Arthritis Care & Research
Vol. 68, No. 2, February 2016, pp 167–178
DOI 10.1002/acr.22804

SPECIAL ARTICLE

The American College of Rheumatology Provisional Composite Response Index for Clinical Trials in Early Diffuse Cutaneous Systemic Sclerosis

DINESH KHANNA,[1] VERONICA J. BERROCAL,[1] EDWARD H. GIANNINI,[2] JAMES R. SEIBOLD,[3] PETER A. MERKEL,[4] MAUREEN D. MAYES,[5] MURRAY BARON,[6] PHILIP J. CLEMENTS,[7] VIRGINIA STEEN,[8] SHERVIN ASSASSI,[5] ELENA SCHIOPU,[1] KRISTINE PHILLIPS,[1] ROBERT W. SIMMS,[9] YANNICK ALLANORE,[10] CHRISTOPHER P. DENTON,[11] OLIVER DISTLER,[12] SINDHU R. JOHNSON,[13] MARCO MATUCCI-CERINIC,[14] JANET E. POPE,[15] SUSANNA M. PROUDMAN,[16] JEFFREY SIEGEL,[17] WENG KEE WONG,[7] ATHOL U. WELLS,[18] AND DANIEL E. FURST[7]

This criteria set has been approved by the American College of Rheumatology (ACR) Board of Directors as Provisional. This signifies that the criteria set has been quantitatively validated using patient data, but it has not undergone validation based on an external data set. All ACR-approved criteria sets are expected to undergo intermittent updates.

The ACR is an independent, professional, medical and scientific society that does not guarantee, warrant, or endorse any commercial product or service.

This article is published simultaneously in the February 2016 issue of *Arthritis & Rheumatology*.

The contents herein are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Supported by the NIH (National Institute of Arthritis and Musculoskeletal and Skin Diseases grant U01-AR-055057). Drs. Khanna and Berrocal's work was supported by the NIH (National Institute of Arthritis and Musculoskeletal and Skin Diseases grant K24-AR-063120). Dr. Johnson's work was supported by the Canadian Institutes of Health Research (Clinician Scientist Award).

[1]Dinesh Khanna, MD, MS, Veronica J. Berrocal, PhD, Elena Schiopu, MD, Kristine Phillips, MD, PhD: University of Michigan, Ann Arbor; [2]Edward H. Giannini, MSc, DrPH: Cincinnati Children's Hospital, Cincinnati, Ohio; [3]James R. Seibold, MD: Scleroderma Research Consultants, Litchfield, Connecticut; [4]Peter A. Merkel, MD, MPH: University of Pennsylvania, Philadelphia; [5]Maureen D. Mayes, MD, MPH, Shervin Assassi, MD, MS: University of Texas Health Science Center at Houston; [6]Murray Baron, MD: Jewish General Hospital and McGill University, Montreal, Quebec, Canada; [7]Philip J. Clements, MD, MPH, Weng Kee Wong, MS, PhD, Daniel E. Furst, MD: University of California, Los Angeles; [8]Virginia Steen, MD: Georgetown University, Washington, DC; [9]Robert W. Simms, MD: Boston University, Boston, Massachusetts; [10]Yannick Allanore, MD, PhD: Paris Descartes University and Cochin Hospital, AP-HP, Paris, France; [11]Christopher P. Denton, MD, PhD, FRCP: Royal Free and University College London Medical School, London, UK; [12]Oliver Distler, MD: University Hospital Zurich, Zurich, Switzerland; [13]Sindhu R. Johnson, MD, PhD: Toronto Western Hospital and University of Toronto, Toronto, Ontario, Canada; [14]Marco Matucci-Cerinic, MD, PhD: Azienda Ospedaliero-Universitaria Careggi (AOUC) and University of Florence, Florence, Italy; [15]Janet E. Pope, MD, MPH, FRCPC: Schulich School of Medicine, Western University, London Campus, and St. Joseph's Health Care, London, Ontario, Canada; [16]Susanna M. Proudman, MBBS, FRACP: Royal Adelaide Hospital and University of Adelaide, Adelaide, South Australia, Australia; [17]Jeffrey Siegel, MD: Genentech/Roche, San Francisco, California; [18]Athol U. Wells, MD: Royal Brompton Hospital, London, UK.

Dr. Khanna has received consulting fees from Bristol-Myers Squibb, Cytori, EMD Serono, Genkyotex, GlaxoSmithKline, Gilead, Medac, and Sanofi-Aventis (less than $10,000 each) and Bayer and Genentech/Roche (more than $10,000 each) and has received research funding from Actelion, Bristol-Myers Squibb, Cytori, Genentech/Roche, Gilead, and United Therapeutics. Dr. Giannini has received consulting fees, speaking fees, and/or honoraria from Pfizer (less than $10,000). Dr. Seibold has received consulting fees from Boehringer-Ingelheim, Biogen Idec, FibroGen, Novartis, Sanofi-Aventis, and Celgene (less than $10,000 each) and Bayer, DART, EMD Serono, InterMune, and Sigma Tau (more than $10,000 each). Dr. Merkel has received consulting fees from Actelion, ChemoCentryx, GlaxoSmithKline, and Sanofi (less than $10,000 each) and has received research funding from Actelion, Bristol-

Objective. **Early diffuse cutaneous systemic sclerosis (dcSSc) is characterized by rapid changes in the skin and internal organs. The objective of this study was to develop a composite response index in dcSSc (CRISS) for use in randomized controlled trials (RCTs).**

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 improved over a period of 1 year. Using the profiles for which raters had reached a consensus on whether the patients were improved versus not improved (79% of the profiles examined), we fit logistic regression models in which the binary outcome referred to whether the patient was improved or not, and the changes in the core set items from baseline to followup 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. In the logistic regression model in which the included core set items were change over 1 year in the modified Rodnan skin thickness score, the forced vital capacity, the patient and physician global assessments, and the Health Assessment Questionnaire disability index, sensitivity was 0.982 (95% confidence interval 0.982–0.983) and specificity was 0.931 (95% confidence interval 0.930–0.932), and the model with these 5 items had the highest face validity. Subjects with a significant worsening of renal or cardiopulmonary involvement were classified as not improved, regardless of improvements in other core items. With use of the index, the effect of methotrexate could be differentiated from the effect of placebo in a 1-year RCT ($P = 0.02$).**

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

INTRODUCTION

Systemic sclerosis (SSc; scleroderma) 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 thickness score (MRSS) (10) 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 (11–13), 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. We therefore undertook the present work to develop a composite response index in dcSSc (CRISS) for use in clinical trials.

Myers Squibb, Celgene, GlaxoSmithKline, and Genentech/Roche. Dr. Mayes has received consulting fees, speaking fees, and/or honoraria from Medtelligence and Cytori (less than $10,000 each). Dr. Steen has received consulting fees from Bayer, Bristol-Myers Squibb, Cytori, and Gilead (less than $10,000 each) and has received research funding from Actelion, Bayer, Celgene, CSL Behring, Cytori, Genentech/Roche, Gilead, InterMune, Sanofi-Aventis/Genzyme, and United Therapeutics. Dr. Simms has received speaking fees from Gilead (less than $10,000) and consulting fees from Actelion and Cytori (less than $10,000 each) and has received research funding from Actelion, Gilead, Medimmune, and InterMune. Dr. Allanore has received consulting fees from Actelion, Bayer, Behring, Biogen Idec, Bristol-Myers Squibb, Genentech/Roche, Inventiva, Medac, Pfizer, Sanofi/Genzyme, Servier, and UCB (less than $10,000 each) and has received research funding from Bristol-Myers Squibb, Genentech/Roche, Inventiva, Pfizer, Sanofi-Genzyme, and Servier. Dr. Denton has received consulting fees from Actelion, Bayer, GlaxoSmithKline, and Roche (less than $10,000 each) and has received research funding from Actelion, Bayer, GlaxoSmithKline, Roche, Genentech/Roche, Pfizer, GlaxoSmithKline, Bristol-Myers Squibb, CSL Behring, Novartis, Sanofi-Aventis, Inventiva, and Biogen Idec. Dr. Distler has received consulting fees from Ergonex, United BioSource, Biovitrium, Novartis, Biogen Idec, and Inventiva (less than $10,000 each) and has received research funding from Actelion, Pfizer, Sanofi-Aventis, and Bayer; he holds a patent for the use of microRNA-29 in the treatment of systemic sclerosis. Dr. Matucci-Cerinic has received consulting fees from GlaxoSmithKline, Actelion, Bristol-Myers Squibb, MSD, and Pfizer (less than $10,000 each) and has received research funding from GlaxoSmithKline, Actelion, Bayer, Behring, Bristol-Myers Squibb, MSD, Pfizer, and UCB. Dr. Proudman has received consulting fees from Actelion (more than $10,000) and has received research funding from Actelion, Bayer, and GlaxoSmithKline. Dr. Siegel owns stock or stock options in Roche. Dr. Furst has received consulting fees from AbbVie, Actelion, Amgen, Bristol-Myers Squibb, Cytori, Janssen, Gilead, GlaxoSmithKline, Novartis, Pfizer, Genentech/Roche, and UCB (less than $10,000 each) and honoraria for CME programs from AbbVie, Actelion, and UCB (less than $10,000 each) and has received research funding from AbbVie, Actelion, Amgen, Bristol-Myers Squibb, Gilead, GlaxoSmithKline, Novartis, Pfizer, Genentech/Roche, and UCB.

Address correspondence to Dinesh Khanna, MD, MSc, University of Michigan, Scleroderma Program, Division of Rheumatology, Department of Internal Medicine, Suite 7C27, 300 North Ingalls Street, SPC 5422, Ann Arbor, MI 48109. E-mail: khannad@med.umich.edu.

Submitted for publication November 11, 2014; accepted in revised form October 30, 2015.

PATIENTS AND METHODS

The index was developed using well-accepted expert consensus (14) and data-driven approaches (Figure 1), including the American College of Rheumatology (ACR) standards for the development of response criteria (15). Details are included in Supplementary Patients and Methods, on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract. The basic process was as follows: 1) We conducted a consensus exercise to select domains and outcome measures (core set of items, referred to below as "core items") for potential inclusion in the composite response index. 2) We then tested the psychometric properties of the core items in a longitudinal cohort of patients followed up for 1 year to assess the items' feasibility, reliability, validity, and sensitivity to change. 3) 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. 4) We performed statistical reduction of the data to the minimum number of domains and core items that retained the maximally responsive index and was acceptable to the experts (face validity). 5) We then tested the ability of the composite response index to discriminate among therapies using results from a previously published randomized controlled trial (RCT). Each of these steps is described in greater detail below.

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; details of the exercise 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). In round 1 the SCTC members were asked to list items in 11 predefined domains, and in round 2 respondents were asked to rate the importance of the chosen items on a 1–9 ordinal scale. This was followed by a face-to-face meeting where, with expert facilitators, consensus about which domains and core items to test in a database (5) was reached, using the nominal group technique (14). During this exercise, the steering committee discussed the feasibility, reliability, redundancy, and validity of the items.

Data collection and evaluation of psychometric properties in a longitudinal observational cohort. Due to a lack of dcSSc trials with positive findings and as a consequence of the fact that previous trials did not include some of the core items chosen in the consensus exercise (16), we assembled a longitudinal observational cohort of patients with early dcSSc (<5 years from first non–Raynaud's phenomenon sign or symptom) at 4 US scleroderma centers (the CRISS cohort) (17). The observational cohort, recruited over a 12-month period, included

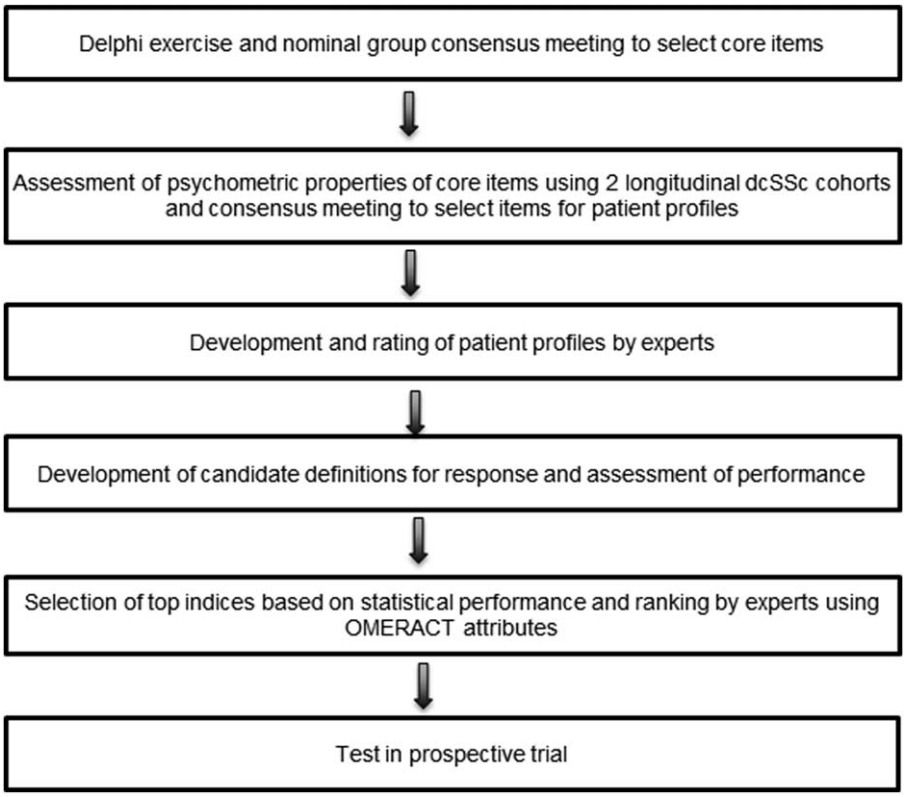

Figure 1. Expert consensus and data-driven approaches used to develop the composite response index in systemic sclerosis (CRISS). dcSSc = diffuse cutaneous systemic sclerosis; OMERACT = Outcome Measures in Rheumatology.

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 up for 12 months, and features were recorded at baseline and 12 months. Exclusion criteria included life expectancy of <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]) (18). Feasibility was defined as completion of the core set item by >50% of subjects at 2 time points, and redundancy was defined as either a Spearman or Pearson correlation coefficient of at least 0.80 at baseline or during followup. Sensitivity to change over the 1-year period was calculated using appropriate patient and physician anchor and transition questions. A modified Likert scale (transition health question) was used by physicians and patients at the 1-year followup visit to determine the change in overall condition during the prior year on a scale of 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, patients with a physician-assigned score of "1" or "2" on the transition question were categorized as improved, and those with a physician-assigned score of "3," "4," or "5" as not improved, according to the physician assessment. Similarly, patients with a self-assigned score of "1" or "2" on the transition question were categorized as improved, and those with a self-assigned score of "3," "4," or "5" as not improved, according to the patient assessment. Effect size was calculated using the transition questions as anchors and Cohen's "rule-of-thumb" for interpreting effect size: values of 0.20–0.49 represent a small change, values of 0.50–0.79 a medium change, and values of ≥0.80 a large change (19). Core items that were significant at a predefined *P* value of <0.20 (for dichotomous measures) or that had an effect size of ≥0.20 in the "improved" group (with respect to either patient or physician assessments) were included in the next stage.

Eight steering committee members (DK, JRS, PAM, MDM, MB, PJC, VS, and DEF) reviewed the data and scored each core item on an ordinal scale (from 1 to 4) for feasibility, reliability, and face, content, and construct validity (including sensitivity to change) using the modified content validity index matrix (20). A score of 4 was assigned when the item referred to a value or an attribute that is well established in the literature or through systematically obtained information, a score of 3 indicated a value or an attribute that is somewhat known and accepted but may need minor alteration or modification, a score of 2 indicated that the rater was unable to assess the attribute without additional information or research, and a score of 1 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 with different aspects of feasibility, reliability, and validity for the item. Scores of 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 led by one of the authors (EHG) via webinar, in which 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 webinar, summary statistics were provided for each core set item, and the moderator encouraged discussion of each item by each committee member and then by the group as a whole. This process ensured that all participants had an opportunity to contribute. Subsequently, each item was rescored (if the committee member believed the score should be changed) and summary statistics were generated. Items that were found to lack feasibility, reliability, and validity (<75% of the raters assigning a score of 3 or better) were excluded from the next step.

Development and ratings of representative patient profiles. We developed 150 paper patient profiles using actual data from the CRISS cohort. To have sufficient data on representative patients, we also obtained data on patients with early dcSSc (defined as a disease duration of <5 years) in the Canadian Scleroderma Research Group database (21), a large observational SSc cohort. Since patient interviews were not performed as part of the consensus meeting (step 1), the medical literature was searched to assess the most prevalent/bothersome issues faced by patients with SSc (22–24). Based on this, pain and fatigue (assessed with the Short Form 36 vitality scale) (25) were included as part of the patient profiles.

Fifty-four international experts in scleroderma 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] versus Europe [n = 21] versus Australia [n = 4]) and years of experience with management of SSc (>10 years [n = 38] versus ≤10 years [n = 16]), to prevent systematic bias in rating due to practice patterns. For each patient profile, the rater was asked 3 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 3 most important core items that influenced your decision regarding change or stability? Consensus was considered to have been met if at least 75% of those who rated the same patient profile agreed that the patient had improved, stabilized, or worsened. When there was lack of consensus, 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 then ≥75%, the case was deemed as having reached consensus. This process yielded a final list of 16 core items. Finally, we sought consensus among SSc experts on the level of change in internal organ involvement that should be used to classify a patient as not improved.

Development of response definitions. Using only profiles for which consensus was reached, we fit logistic regression models to the binary outcome measure, i.e., whether a patient had been rated by experts as being improved (recorded as 1) versus not improved (recorded as 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 with the Brier score (26); 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 that were rated improved by the experts and those that were rated not improved. The difference in the 2 distributions was assessed with the nonparametric 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 whom the predicted probability is above the chosen threshold and "not improved" for all patients for whom 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 (i.e., 1) or not improved (i.e., 0). We plotted 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 the degree of their ability 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 (27).

Preliminary evaluation in an independent cohort. The composite index was tested in an RCT of methotrexate versus placebo for the treatment of early dcSSc (28). This trial was chosen because individual patient data were recorded, and all final core items were available in this database. We applied the CRISS to the patients with complete data and for each patient, derived the predicted probability that the individual was improved, using the predicted probability equation (see below). We transformed the continuous predicted probabilities ranging from 0 to 1 into a binary classification, by defining each patient as improved or not improved depending on whether the predicted probability was above the threshold with the highest sensitivity and specificity (identified in step 4). We then tested whether the probability of being improved was independent of methotrexate therapy (i.e., whether the probability of being improved was the same in the methotrexate-treated and the placebo-treated groups), by chi-square testing. We also assessed, by Mann-Whitney test, whether the distributions of the predicted probabilities differed between the patients who received methotrexate and those who received placebo.

RESULTS

Identification of domains and core items via structured consensus exercise. 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 the 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 a 2008 OMERACT conference (4,29) provided input during the consensus exercise.

Characteristics of the longitudinal observational cohort (CRISS cohort) and evaluation of core item psychometric properties in the cohort. Two hundred patients with early dcSSc were recruited at baseline. For 150 of these patients, both baseline and 1-year data were available. The mean ± SD age of the 150 patients at baseline was 50.4 ± 11.7 years, and 74.7% were female. Seventy-eight percent were white and 10.7% were Hispanic. The mean duration of disease from the time of the first non–Raynaud's phenomenon sign or symptom was 2.3 ± 1.5 years, the mean MRSS was 21.4 ± 10.1, the mean forced vital capacity (FVC; % predicted) was 82.3 ± 18.5, and the mean Health Assessment Questionnaire (HAQ) disability index (DI) (30) was 1.0 ± 0.8 (Table 1). Core items that lacked feasibility due to low completion rate (<50%) at 1 year included durometry (a device to measure the skin hardness) (31), right-sided heart catheterization, Borg dyspnea scale (32), 6-minute walk test, and Raynaud's Condition Score (33) (which required daily patient diary records).

When patient global assessment was used as the metric to classify patients as improved versus not improved, 57% were rated as improved and 43% as not improved. Using physician global assessment, 58% were rated as improved and 42% as not improved. The Spearman correlation among the definitions was 0.46, supporting use of 2 global transition questions. Using these transition questions, 6 items were found to be not responsive to change or occurred in <10% of the cohort: tender joint count, presence of renal crisis, estimated glomerular filtration rate, body mass index, presence of digital ulcers, and erythro-

Table 1. Baseline demographic characteristics of the patients in the CRISS cohort with available baseline and 1-year data*

Age (n = 150)	50.4 ± 11.7
Race, no. (%) (n = 150)	
White	117 (78)
African American	13 (9)
Asian	11 (7)
Other or not reported	9 (6)
Ethnicity, no. (%) (n = 150)	
Hispanic	16 (11)
Non-Hispanic	134 (89)
Disease duration, years (n = 144)	1.59 ± 1.34
Years since first RP symptom (n = 128)	2.87 ± 2.49
Years since first non-RP symptom (n = 129)	2.32 ± 1.5
Body mass index, kg/m^2 (n = 96)	26.02 ± 7.1
MRSS (n = 150)	21.4 ± 10.1
Durometry result (n = 113)	272.4 ± 64.5
FVC % predicted (n = 140)	82.32 ± 18.5
Total lung capacity % predicted (n = 109)	87.83 ± 20.4
DLCO % predicted (n = 140)	65.05 ± 20.9
HRCT consistent with ILD, no. (%) (n = 99)	79 (80)
6-minute walking distance, meters (n = 50)	421.6 ± 139.2
Borg scale, 0–10 (n = 46)	1.92 ± 1.51
Tendon friction rubs, no. (%) (n = 140)	40 (29)
Small joint contractures, no. (%) (n = 133)	78 (59)
Large joint contractures, no. (%) (n = 133)	39 (29)
Digital ulcers, no. (%) (n = 150)	15 (10)
HAQ DI (n = 150)	1.0 ± 0.8
Patient assessment of digital ulcers, 0–150 VAS (n = 134)	20.9 ± 40.9
Patient assessment of RP, 0–150 VAS (n = 135)	32.7 ± 40.8
Patient assessment of breathing, 0–150 VAS (n = 138)	23.1 ± 36.7
Patient assessment of GI condition, 0–150 VAS (n = 136)	22.6 ± 34.4
Patient assessment of disease severity, 0–150 VAS (n = 138)	56.4 ± 42.9
Pain, 0–10 VAS (n = 140)	4.0 ± 2.8
SF-36 PCS (n = 138)	37.6 ± 12.9
SF-36 MCS (n = 138)	44.2 ± 6.0
Physician global assessment, 0–10 VAS (n = 143)	4.4 ± 2.2
Antinuclear antibody positive, no. (%)	94 (81)
Anti–Scl-70 positive, no. (%)	34 (30)
Serum CPK, IU/liter	143.9 ± 184.5
Serum platelets, ×1,000/μl	315.2 ± 102.5
Serum brain natriuretic peptide, pg/ml	161.3 ± 824.0
ESR, mm/hour	23.4 ± 22.6
Serum CRP, mg/dl	2.1 ± 4.9

* Except where indicated otherwise, values are the mean ± SD; n values are the number of patients with baseline data included in the table. CRISS = composite response index in diffuse cutaneous systemic sclerosis; RP = Raynaud's phenomenon; MRSS = modified Rodnan skin thickness score; FVC = forced vital capacity; DLCO = diffusing capacity for carbon monoxide; HRCT = high-resolution computed tomography; ILD = interstitial lung disease; HAQ DI = Health Assessment Questionnaire disability index; VAS = visual analog scale; GI = gastrointestinal; SF-36 = Short Form 36; PCS = physical component summary; MCS = mental component summary; CPK = creatine phosphokinase; ESR = erythrocyte sedimentation rate; CRP = C-reactive protein.

cyte sedimentation rate. A modified nominal group review was performed, in which consensus was achieved on 16 core items that should be used for the development of paper patients. It was decided to retain renal crisis and presence/absence of digital ulcers as core items due to their impact on prognosis in early dcSSc. No redundancy in the core items was noted at baseline or in the change scores, as assessed using correlation coefficients (Supplementary Tables 1 and 2, on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract).

Rating of paper patients as improved, worsened, or stable over time, and ranking of core items used in making this assessment. 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 Supplementary Tables 3–5, http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract). The median number of experts who rated a profile was 6 (range 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

Table 2. Predictive characteristics of the final CRISS model consisting of the 5 core items with the highest face validity*

Overall area under the curve	0.9861
Overall sensitivity (95% CI)	0.9821 (0.9816–0.9827)
Overall specificity (95% CI)	0.9310 (0.9300–0.9321)
Unadjusted beta coefficient (by core item)	
MRSS	−0.81
FVC % predicted	0.21
HAQ DI	−0.40
Patient global assessment	−0.44
Physician global assessment	−3.41
Standard error (by core item)	
MRSS	0.21
FVC % predicted	0.08
HAQ DI	0.24
Patient global assessment	0.26
Physician global assessment	1.75

* CRISS = composite response index in diffuse cutaneous systemic sclerosis; 95% CI = 95% confidence interval; MRSS = modified Rodnan skin thickness score; FVC = forced vital capacity; HAQ DI = Health Assessment Questionnaire disability index.

time, followed by FVC % predicted (14.5%), patient global assessment (11.0%), physician global assessment (9.1%), and HAQ DI (8.0%). All other core items were ranked as most influential in the decision making <2% of the time.

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

Results of modeling of changes in core items to develop response definitions. *Logistic regression models.* The 118 patient profiles on which consensus was reached were used in the statistical models to examine response definitions regarding improvement based on change in the 16 core items. In 1−core item models (in which only 1 covariate was included), the AUC ranged from 0.48 (for the model including as the single covariate the change in presence/absence of new digital ulcers) to 0.92 (for the model including as the single covariate the change in MRSS) (Supplementary Table 6, http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract). In a 2−core item model, change in MRSS and change in FVC % predicted yielded the highest AUC (0.96) (Supplementary Table 7, http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract) 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 AUCs, sensitivity, and specificity were discussed by the steering committee (data available upon request 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). A clustering analysis, performed to assess whether core items clustered in groups with similar characteristics with respect to usefulness in inferring a patient's 1-year followup status, supported a 5–core item model with the following 5 items: MRSS, FVC % predicted, patient global assessment, physician global assessment, and HAQ DI, all belonging to the same cluster. The remaining core items all belonged to a second cluster (Table 3). The 5–core item model with MRSS, FVC % predicted, patient global assessment, physician global assessment, and HAQ DI as predictors had a sensitivity of 0.9821 (95% confidence interval [95% CI] 0.9816–0.9827), a specificity of 0.9310 (95% CI 0.9300–0.9321), and an AUC of 0.9861. The Brier score was 0.038 (lower score indicates better predictive performance). As the data were

Table 3. Ranking of the 16 core items by scleroderma experts, and results of the cluster analysis

Core item*	Rank 1, no. (%)†	Rank 2, no. (%)†	Rank 3, no. (%)†	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
SF-36 vitality scale	12 (1.4)	37 (4.4)	101 (11.9)	2
Patient GI assessment (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
Patient breathing assessment (VAS)	13 (1.5)	25 (3)	32 (3.8)	2
Patient digital ulcers assessment (VAS)	7 (0.8)	38 (4.5)	17 (2)	2
Patient RP assessment (VAS)	11 (1.3)	18 (2.1)	43 (5.1)	2
Patient-reported skin interference with activities in 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 thickness score; FVC = forced vital capacity; HAQ DI = Health Assessment Questionnaire disability index; SF-36 = Short Form 36; GI = gastrointestinal; VAS = visual analog scale; RP = Raynaud's phenomenon.
† The number is the number of times the item was assigned the given rank.

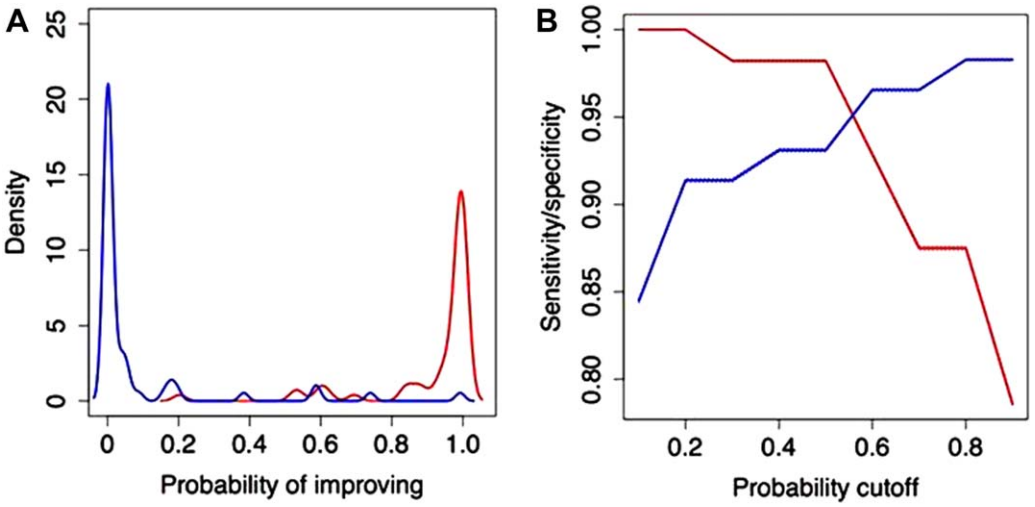

Figure 2. A, Distribution of the predicted probability of improving among patients rated by the experts as improved (red curve) and patients rated by the experts as not improved (blue curve). **B,** Sensitivity and specificity of the predicted classification of patients as improved or not improved as a function of the predicted probability cutoff. The cutoffs considered were 0.1, 0.2, 0.3, … 0.9, and the predicted classifications were derived as follows: if the predicted probability for a patient is greater than the probability cutoff, the patient is rated as improved; otherwise, the patient is rated as not improved.

not normally distributed, nonparametric tests were used to assess whether the distributions of the predicted probability of improving were different between subjects who improved and those who did not (Figure 2A). The distributions of predicted improvement probability were found to differ significantly ($P < 0.0001$). Using depiction of sensitivity versus specificity for identifying the improved group versus the not improved group, a threshold of 0.6 was found to have 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 an individual has equal odds of improving as not improving (i.e., predicted probability of 0.5 or 50%) while a positive (negative) log odds means that an individual 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 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, the MRSS explained 66.3% of the variation, the FVC % predicted explained 36.1%, the physician global assessment explained 24.5%, the patient global assessment explained 23.7%, and the HAQ DI explained 28.5%.

We assessed how changes in the core items were related to the predicted probability of improvement for each patient profile. The changes (from baseline to 12 months) in the MRSS, FVC % predicted, patient global assessment, physician global assessment, and HAQ DI versus the predicted probabilities for the 118 patient profiles are depicted in Supplementary Figure 1, on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract. Changes in the MRSS, FVC, and HAQ DI were strong indicators of whether a patient was likely to be improved. In each scenario, a decrease in the MRSS or HAQ DI from baseline to followup and an increase in the FVC % predicted corresponded to very high probabilities of improving. For patient and physician global assessments, the association between probability of improving and change in these 2 core components was less evident.

Defining a patient as not improved irrespective of improvement in core items. The steering committee considered circumstances in which a patient may improve in a particular outcome measure (such as MRSS or FVC) 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 in a clinical trial, such patients should be defined as not improved. The steering committee voted and determined that the following items met this definition: new onset of renal crisis, new onset or worsening of lung fibrosis, new onset of pulmonary arterial hypertension, or new onset of left ventricular failure (Figure 3). The international experts subsequently endorsed these definitions as well.

Preliminary evaluation in a randomized controlled clinical trial. We used the individual patient data from a clinical trial that compared treatment of dcSSc with methotrexate versus placebo (28) to assess our definition of response. Data on change in MRSS, FVC % predicted, patient global assessment, physician global assessment, and HAQ DI were 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 as improved or not improved using a probability cutoff of 0.6 (determined

CRISS is a 2-step process.

Step 1: Subjects who develop new or worsening 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 (47)
- Decline in FVC % predicted $\geq$15% (relative), confirmed by another FVC test within a month, HRCT to confirm ILD (if previous HRCT of chest did not show ILD) and FVC <80% of predicted*
- New onset of left ventricular failure (defined as left ventricular ejection fraction $\leq$45%) requiring treatment*
- New onset of PAH on right-sided heart catheterization (48) requiring treatment*

* 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 Δ_{MRSS} indicates the change in MRSS from baseline to followup, $\Delta_{\text{FVC\%}}$ denotes the change in FVC % predicted from baseline to followup, $\Delta_{\text{Pt-glob}}$ indicates the change in patient global assessment, $\Delta_{\text{MD-glob}}$ denotes the change in physician global assessment, and $\Delta_{\text{HAQ-DI}}$ is the change in HAQ-DI. All changes are absolute change ($\text{Time}_2 - \text{Time}_{\text{baseline}}$).

Figure 3. Application of the composite response index in diffuse cutaneous systemic sclerosis (CRISS) in a clinical trial. Scleroderma renal crisis is defined as shown in Supplementary Table 8 (on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract). FVC = forced vital capacity; HRCT = high-resolution computed tomography; ILD = interstitial lung disease; PAH = pulmonary arterial hypertension (defined as mean pulmonary artery pressure $\geq$25 mm Hg at rest and end-expiratory pulmonary artery wedge pressure $\leq$15 mm Hg and pulmonary vascular resistance >3 Wood units); MRSS = modified Rodnan skin thickness score; HAQ DI = Health Assessment Questionnaire disability index.

analytically in step 4). With this criterion, 11 of 19 patients who received methotrexate were rated as improved, whereas 3 of 16 patients in the placebo group were rated as improved ($P = 0.04$) (Supplementary Figure 2, http://onlinelibrary.wiley.com/doi/10.1002/acr.22804/abstract). When the data were assessed as a continuous measure, the distribution of the predicted probability for improvement was significantly different between the placebo and methotrexate groups ($P = 0.02$).

Application in a clinical trial. The CRISS was developed with the goal of summarizing changes in clinical and patient-reported outcomes in a single composite score that conveys the likelihood (or probability) that a patient with dcSSc has improved. If there is an effective agent for treatment of dcSSc, the assumption is that a patient treated with the agent will have a higher probability of improvement as summarized by the CRISS versus a patient treated with placebo or an ineffective agent. The CRISS is a 2-step process for use in a clinical trial and is described in Figure 3. In step 1, patients who develop new onset of renal crisis, new onset or worsening of 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 patients with complete data, step 2 involves computing the predicted probability of improving for each individual, using the equation shown in Figure 3. Subjects for whom the predicted probability is $\geq$0.60 are considered improved, while subjects for whom the predicted probability is <0.60 are considered not improved. The 2 groups (study drug versus placebo or active comparator) can then be compared in a 2 $\times$ 2 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 receiving study drug versus placebo can be compared using nonparametric tests.

The CRISS was developed using data from 12 months of treatment. Therefore, with regard to trials that incorporate components of the CRISS at multiple time points, there is a lack of data to support its performance at earlier time periods. We recommend using 12-month findings as primary/secondary outcome measures and using data from other time points, such as baseline to 3, 6, and/or 9 months, as exploratory outcomes. We recommend capturing the data during each patient visit, using specific case report forms for organ involvement. We also encourage inclusion of an adjudication committee that can help with validating the occurrence of cardiopulmonary or renal involvement. If case report forms are not developed and included in the trial, this information should be captured as part of the accounting of adverse events (all of these occurrences should be classified as serious adverse events). Nonavailability of these data on specific case report forms (i.e., if such forms were not developed prospectively for use in the trial) should not be taken as missing data as, again, these occurrences should be captured as serious adverse events. If there are 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 through month 12 before calculating the score.

DISCUSSION

We have developed a composite response index for trials of early dcSSc (the CRISS) using well-established consensus and data-driven approaches. The CRISS includes core items that assess change in 2 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 worsening of internal organ involvement requiring treatment, that classifies the patient as having not improved (regardless of changes in other parameters) during the clinical trial. We subsequently tested the CRISS using data from a clinical trial and, using this index, identified different probabilities of improvement among methotrexate-treated versus placebo-treated patients with early dcSSc. The findings of this analysis suggested that methotrexate has the potential to improve the overall health of patients with dcSSc after 1 year of treatment.

Traditionally, trials in early dcSSc have focused on skin or lung involvement (34,35). The MRSS has been used as the primary outcome measure in the trials of skin fibrosis (6). It meets the OMERACT criteria as a fully validated measure of outcome (36), but is also a surrogate for internal organ involvement and mortality in early dcSSc (37,38). However, clinical trials in dcSSc to date have largely yielded negative results, and the MRSS has been questioned as a primary outcome measure when post hoc analysis of "negative" trials has shown stability/improvement in the MRSS over time (15,39). The CRISS incorporates multisystem involvement in dcSSc and includes the patient perspective and the impact of the disease on functional disability. It is calculated as a 2-step process (Figure 3). The first step evaluates clinically significant worsening of renal or cardiopulmonary involvement that requires treatment; if this is present, the patient is classified as not improved. The definitions chosen for internal organ involvement were based on published data and expert opinion regarding involvement that is 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 the use of a data-driven definition as suggested by the ACR Criteria Subcommittee (14). In addition, data-driven definitions of disease activity have been successfully used for regulatory approval in other rheumatic diseases (40,41).

The purpose of the CRISS is to assess whether new pharmacologic agents have an impact on overall disease activity/severity. Our hope is that its use in clinical trials of dcSSc will greatly facilitate the interpretation of results and form the basis for drug approvals. Rather than using numerous outcome measures that vary from trial to trial, the core set of items used in the CRISS will produce a single efficacy measure. This process will lessen the ambiguity associated with presentation of multiple test statistics, some of which may be significant and others not, and facilitate meta-analyses. It will likely also allow a reduction in the number of patients needed for appropriately powered clinical trials, as has been the case with other composite indices in rheumatoid arthritis. It should be noted that use of the 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 the 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, similar to the comprehensive structure of the British Isles Lupus Assessment Group and Systemic Lupus Erythematosus Disease Activity Index measures used in trials of systemic lupus erythematosus (42,43). 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 that was selected.

Other indices for SSc have been described. The European Scleroderma Study Group (44) 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 (45), a measure that encompasses disease activity and damage, has been proposed and can be used in trials to complement the 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 for use in patients with 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 able to test the index in only a single, small RCT in which a substantial number of patients were lost to followup; therefore, further validation of the CRISS in a prospective RCT of adequate size is needed.

The study is also 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 the MRSS. The majority of past and ongoing 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 limi-

tation and searched the literature for patient input regarding scleroderma (22,23); 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, 2 of the constituent core items of the CRISS include patient global assessment and patient-reported functional assessment.

We also note that the CRISS should be considered as a preliminary index. Although it was tested in an RCT, missing data in that trial (>50%) precludes definitive conclusions, and the CRISS may need to be revised as more data from future trials become available. We had 118 paper patient profiles for which there was expert consensus, and these profiles were used to develop different response definitions. Although this is standard methodology, it may be suboptimal for testing 16 core set items. This may also explain the high AUC of 0.986 for the index.

Last, 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% improvement criteria for rheumatoid arthritis (13) or the ACR 30% improvement criteria for juvenile idiopathic arthritis (46) also address only changes in core items, and not baseline values. Although baseline scores can influence the change 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.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Khanna had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Khanna, Berrocal, Giannini, Seibold, Merkel, Clements, Phillips, Simms, Denton, Johnson, Matucci-Cerinic, Pope, Siegel, Wong.

Acquisition of data. Khanna, Seibold, Merkel, Mayes, Baron, Clements, Steen, Assassi, Schiopu, Phillips, Simms, Denton, Johnson, Matucci-Cerinic, Pope, Siegel.

Analysis and interpretation of data. Khanna, Berrocal, Giannini, Seibold, Merkel, Baron, Clements, Steen, Assassi, Phillips, Simms, Allanore, Denton, Distler, Johnson, Matucci-Cerinic, Pope, Proudman, Siegel, Wong, Wells, Furst.

ADDITIONAL DISCLOSURES

Author Siegel is an employee of Genentech/Roche.

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;118:2–10.
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;51:1017–26.
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 Suppl 65:33–9.
4. Khanna D, Distler O, Avouac J, Behrens F, Clements PJ, Denton C, et al, for the Investigators in CRISS and EPOSS. 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;36:2356–61.
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;67:703–9.
6. Khanna D, Merkel PA. Outcome measures in systemic sclerosis: an update on instruments and current research. Curr Rheumatol Rep 2007;9:151–7.
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 Suppl 71:97–102.
8. Abraham DJ, Varga J. Scleroderma: from cell and molecular mechanisms to disease models. Trends Immunol 2005;26:587–95.
9. Nagaraja V, Denton CP, Khanna D. Old medications and new targeted therapies in systemic sclerosis. Rheumatology (Oxford) 2015;54:1944–53.
10. Clements P, Lachenbruch P, Seibold J, White B, Weiner S, Martin R, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol 1995;22:1281–5.
11. Van der Heijde DM, van '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;49:916–20.
12. Paulus HE, Egger MJ, Ward JR, Williams HJ, and the Cooperative Systematic Studies of Rheumatic Diseases Group. Analysis of improvement in individual rheumatoid arthritis patients treated with disease-modifying antirheumatic drugs, based on the findings in patients treated with placebo. Arthritis Rheum 1990;33:477–84.
13. 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;38:727–35.
14. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011;41:95–105.
15. Classification and Response Criteria Subcommittee of the American College of Rheumatology Committee on Quality Measures. Development of classification and response criteria for rheumatic diseases. Arthritis Rheum 2006;55:348–52.
16. Merkel PA, Silliman NP, Clements PJ, Denton CP, Furst DE, Mayes MD, et al, for the Scleroderma Clinical Trials Consortium. 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;64:3420–9.
17. 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;66:1731–9.

18. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73–5.
19. Cohen J. The analysis of variance and covariance. In: Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988. p. 273–406.
20. 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: 768–82.
21. Fan X, Pope J, the Canadian Scleroderma Research Group, 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;30:1205–10.
22. 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: 762–67.
23. Suarez-Almazor ME, Kallen MA, Roundtree AK, Mayes M. Disease and symptom burden in systemic sclerosis: a patient perspective. J Rheumatol 2007;34:1718–26.
24. 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: 1074–9.
25. Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute, New England Medical Center; 1993.
26. Gneiting T, Raftery A. Strictly proper scoring rules. J Am Stat Assoc 2007;102:359–78.
27. Nagelkerke NG. A note on a general definition of the coefficient of determination. Biometrika 1991;78:691–2.
28. 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;44:1351–8.
29. 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;34:1194–200.
30. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
31. Merkel PA, Silliman NP, Denton CP, Furst DE, Khanna D, Emery P, et al, for the CAT-192 Research Group and the Scleroderma Clinical Trials Consortium. Validity, reliability, and feasibility of durometer measurements of scleroderma skin disease in a multicenter treatment trial. Arthritis Rheum 2008;59:699–705.
32. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. ATS statement: guidelines for the six-minute walk test. Am J Respir Crit Care Med 2002;166:111–7.
33. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al, for the Scleroderma Clinical Trials Consortium. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410–20.
34. Khanna D, Clements PJ, Furst DE, Korn JH, Ellman M, Rothfield N, et al, for the Relaxin Investigators and the Scleroderma Clinical Trials Consortium. Recombinant human relaxin in the treatment of systemic sclerosis with diffuse cutaneous involvement: a randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2009;60:1102–11.
35. 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;354:2655–66.
36. 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;30:1630–47.
37. 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;43:2445–54.
38. Steen VD, Medsger TA Jr. Severe organ involvement in systemic sclerosis with diffuse scleroderma. Arthritis Rheum 2000;43:2437–44.
39. Amjadi S, Maranian P, Furst DE, Clements PJ, Wong WK, Postlethwaite AE, et al, for the Investigators of the D-Penicillamine, Human Recombinant Relaxin, and Oral Bovine Type I Collagen Clinical Trials. 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: 2490–8.
40. 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; 51:177–81.
41. Luijten KM, Tekstra J, Bijlsma JW, Bijl M. The Systemic Lupus Erythematosus Responder Index (SRI); a new SLE disease activity assessment. Autoimmun Rev 2012;11:326–9.
42. Symmons DP, Coppock JS, Bacon PA, Bresnihan B, Isenberg DA, Maddison P, et al, and Members of the British Isles Lupus Assessment Group (BILAG). Development and assessment of a computerized index of clinical disease activity in systemic lupus erythematosus. Q J Med 1988;69:927–37.
43. Bombardier C, Gladman DD, Urowitz MB, Caron D, Chang DH, and the Committee on Prognosis Studies in SLE. Derivation of the SLEDAI: a disease activity index for lupus patients. Arthritis Rheum 1992;35:630–40.
44. 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;62:904–5.
45. 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;26:2159–67.
46. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997;40:1202–9.
47. Steen VD, Mayes MD, Merkel PA. Assessment of kidney involvement. Clin Exp Rheumatol 2003;21 Suppl 29:29–31.
48. 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;62:42–50.