Summed and Weighted Summary Scores for the Medsger Disease Severity Scale Compared with the Physician's Global Assessment of Disease Severity in Systemic Sclerosis

Daphna Harel, Marie Hudson, Alexandra Iliescu, Murray Baron, Canadian Scleroderma Research Group, and Russell Steele

ABSTRACT. ***Objective.*** To develop a weighted summary score for the Medsger Disease Severity Scale (DSS) and to compare its measurement properties with those of a summed DSS score and a physician's global assessment (PGA) of severity score in systemic sclerosis (SSc).

Methods. Data from 875 patients with SSc enrolled in a multisite observational research cohort were extracted from a central database. Item response theory was used to estimate weights for the DSS weighted score. Intraclass correlation coefficients (ICC) and convergent, discriminative, and predictive validity of the 3 summary measures in relation to patient-reported outcomes (PRO) and mortality were compared.

Results. Mean PGA was 2.69 (SD 2.16, range 0–10), mean DSS summed score was 8.60 (SD 4.02, range 0–36), and mean DSS weighted score was 8.11 (SD 4.05, range 0–36). ICC were similar for all 3 measures [PGA 6.9%, 95% credible intervals (CrI) 2.1–16.2; DSS summed score 2.5%, 95% CrI 0.4–6.7; DSS weighted score 2.0%, 95% CrI 0.1–5.6]. Convergent and discriminative validity of the 3 measures for PRO were largely similar. In Cox proportional hazards models adjusting for age and sex, the 3 measures had similar predictive ability for mortality (adjusted R^2 13.9% for PGA, 12.3% for DSS summed score, and 10.7% DSS weighted score).

Conclusion. The 3 summary scores appear valid and perform similarly. However, there were some concerns with the weights computed for individual DSS scales, with unexpected low weights attributed to lung, heart, and kidney, leading the PGA to be the preferred measure at this time. Further work refining the DSS could improve the measurement properties of the DSS summary scores. (First Release June 15 2016; J Rheumatol 2016;43:1510–18; doi:10.3899/jrheum.151440)

Key Indexing Terms:
SYSTEMIC SCLEROSIS
SCLERODERMA
DISEASE SEVERITY SCORE
DISEASE SEVERITY
OUTCOME ASSESSMENT

Systemic sclerosis (SSc) is a chronic, heterogeneous multisystem disease. A barrier to the study of SSc has been the difficulty in measuring disease status[1,2,3]. Disease activity measures the potentially reversible aspects of disease that vary over time[3,4,5,6]. Disease damage measures the irreversible tissue injury[3,4,5,6]. Our study focused on measuring disease severity, the total effect of disease on organ function including both reversible and irreversible components[5].

From the Center for the Promotion of Research Involving Innovative Statistical Methodology (PRIISM) Applied Statistics Center, New York University; Department of Humanities and Social Sciences in the Professions, New York University, New York, New York, USA; Division of Rheumatology, Jewish General Hospital; Lady Davis Institute, Jewish General Hospital; Department of Medicine, McGill University; Mathematics and Statistics, McGill University, Montreal, Quebec, Canada.

The Canadian Scleroderma Research Group (CSRG) is funded by the Canadian Institutes of Health Research (CIHR; grant #FRN 83518), the Scleroderma Society of Canada and its provincial Chapters, the Scleroderma Society of Ontario, the Scleroderma Society of Saskatchewan, Sclérodermie Québec, and the Cure Scleroderma Foundation, and receives support from INOVA Diagnostics Inc., Dr. Fooke Laboratorien GmbH, Euroimmun, and Mikrogen GmbH. The CSRG has also received educational grants from Pfizer and Actelion pharmaceuticals. Dr. Hudson is funded by the Fonds de la recherche en Santé du Québec. Alexandra Iliescu was funded by a CIHR Undergraduate research award. Dr. Harel's work was funded by the CSRG. Dr. Steele's work was funded in part by the Natural Sciences and Engineering Research Council of Canada.

D. Harel, PhD, PRIISM Applied Statistics Center, New York University, and the Department of Humanities and Social Sciences in the Professions, New York University; M. Hudson, MD, MPH, Division of Rheumatology, Jewish General Hospital, and the Lady Davis Institute, Jewish General Hospital, and the Department of Medicine, McGill University; A. Iliescu, BS, Lady Davis Institute, Jewish General Hospital, Montreal; M. Baron, MD, Division of Rheumatology, Jewish General Hospital, and the Lady Davis Institute, Jewish General Hospital, and the Department of Medicine, McGill University; R. Steele, PhD, Lady Davis Institute, Jewish General Hospital, Montreal, and Mathematics and Statistics, McGill University.

Address correspondence to Dr. M. Hudson, Jewish General Hospital, Room A-725, 3755 Côte Sainte-Catherine Road, Montreal, H3T 1E2 Quebec, Canada. E-mail: marie.hudson@mcgill.ca

Accepted for publication April 21, 2016.

Two common measures for severity in SSc are the Scleroderma Disease Severity Scale (DSS) developed by Medsger, *et al*[5,7], and the physician's global assessment (PGA) of severity. The DSS rates the severity of SSc in 9 organ systems, each scored separately depending on the level of involvement (no, mild, moderate, severe, or endstage). The PGA reflects a physician's judgment of the subject's overall disease severity using the visual analog scale or the numerical rating scale (NRS) while considering all information available. In the absence of a gold standard, the DSS and PGA are commonly used to estimate disease status[4], and despite not having been extensively validated, are nevertheless believed to be accurate[8] and are widely used both in SSc[9,10] and in other rheumatic diseases[11].

Choosing between the SSc severity measures requires a careful examination of advantages and disadvantages, both practical and numerical. An important limitation of the DSS, specifically acknowledged by the authors of the scale, is that it results in 9 separate scores[5]. Nonetheless, a simple summed score of the original 9 or modified versions of the DSS scores have been used without validation[12,13,14,15].

A summed score requires an assumption that each of the items, for example lung and joint/tendon severity, provide equal amounts of discrimination for disease severity. Weighted alternatives to the summed score drop this assumption while maintaining the simplicity of a single-number summary. Alternatively, the PGA is simple and highly feasible, but inherently incorporates subjective physician opinion, which may inject additional heterogeneity into the measure. We undertook this study to develop a weighted summary score (WSS) for the DSS and compare its measurement properties with those of a PGA and a summed DSS score.

MATERIALS AND METHODS

Study subjects. The Canadian Scleroderma Research Group (CSRG) includes subjects with SSc recruited from 16 centers. Ethics committee approval for the CSRG data collection and study protocols was obtained at McGill University (Montreal, Quebec, Canada) and at all participating study sites. All subjects provided informed written consent to participate. Our study did not require additional ethical approval.

All subjects in the registry must have a diagnosis of SSc confirmed by a rheumatologist, be ≥ 18 years of age, and be fluent in English, French, or Spanish. Over 98% of the cohort meets the 2013 American College of Rheumatology/European League Against Rheumatism classification criteria for SSc[16]. Subjects have been recruited since 2004 and were seen at baseline and yearly thereafter. The subjects in our study included those whose baseline visit was between September 2004 and February 2013, and who had complete data for both the DSS and PGA. Data were collected for all study instruments and variables at the baseline visit.

Study instruments. Measures of disease severity were the DSS[5,7] and the PGA. The DSS assesses disease severity in 9 organ systems: general health, peripheral vascular, skin, joint/tendon, muscle, gastrointestinal (GI) tract, lungs, heart, and kidneys. Each organ is scored separately from 0 to 4 depending on whether there is no, mild, moderate, severe, or endstage involvement. For the purposes of our study, some adaptations were made. The results of any investigation not requested by the physician were considered "normal"[5]. For the skeletal muscle system, physicians assessed muscle strength in the neck flexors and the right and left, upper and lower proximal extremities using the British Medical Research Council scale[17], and calculated as reported previously[18]. The Health Assessment Questionnaire (HAQ; described below) was used to assess the patient's use of ambulation aids needed to assign endstage severity for the skeletal muscle system. To score the GI system, in addition to the standard tests (an abnormal esophagram, abnormal esophageal manometry, or abnormal small bowel series), subjects were also given a score of 1 for mild GI disease severity if they reported difficulty swallowing, acid taste in their mouth, choking at night, burning sensation, feeling of being full shortly after eating, or taking gastroprotective or promotility agents. If malabsorption, episodes of pseudo-obstruction, or abnormal hydrogen breath test were present, a score of 3 for severe GI disease severity was given. To score the heart system, physicians also considered electrocardiogram results, left ventricular ejection fraction values, presence of conduction abnormalities, distended neck veins, and arrhythmias. Full details on these adaptations can be found elsewhere[18,19]. Study physicians recorded the PGA of disease severity using an 11-point NRS ranging from 0 (no disease) to 10 (very severe disease).

Study variables. Disease duration was measured from the onset of both the first Raynaud and first non-Raynaud disease manifestation to baseline study visit. Subjects were classified into limited (skin involvement of the arms and/or legs distal to elbows or knees, with or without facial involvement) and diffuse (skin involvement of the proximal limbs and/or trunk) cutaneous subsets (lcSSc and dcSSc, respectively)[20] according to the maximum extent of skin involvement at any time during their participation in the cohort. SSc sine scleroderma was classified as lcSSc[21].

Mortality was assessed at any point in the study period (2004–2013) based on information provided by the physicians, or notice of death.

Function was assessed using the HAQ Disability Index[22], with scores ranging from 0 (no disability) to 3 (severe disability), and patient ratings of a series of SSc symptoms from the Scleroderma HAQ (SHAQ)[23,24,25,26].

Health-related quality of life (HRQOL) was measured using the Medical Outcomes Study Short Form-36 (SF-36)[27]. The SF-36 is a self-administered, generic HRQOL questionnaire covering 8 domains. Each domain is scored separately and combined into physical (PCS) and mental component summary (MCS) scores and normalized based on a general population sample.

Similarly to the physicians, subjects were asked to rate the severity of their disease on a scale from 0 to 10, yielding the patient's global assessment (PtGA) of disease severity.

Statistical analysis. Descriptive statistics summarized the baseline characteristics of the study subjects. Three composite measures of disease severity were compared: the DSS summed score, the DSS WSS, and the PGA. The summed score was calculated by adding the scores of the 9 organ systems for each individual variable, thus ranging from 0 (lower severity) to 36 (higher severity). Weights for the WSS were obtained using item response theory (IRT) to estimate organ-specific discrimination variables by fitting a generalized partial credit model (GPCM)[28] to the 9 organ system subscales of the DSS. For each organ system, the GPCM estimates both the level of severity at which a patient is more likely to be categorized in 1 category instead of the 1 below, and a discrimination variable that measures the strength of the relationship between the organ system and severity. The WSS, which weights each organ system's score by the organ system's discrimination variable, was then calculated and scaled to range from 0 to 36, allowing for direct comparison with the summed score[29,30]. The score for the PGA was the number recorded by the study physician between 0–10.

Because the PGA incorporates the physician's subjective opinion, interrater reliability was assessed by the intraclass correlation coefficient (ICC). The ICC was computed for each composite measure, which represents the magnitude of variability introduced by individual physicians. Because of the small number of physicians, 31 in total, we used a Bayesian hierarchical model to obtain 95% credible intervals (CrI) for the 3 ICC.

The convergent and discriminative construct validity of each composite measure was assessed. For convergent validity, correlations were computed to compare associations of the study instruments with patient-reported outcomes (PRO). Both nonparametric Kendall tau and Spearman rank correlation were used to account for their differing emphases.

For discriminative validity, dichotomous subsets were constructed to identify subjects with less and more severe disease based on the median values of various PRO. The mean disease severity scores of each subset were computed and the differences in these means were tested using the Wilcoxon rank-sum test.

Cox proportional hazards models were fit to assess the extent to which each composite measure was predictive of mortality by estimating the proportional change that can be expected in the hazard related to changes in the composite measure. First, a baseline model was fit, controlling for age and sex. Cox proportional hazard models additionally adjusting for 1 of the composite measures were compared. The relative predictive ability of each measure was assessed through a comparison of R^2 values. For each model, the proportional hazards assumption was tested using a chi-square test of the scaled Schoenfeld residuals. For each composite measure, dichotomous subsets of the subjects were constructed by splitting subjects into 2 groups based on the median values and log-rank tests assessed whether there was a statistically significant difference in mortality between those with low and high values.

To assess statistical significance, we applied a posthoc Bonferroni correction factor for each of the 54 independent convergent validity comparisons ($p < 0.0009$) and each of the 27 independent discriminative validity comparisons ($p < 0.002$).

All analyses were done using R version 3.1.1[31]. The GPCM was fit using the ltm package[32]. The Bayesian models were fit using JAGS and the R2jags packages[33,34]. The proportional hazards models were fit using the survival package[35].

RESULTS

The study included 875 subjects (Table 1). About 86% were women with a mean age of about 55 years. Mean disease duration was 11.1 years since the first non-Raynaud symptom and 14.6 years since the first Raynaud symptom. About 37% of subjects had dcSSc. Disease severity, measured by organ system, was mild to moderate, with the GI tract (mean 1.95, SD 0.81), peripheral vascular system (1.58, SD 1.24), lungs (1.41, SD 1.11), and skin (1.24, SD 0.66) being the most severe.

The discrimination variables and the rescaled weights for the 9 DSS scores estimated using the GPCM are presented in Table 2. The skin scale was most discriminating among subjects, being weighted 2.47× higher in the weighted compared with the summed score. The general system, joint/tendon, GI, and muscle systems had weights about equal to 1, and the peripheral vascular, heart, lung, and kidney scales received weights below 1.

The mean summed score was 8.60 (SD 4.02) and the mean WSS was 8.11 (SD 4.05), both compared with a maximum score of 36. The mean PGA was 2.69 (SD 2.16), compared with a maximum of 10. There were 578 unique sets of scores on the 9 organ subscales, resulting in observing 26 of the possible 36 unique values of the summed score (26 unique values observed were 0 to 24 and 26) and 578 of the possible 875 unique values of the WSS (range 0–25.85). All 11 possible values of the PGA were observed. Figure 1 shows summary plots of the 3 composite measures. As expected, the WSS and summed score were highly correlated (Figure 1A). Nevertheless, there was substantial variation in the center of the distribution (e.g., the WSS for subjects with a summed score of 10 ranged from 5.78 to 13.19), suggesting that the measures would not yield exactly the same ordering of subjects.

Table 1. Baseline characteristics of the study cohort (n = 875). Low values for DSS scores, global assessments, and HAQ represent better outcomes, and high values represent worse outcomes. Low values for SF-36 PCS and MCS represent worse outcomes and high values represent better outcomes. Values are mean (SD) unless otherwise specified.

Characteristics	Values
Female, n (%)	755 (86.3)
Age, yrs	55.2 (12.0)
Disease duration, yrs	
First non-Raynaud symptom	11.1 (9.7)
First Raynaud symptom	14.6 (12.4)
Disease subsets by extent of cutaneous involvement, n (%)	
Limited	545 (62.7)
Diffuse	323 (37.2)
DSS scores, range 0–4	
General health	0.87 (1.19)
GI tract	1.95 (0.81)
Heart	0.49 (1.00)
Joint/tendon	0.73 (1.21)
Kidneys	0.11 (0.60)
Lungs	1.41 (1.11)
Muscle	0.22 (0.70)
Peripheral vascular	1.58 (1.24)
Skin	1.24 (0.66)
PtGA, range 0–10	
Pain	3.58 (2.78)
Raynaud phenomenon	2.88 (2.87)
Finger ulcers	1.99 (2.99)
GI problems	1.74 (2.60)
Breathing	2.01 (2.54)
Disease severity	3.58 (2.60)
SF-36 PCS score	37.1 (10.02)
SF-36 MCS score	48.7 (11.91)
HAQ, range 0–3	0.77 (0.69)
Deaths observed, n (%)	120 (13.7)
Time to death, yrs	2.89 (1.99)
DSS summed score, range 0–36	8.60 (4.02)
DSS WSS, range 0–36	8.11 (4.05)
PGA of severity, range 0–10	2.69 (2.16)

DSS: Medsger Disease Severity Scale; HAQ: Health Assessment Questionnaire; SF-36: Medical Outcomes Study Short Form-36; PCS: physical component summary; MCS: mental component summary; GI: gastrointestinal; PtGA: patient's global assessment; WSS: weighted summed score; PGA: physician's global assessment.

Assessing between-physician heterogeneity. The ICC for the WSS was 2.0% (95% Bayesian CrI 0.1–5.6), for the DSS summed score it was 2.5% (95% CrI 0.4–6.7), and for the PGA it was 6.9% (95% CrI 2.1–16.2). Although the measured ICC for the PGA was the largest, its absolute magnitude was still small, indicating that it still did not represent a substantial part of the variability of the PGA. Therefore, there were no meaningful differences in the subjective contribution of the physician to the 3 measures.

Table 2. Discrimination variables from the generalized partial credit model and weights for the WSS.

Medsger Disease Severity Scale	Discrimination Variables (95% CI)	Multiplicative Weight for WSS (95% CI)
General	0.57 (0.41–0.73)	1.18 (0.85–1.52)
Peripheral vascular	0.20 (0.12–0.28)	0.42 (0.25–0.59)
Skin	1.19 (0.64–1.74)	2.47 (1.33–3.61)
Joint/tendon	0.45 (0.32–0.58)	0.93 (0.66–1.21)
Muscle	0.59 (0.38–0.80)	1.23 (0.80–1.66)
GI tract	0.50 (0.31–0.68)	1.03 (0.65–1.41)
Lung	0.34 (0.23–0.45)	0.71 (0.48–0.94)
Heart	0.25 (0.15–0.35)	0.52 (0.30–0.73)
Kidney	0.24 (0.09–0.40)	0.51 (0.18–0.83)

WSS: weighted summed score; GI: gastrointestinal.

Construct and discriminative validity. Kendall τ and Spearman ρ correlations of all 3 composite measures with the SF-36 PCS, the HAQ and PtGA of pain, GI problems, breathing, and severity were statistically significant and moderate in strength (Table 3). However, all correlations with the SF-36 MCS and the Kendall τ between the PGA and Raynaud phenomenon global assessments were weak and nonstatistically different from 0 under the Bonferroni correction. For each outcome and correlation considered, with the exception of the finger ulcer global assessment, the bootstrap CI for the 3 composite measures overlapped, indicating no difference in the strength of association with any of the 3 composite measures. All 3 composite measures were able to discriminate between subjects with better or worse scores on all PRO, with the exception of the PGA on the GI problem global assessment (Table 4).

Predictive validity for mortality. Death was observed for 120 patients (13.7%) in the study, with mean time to death of 2.89 years (SD 1.99 yrs). A reference Cox proportional hazards survival model that included age and sex as baseline covariates yielded an R^2 value of 4.0% and a concordance probability of 0.66. Three further Cox proportional hazards models, each adjusting for 1 composite measure and age and sex, were generated. The model with the PGA had an R^2 of

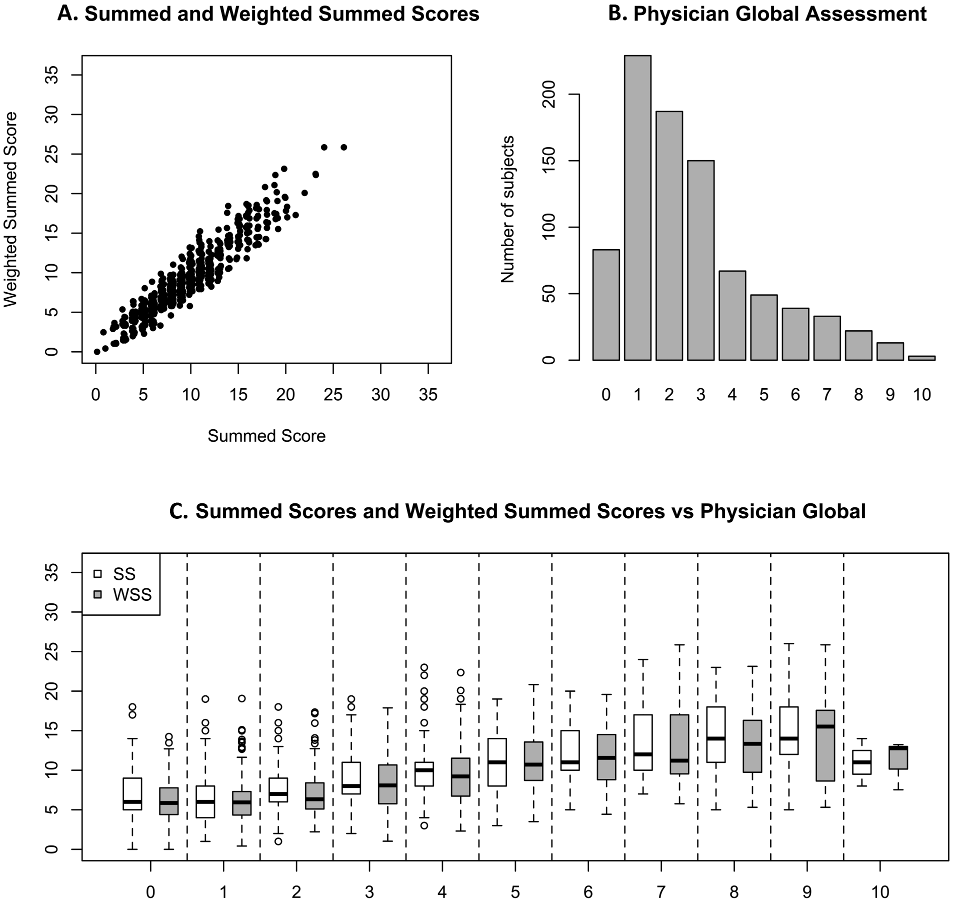

Figure 1. Comparison of the 3 composite measures of disease severity. A. A plot of the simple SS versus the WSS. B. The counts of individuals at each value of the PGA of severity. C. Boxplots of the simple SS (white) and WSS (grey) for subjects at each level of the PGA. SS: summed score; WSS: weighted summed score; PGA: physician's global assessment.

Table 3. Convergent validity. Nonparametric correlations between composite severity scores and outcome measures. Nonparametric correlations between each outcome and each composite measure with 95% bootstrap CI. Bonferroni p value < 0.0009.

Measures	DSS Summed Score		DSS WSS		PGA	
	Kendall τ	p	Kendall τ	p	Kendall τ	p
SF-36 PCS	–0.26 (–0.30 to –0.22)	< 0.0001	–0.25 (–0.29 to –0.20)	< 0.0001	–0.31 (–0.36 to –0.26)	< 0.0001
SF-36 MCS	–0.07 (–0.12 to –0.02)	0.005	–0.05 (–0.10 to –0.01)	0.023	–0.07 (–0.12 to –0.02)	0.004
HAQ	0.32 (0.28–0.36)	< 0.0001	0.32 (0.27–0.36)	< 0.0001	0.30 (0.26–0.34)	< 0.0001
Patient-reported						
Pain	0.19 (0.13–0.23)	< 0.0001	0.18 (0.13–0.23)	< 0.0001	0.20 (0.16–0.25)	< 0.0001
Raynaud phenomenon	0.15 (0.10–0.21)	< 0.0001	0.12 (0.07–0.16)	< 0.0001	0.09 (0.04–0.14)	0.001
Finger ulcers	0.29 (0.24–0.33)	< 0.0001	0.22 (0.17–0.27)	< 0.0001	0.15 (0.09–0.20)	< 0.0001
GI problems	0.16 (0.11–0.21)	< 0.0001	0.12 (0.07–0.17)	< 0.0001	0.10 (0.04–0.15)	< 0.005
Breathing	0.21 (0.16–0.26)	< 0.0001	0.16 (0.12–0.21)	< 0.0001	0.21 (0.16–0.27)	< 0.0001
Disease severity	0.28 (0.23–0.33)	< 0.0001	0.27 (0.22–0.31)	< 0.0001	0.30 (0.26–0.34)	< 0.0001
Measures	Spearman ρ	p	Spearman ρ	p	Spearman ρ	p
SF-36 PCS	–0.37 (–0.43 to –0.31)	< 0.0001	–0.37 (–0.43 to –0.31)	< 0.0001	–0.43 (–0.49 to –0.37)	< 0.0001
SF-36 MCS	–0.09 (–0.16 to –0.03)	0.005	–0.08 (–0.15 to –0.01)	0.023	–0.10 (–0.16 to –0.03)	0.005
HAQ	0.43 (0.38–0.49)	< 0.0001	0.44 (0.38–0.49)	< 0.0001	0.40 (0.34–0.45)	< 0.0001
Patient-reported						
Pain	0.25 (0.19–0.31)	< 0.0001	0.25 (0.18–0.31)	< 0.0001	0.27 (0.20–0.33)	< 0.0001
Raynaud phenomenon	0.20 (0.13–0.26)	< 0.0001	0.16 (0.10–0.23)	< 0.0001	0.11 (0.04–0.17)	< 0.0001
Finger ulcers	0.36 (0.30–0.42)	< 0.0001	0.29 (0.23–0.35)	< 0.0001	0.18 (0.12–0.25)	< 0.0001
GI problems	0.20 (0.13–0.27)	< 0.0001	0.17 (0.10–0.24)	< 0.0001	0.12 (0.06–0.18)	< 0.005
Breathing	0.27 (0.21–0.33)	< 0.0001	0.22 (0.15–0.28)	< 0.0001	0.27 (0.20–0.33)	< 0.0001
Disease severity	0.37 (0.31–0.43)	< 0.0001	0.37 (0.32–0.42)	< 0.0001	0.39 (0.34–0.45)	< 0.0001

DSS: Medsger Disease Severity Scale; WSS: weighted summed score; PGA: physician's global assessment; SF-36: Medical Outcomes Study Short Form-36; PCS: physical component summary; MCS: mental component summary; HAQ: Health Assessment Questionnaire; GI: gastrointestinal.

13.9%, followed by the model with the DSS summed score (12.3%) and that with the WSS (10.7%). While each composite measure provided some additional explanatory power over the reference model, the differences in explanatory power among the 3 were small. There was insufficient evidence to reject the proportional hazards assumption for all 3 models ($p > 0.05$), indicating that the model assumptions were satisfied. Similarly, mortality between subjects with low and high values on each of the measures was significantly different ($p < 0.05$), indicating that all 3 composite measures were predictive of mortality. Thus, in so far as predictive validity, the 3 measures were again about similar.

Posthoc analysis of the DSS organ scale weights. The unexpected low weights of the DSS lung, heart, and kidney scales in the WSS led to some posthoc analyses. First, box plots of the PGA at each DSS skin scale level indicated that the median score of the PGA was visibly different across levels (data not shown). However, for lung, heart, and kidney, the relationship between the DSS and PGA scores was not monotonically increasing (Figures 2A, 2B, and 2C), illustrating that, unlike the skin scale, the lung, heart, and kidney scales had poor discriminatory ability for the latent trait of severity. This provides an explanation, at least in part, for their weights.

Three variables compose the DSS lung scale: systolic pulmonary artery pressure (sPAP), forced vital capacity (FVC), and DLCO. Tables cross-classifying patients indicated significant heterogeneity in these 3 measures for subjects at the same level of the DSS lung scale (data not shown). In addition, box plots of the PGA scores against each of these 3 variables showed that the bottom categories of both sPAP and DLCO and the top categories of the FVC did not provide meaningful discrimination for different values of PGA (Figures 2D, 2E, and 2F), further demonstrating the poor discriminatory ability of the DSS lung for severity, as measured by the PGA.

Eighty-seven percent of subjects had normal or mild scores for heart severity and over 96% were normal for kidney severity, indicating a lack of endorsement of the higher categories. Cross-tabulations between each of the DSS organ scales showed that subjects with high scores for kidney and heart did not generally have high scores on the other DSS organ scales (data not shown).

Finally, we performed a sensitivity analysis based on disease subset and disease duration since the first non-Raynaud symptom. When stratifying by disease subset, there were no statistically significant differences in the weights for lcSSc or dcSSc compared with those calculated on all subjects. When stratifying by short (≤ 3 yrs) versus long disease duration, only the weight for the GI system was statistically lower than for all subjects. In comparison, the PGA performed similarly among all subsets of patients.

Table 4. Discriminative validity of the composite measures (Bonferroni p value < 0.002).

Measures	n	DSS Summed Score		DSS WSS		PGA	
		Mean (SD)	p	Mean (SD)	p	Mean (SD)	p
SF-36 PCS	837						
≤ 37.18		10.03 (4.34)	< 0.001	9.56 (4.53)	< 0.001	3.53 (2.40)	< 0.001
> 37.18		7.32 (3.17)		6.81 (3.04)		1.90 (1.53)	
SF-36 MCS	837						
≤ 50.84		9.24 (4.29)	< 0.001	8.67 (4.37)	< 0.001	2.97 (2.30)	< 0.001
> 50.84		8.11 (3.68)		7.70 (3.74)		2.46 (2.01)	
HAQ	844						
≤ 0.625		7.23 (3.14)	< 0.001	6.67 (2.93)	< 0.001	1.99 (1.72)	< 0.001
> 0.625		10.25 (4.30)		9.82 (4.52)		3.50 (2.34)	
Patient-reported							
Pain	843						
≤ 3		7.52 (3.52)	< 0.001	7.28 (3.42)	< 0.001	2.36 (2.10)	< 0.001
> 3		9.14 (4.14)		9.24 (4.54)		3.12 (2.20)	
Raynaud phenomenon	844						
≤ 2		8.06 (3.86)	< 0.001	7.64 (3.89)	< 0.001	2.52 (2.15)	< 0.001
> 2		9.39 (4.13)		8.80 (4.23)		2.94 (2.19)	
Finger ulcers	840						
= 0		7.58 (3.58)	< 0.001	7.28 (3.62)	< 0.001	2.48 (2.14)	< 0.001
> 0		10.16 (4.15)		9.40 (4.38)		3.03 (2.19)	
GI problems	840						
= 0		7.95 (3.64)	< 0.001	7.60 (3.71)	< 0.001	2.54 (2.14)	0.003
> 0		9.54 (4.33)		8.87 (4.43)		2.91 (2.19)	
Breathing	839						
≤ 1		7.79 (3.56)	< 0.001	7.45 (3.66)	< 0.001	2.21 (1.81)	< 0.001
> 1		9.79 (4.32)		9.09 (4.40)		3.35 (2.43)	
Disease severity	841						
≤ 3		7.51 (3.42)	< 0.001	7.00 (3.36)	< 0.001	2.06 (1.73)	< 0.001
> 3		9.96 (4.28)		9.47 (4.44)		3.44 (2.39)	

DSS: Medsger Disease Severity Scale; WSS: weighted summed score; PGA: physician's global assessment; SF-36: Medical Outcomes Study Short Form-36; PCS: physical component summary; MCS: mental component summary; HAQ: Health Assessment Questionnaire; GI: gastrointestinal.

DISCUSSION

We have shown that the DSS summed and WSS and a PGA of severity each showed moderate levels of convergent and discriminative validity and predictive validity for mortality. Although the PGA had the potential to and did contain more between-physician heterogeneity than the other 2 measures, the amount of physician-specific heterogeneity relative to the total variability of the measure was small and did not impair the performance of the PGA in terms of construct or predictive validity.

To construct the WSS, a GPCM was used to obtain weights for disease severity, allowing for the weights of the 9 organ scales to be internal to the instrument. Thus, the WSS based on these weights can be used regardless of what other measures it may be compared to. While multivariate linear regression procedure could have been used to generate weights, any weights obtained would be specifically tuned to a particular outcome and would not necessarily be generalizable. Principal components would not have been an appropriate alternative either because they require continuous outcomes and would not have respected the categorical design of the DSS scales.

The weights for the WSS are obtained from the GPCM using maximum likelihood providing the best 1-dimensional summary of the 9 DSS organ subscales under the restriction that an increasing latent severity score cannot result in a decreasing expected organ subscale score for any organ[30,36]. Although disease severity is poorly represented through a unidimensional latent construct, we rather present the WSS as a more flexible, 1-dimensional alternative to the summed score that removes the naive assumption that all organ systems are equally discriminative of disease severity. Note that even though in other situations multidimensional summaries of disease severity might be found to be more useful, they require larger sample sizes for estimating variables and are more difficult to interpret.

The weights obtained from the IRT models were unexpectedly low for the lung, heart, and kidney systems. These 3 DSS scales did not discriminate well among subjects with higher and lower disease severity (Figure 2). The proposed cutoffs for FVC, DLCO, and sPAP, when examined separately, did not adequately discriminate between different degrees of severity, leading to considerable heterogeneity (Figure 2). The low weights for the heart and kidney scales may have occurred because of deviation from the assumption of unidimensionality of disease severity required by the

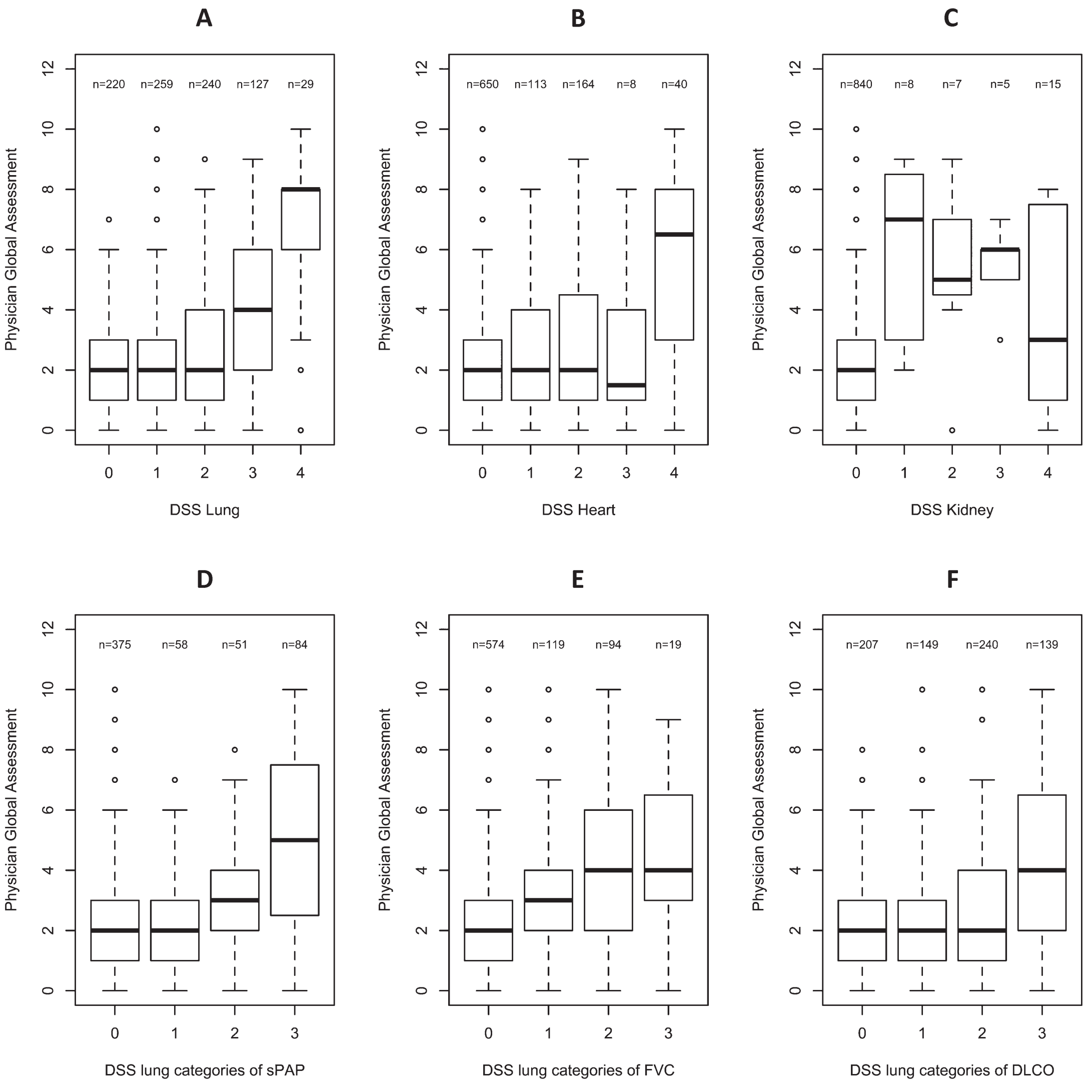

Figure 2. A, B, and C: boxplots of the PGA scores at each level of the DSS lung, heart, and kidney scales. D, E, and F: boxplots of the PGA scores at each level of the components of the DSS lung score. PGA: physician's global assessment; DSS: Medsger Disease Severity Scale; sPAP: systolic pulmonary artery pressure; FVC: forced vital capacity.

GPCM. Subjects with extreme scores on these scales did not systematically have high scores on other organ scales. However, the summed score would be susceptible to the same problem, because it also assumes unidimensionality. Alternatively, the low weights may also have been because of low rates of endorsement across the spectrum of severity, a possible interaction with time not identified in the DSS score, or from those with asymptomatic disease obtaining intermediate scores.

Because all 3 measures appear to be valid, it is of interest to consider whether 1 measure should be preferred. The PGA had slightly lower correlations with 3 of the PtGA on the SHAQ than the DSS summed and WSS. This could be because of the way in which the DSS more directly accounts for these symptoms, rather than a shortcoming in the PGA ability to measure disease severity. Therefore, because the PGA is the simplest to record and its measurement properties were similar to those of the more complex DSS summed score and the WSS, it appears as the preferred measure for global disease severity in SSc, particularly if some variables required in the DSS were not collected. However, while the DSS is a cross-sectional measure based on objective criteria, the PGA is inherently subjective and allows the physician to include information about the observed and potential disease trajectory. Because all CSRG investigators are experienced clinicians in SSc, the relative benefit of the PGA over the DSS summed score may be due to the high levels of familiarity with the disease; inexperienced physicians may benefit from using the more objective DSS summed score.

Further work refining the DSS, in particular scoring some rare but severe renal and cardiac manifestations differently, revising the DSS lung scale by using different cutoffs or separating variables that measure different aspects of cardiopulmonary disease (e.g., FVC, DLCO, sPAP), and determining the effect of disease duration on severity could potentially improve the measurement properties of a revised DSS summed score or WSS. Last, further work studying the 3 measures longitudinally could result in additional information regarding their relative use.

The DSS summed and WSS and a PGA of severity using an NRS ranging from 0–10 had moderate levels of construct and predictive validity and low levels of between-physician heterogeneity. The PGA is the simplest measure for global disease severity to record and may be the preferred measure for experienced clinicians in SSc at this time.

ACKNOWLEDGMENT

We are grateful for a critical review of the manuscript and the helpful suggestions of Dr. Thomas Medsger.

APPENDIX 1.

List of study collaborators. Investigators of the Canadian Scleroderma Research Group: J. Pope, London, Ontario; J. Markland, Saskatoon, Saskatchewan (deceased); D. Robinson, Winnipeg, Manitoba; N. Jones, Edmonton, Alberta; N. Khalidi, Hamilton, Ontario; P. Docherty, Moncton, New Brunswick; E. Kaminska, Calgary, Alberta; A. Masetto, Sherbrooke, Quebec; E. Sutton, Halifax, Nova Scotia; J.P. Mathieu, Montreal, Quebec; S. Ligier, Montreal, Quebec; T. Grodzicky, Montreal, Quebec; S. LeClercq, Calgary, Alberta; C. Thorne, Newmarket, Ontario; G. Gyger, Montreal, Quebec; D. Smith, Ottawa, Ontario; P.R. Fortin, Quebec City, Quebec; M. Larché, Hamilton, Ontario; M. Abu-Hakima, Calgary, Alberta; T.S. Rodriguez-Reyna, Mexico City, Mexico; A.R. Cabral, Mexico City, Mexico; M. Fritzler, Calgary, Alberta.

REFERENCES

1. Clements PJ. Measuring disease activity and severity in scleroderma. Curr Opin Rheumatol 1995;7:517-21.
2. Medsger TA Jr. Assessment of damage and activity in systemic sclerosis. Curr Opin Rheumatol 2000;12:545-8.
3. Medsger TA Jr. Natural history of systemic sclerosis and the assessment of disease activity, severity, functional status, and psychologic well-being. Rheum Dis Clin North Am 2003;29:255-73.
4. Symmons DP. Disease assessment indices: activity, damage and severity. Baillieres Clin Rheumatol 1995;9:267-85.
5. 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.
6. Valentini G, Silman AJ, Veale D. Assessment of disease activity. Clin Exp Rheumatol 2003;21 Suppl 29:S39-41.
7. Medsger TA Jr, Bombardieri S, Czirjak L, Scorza R, Della Rossa A, Bencivelli W. Assessment of disease severity and prognosis. Clin Exp Rheumatol 2003;21 Suppl 29:S42-6.
8. Clements PJ, Seibold JR, Furst DE, Mayes M, White B, Wigley F, et al. High-dose versus low-dose D-penicillamine in early diffuse systemic sclerosis trial: lessons learned. Semin Arthritis Rheum 2004;33:249-63.
9. Hudson M, Impens A, Baron M, Seibold JR, Thombs BD, Walker JG, et al; Canadian Scleroderma Research Group. Discordance between patient and physician assessments of disease severity in systemic sclerosis. J Rheumatol 2010;37:2307-12.
10. Fan X, Pope J, Baron M; Canadian Scleroderma Research Group. 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 2010;30:1205-10.
11. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729–40.
12. Rimar D, Rosner I, Nov Y, Slobodin G, Rozenbaum M, Halasz K, et al. Brief report: lysyl oxidase is a potential biomarker of fibrosis in systemic sclerosis. Arthritis Rheumatol 2014;66:726-30.
13. Gheita T, Hussein H. Cartilage Oligomeric Matrix Protein (COMP) in systemic sclerosis (SSc): role in disease severity and subclinical rheumatoid arthritis overlap. Joint Bone Spine 2012;79:51-6.
14. Hinchcliff ME, Beaumont JL, Carns MA, Podlusky S, Thavarajah K, Varga J, et al. Longitudinal evaluation of PROMIS-29 and FACIT-dyspnea short-forms in systemic sclerosis. J Rheumatol 2015;42:64-72.
15. Hinchcliff ME, Beaumont JL, Thavarajah K, Varga J, Chung A, Podlusky S, et al. Validity of two new patient-reported outcome measures in systemic sclerosis: Patient-Reported Outcomes Measurement Information System 29-item Health Profile and Functional Assessment of Chronic Illness Therapy-Dyspnea short form. Arthritis Care Res 2011;63:1620-8.
16. Alhajeri H, Hudson M, Fritzler M, Pope J, Tatibouet S, Markland J, et al. 2013 American College of Rheumatology/European League Against Rheumatism classification criteria for systemic sclerosis outperform the 1980 criteria: data from the Canadian Scleroderma Research Group. Arthritis Care Res 2015;67:582-7.
17. Matthews WB. Medical Research Council. Aids to the examination of the peripheral nervous system [abstract]. [Internet. Accessed May 2, 2016.] Available from: www.jns-journal.com/article/0022-510X(77)90205-2/abstract
18. Santiago M, Baron M, Hudson M, Burlingame RW, Fritzler MJ. Antibodies to RNA polymerase III in systemic sclerosis detected by ELISA. J Rheumatol 2007;34:1528-34.
19. Hudson M, Walker JG, Fritzler M, Taillefer S, Baron M. Hypocomplementemia in systemic sclerosis—clinical and serological correlations. J Rheumatol 2007;34:2218-23.
20. LeRoy EC, Medsger TA. Criteria for the classification of early systemic sclerosis. J Rheumatol 2001;28:1573-6.
21. Diab S, Dostrovsky N, Hudson M, Tatibouet S, Fritzler MJ, Baron M, et al. Systemic sclerosis sine scleroderma: a multicenter study of 1417 subjects. J Rheumatol 2014;41:2179-85.
22. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
23. Steen VD, Medsger TA Jr. The value of the Health Assessment Questionnaire and special patient-generated scales to demonstrate change in systemic sclerosis patients over time. Arthritis Rheum 1997;40:1984-91.
24. Van Tubergen A, Debats I, Ryser L, Londoño J, Burgos-Vargas R, Cardiel MH, et al. Use of a numerical rating scale as an answer modality in ankylosing spondylitis-specific questionnaires. Arthritis Rheum 2002;47:242-8.
25. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022-4.
26. Johnson SR, Hawker GA, Davis AM. The health assessment questionnaire disability index and scleroderma health assessment questionnaire in scleroderma trials: an evaluation of their measurement properties. Arthritis Care Res 2005;52:256-62.
27. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
28. Muraki E. A generalized partial credit model: application of an EM

algorithm. Appl Psychol Meas 1992;16:159-76.

29. Glas CA. Detection of differential item functioning using Lagrange multiplier tests. Stat Sin 1998;8:647-67.

30. Harel D. The effect of model misspecification for polytomous logistic adjacent category item response theory models [dissertation]. Montreal: McGill University; 2014:142.

31. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet. Accessed May 2, 2016.] Available from: www.R-project.org

32. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. J Stat Softw 2006;17:1-25.

33. Plummer M. rjags: Bayesian Graphical Models using MCMC. R package version 3-15. [Internet. Accessed May 2, 2016.] Available from: CRAN.R-project.org/package=rjags

34. Su Y, Yajima M. R2jags: Using R to run 'JAGS'. R package version 0.5-6. [Internet. Accessed May 2, 2016.] Available from: CRAN.R-project.org/package=R2jags

35. Therneau T. A Package for survival analysis in S. R package version 2.37-7. [Internet. Accessed May 2, 2016.] Available from: cran.r-project.org/package=survival

36. White H. Maximum likelihood estimation of misspecified models. Econometrica 1982;50:1-25.