic regression models based on their respective predictive power, which we quantified using the Brier score, a generalization of the Mean Squared Error metric for the binary core set items. Specifically, if 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:

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

Among all the logistic regression models considered – from models with each core set item individually as a predictor, to models with two core set items, etc. – the model that included change from baseline to follow-up in MRSS, FVC% predicted, physician global assessment, patient global assessment, and HAQ-DI was considered by the steering committee as the model with the best face validity. This model had a sensitivity of 0.9821 (95% CI: [0.9816, 0.9827]), specificity of 0.9310 (95% CI: [0.9300, 0.9321]), an AUC of 0.9861 and a Brier score was 0.038 (lower score indicates better predictive performance).

Having decided the core set items that are part of the response definition, we derived the predicted probability of improvement at 1-year follow up for each patient profile using the logistic regression model with the 5 core set items identified above. We tested whether there was a significant difference in the distribution of the probability of improvement at 1-year follow-up among the improved and not-improved subjects using the Kolmogorov-Smirnov test. The continuous predicted probability of improvement was translated into a binary classification of patients as “improved” and “not improved” using a cutoff *c* for the predicted probability. In other words, a patient will be deemed “improved” at the 1-year follow-up if the predicted probability of improvement for the patient is greater than *c*, otherwise he/she is considered “not-improved”. To establish the cutoff *c* to use to transform the predicted probabilities into an “improved”-“not improved” classification, we derived the specificity and sensitivity of the corresponding binary classification for various values of “*c*”. By plotting the sensitivity and specificity of the classification as a function of the probability threshold “c”, we determined the predicted probability cutoff that yielded the best balance of sensitivity and specificity as 0.6.