Responsiveness of the SF-36 and the Health Assessment Questionnaire Disability Index in a Systemic Sclerosis Clinical Trial

DINESH KHANNA, DANIEL E. FURST, PHILIP J. CLEMENTS, GRACE S. PARK, RON D. HAYS, JEONGLIM YOON, JOSEPH H. KORN, PETER A. MERKEL, NAOMI ROTHFIELD, FREDRICK M. WIGLEY, LARRY W. MORELAND, RICHARD SILVER, VIRGINIA D. STEEN, MICHAEL WEISMAN, MAUREEN D. MAYES, DAVID H. COLLIER, THOMAS A. MEDSGER Jr, and JAMES R. SEIBOLD, for the Relaxin Study Group and the Scleroderma Clinical Trials Consortium

ABSTRACT. ***Objective.*** This study compares the responsiveness to change of the Medical Outcomes Study Short Form Health Survey (SF-36), a measure of health related quality of life (HRQOL), and the Health Assessment Questionnaire Disability Index (HAQ-DI), a function instrument, in a randomized clinical trial for treatment of systemic sclerosis (SSc).

Methods. A phase 2/3, multicenter, prospective, placebo controlled trial was conducted to evaluate human recombinant relaxin treatment in patients with diffuse SSc over 24 weeks. At baseline, subjects had stable, moderately severe, diffuse SSc of disease duration ≤ 5 years, modified Rodnan skin score ≥ 20, serum creatinine < 2.0 mg/dl, percentage forced vital capacity (% FVC) predicted ≥ 50%, and % DLCO predicted ≥ 40% and were not receiving concomitant disease modifying therapies. Internal consistency reliability of multi-item scales was estimated using Cronbach's alpha. Responsiveness to change of the SF-36 and HAQ-DI was computed between Weeks 0 and 24. Subjects were classified as unchanged or having a meaningful change in 4 different external measures: Change in (1) skin score ≥ 30%; (2) % FVC predicted of ≥ 15%; (3) self-reported patient global assessment by visual analog scale (VAS) ≥ 20%; and (4) physician global assessment by VAS of ≥ 20%. Responsiveness indices were computed and Cohen's effect size criteria were used to assess the magnitude of change.

Results. A total of 239 patients participated in this trial, with 196 completing the 24 week trial. Cronbach's alpha for the SF-36 scales ranged from 0.76 to 0.93 and for the HAQ-DI ranged from 0.69 to 0.91 (good to excellent). The SF-36 had a larger magnitude of responsiveness in overall disease (patient and physician global assessment) compared to the HAQ-DI, while the HAQ-DI had a larger magnitude of responsiveness in clinical measures (i.e., change in skin score and % FVC predicted) than the SF-36.

Conclusion. These data support inclusion of both the SF-36 and HAQ-DI as outcome measures in future clinical trials of diffuse SSc. (J Rheumatol 2005;32:832–40)

Key Indexing Terms:
HEALTH RELATED QUALITY OF LIFE SYSTEMIC SCLEROSIS SF-36
RESPONSIVENESS HEALTH ASSESSMENT QUESTIONNAIRE DISABILITY INDEX

From the Division of Rheumatology, Department of Medicine, and Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine, Los Angeles, California; Section of Rheumatology, Boston University, Boston, Massachusetts; University of Connecticut Health Center, Farmington, Connecticut; Johns Hopkins University, Baltimore, Maryland; University of Alabama at Birmingham, Birmingham, Alabama; University of South Carolina, Charleston, South Carolina; Division of Rheumatology, Georgetown University Medical Center, Washington, DC; Cedars-Sinai Medical Center, Los Angeles, California; University of Texas Health Science Center, Houston, Texas; University of Colorado, Denver, Colorado; University of Pittsburgh, Pittsburgh, Pennsylvania; and University of Michigan Scleroderma Program, Ann Arbor, Michigan, USA.

Dr. D. Khanna was supported by the Arthritis and Scleroderma Foundations (Physician Scientist Development Award) and a grant from the Scleroderma Clinical Trial Consortium. Dr. R.D. Hays was supported in part by the UCLA/DREW Project EXPORT, National Institutes of Health, National Center on Minority Health & Health Disparities (P20-MD00148-01) and the UCLA Center for Health Improvement in Minority Elders/Resource Centers for Minority Aging Research, National Institutes of Health, National Institute of Aging (AG-02-004). Dr. P.A. Merkel was supported in part by a Mid-Career Development Award in Clinical Investigation (NIH-NIAMS: K24 AR2224-01A1) and the National Center for Research Resources (NCRR), Boston University General Clinical Research Center Program Grant (NIH-NCRR: MO 1RR00533). The clinical trial was supported by Connetics Corporation, Palo Alto, California.

D. Khanna MD, MSc, Assistant Professor of Medicine; D.E. Furst, MD, Professor of Medicine; P.J. Clements, MD, MPH, Professor of Medicine; G.S. Park, MPH, Biostatistician; J. Yoon, BA, Master's Student in Biostatistics, Division of Rheumatology, Department of Medicine, David Geffen School of Medicine; R.D. Hays, PhD, Professor of Medicine, Division of General Internal Medicine and Health Services Research, David Geffen School of Medicine; J.H. Korn, MD, Professor of Medicine; P.A. Merkel, MD, MPH, Associate Professor of Medicine, Section of Rheumatology, Boston University; N. Rothfield, MD, Professor of Medicine, University of Connecticut Health Center; F.M. Wigley, MD,

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

Professor of Medicine, Johns Hopkins University; L.W. Moreland, MD, Professor of Medicine, University of Alabama at Birmingham; R. Silver, MD, Professor of Medicine, University of South Carolina; V.D. Steen, MD, Professor of Medicine, Division of Rheumatology, Georgetown University Medical Center; M. Weisman, MD, Professor of Medicine, Cedars-Sinai Medical Center; M.D. Mayes, MD, MPH, Professor of Medicine, University of Texas Health Science Center; D.H. Collier, MD, Professor of Medicine, University of Colorado; T.A. Medsger Jr, MD, Professor of Internal Medicine, University of Pittsburgh; J.R. Seibold, MD, Professor of Internal Medicine, University of Michigan Scleroderma Program.

The following investigators also participated in the study: M. Ellman, MD, University of Chicago, Chicago, IL; Y. Kim, MD, Stanford University, Stanford, CA; G.S. Firestein, MD; A.F. Kavanaugh, MD, University of California, San Diego, CA; M.E. Csuka, MD, Medical College of Wisconsin, Milwaukee, WI; R. Simms, MD, Boston University Medical Center, Boston, MA; V.M. Hsu, University of Medicine and Dentistry, Scleroderma Program, Newark, NJ.

Address reprint requests to Dr. D. Khanna, Division of Immunology, Department of Medicine, University of Cincinnati, PO Box 670563, Cincinnati, OH 45267-0563. E-mail: dinesh.khanna@uc.edu

Accepted for publication December 20, 2004.

Patient functioning and well being, health related quality of life (HRQOL), is an important outcome of treatment for chronic diseases[1]. Systemic sclerosis (scleroderma, SSc) is a chronic multisystem disease with a potentially important impact on HRQOL[2], but few studies have been conducted[3–7]. While physiologic measures provide information to clinicians, they often correlate poorly with HRQOL; patients are interested in both kinds of outcome measures[1].

Consequently, it is important to use a HRQOL instrument that is reliable (i.e., produces the same results repeatedly) and valid (it measures what it is intended to measure) including being responsive to changes in HRQOL over the course of observation[8,9]. A responsive measure should discriminate between patients who have improved, deteriorated, or remained stable[8].

There is a need to evaluate both generic HRQOL and disease-targeted instruments in SSc. This study assesses the Medical Outcomes Study Short Form Health Survey (SF-36, a generic measure)[10] and the Health Assessment Questionnaire Disability Index (HAQ-DI; a musculoskeletal-targeted measure)[11] in a prospective, randomized, placebo controlled trial evaluating recombinant human relaxin in the management of SSc with diffuse cutaneous involvement. Both the SF-36 and HAQ-DI have been used successfully in other rheumatic diseases[12-16], but only recently evaluated in SSc[3,5-7,17-21]. The objective of the original trial was to evaluate the efficacy and safety of relaxin in SSc. The present analysis evaluates the responsiveness to change of the SF-36 and HAQ-DI using data from the relaxin trial.

MATERIALS AND METHODS

Population and sample. The subjects were participants in a double blind, placebo controlled, multicenter randomized clinical trial evaluating the safety and efficacy of continuous subcutaneously infused recombinant human relaxin in patients with diffuse SSc over a period of 24 weeks. Sixteen centers throughout the United States participated in the trial. Subjects were randomized to either relaxin 25 µg/kg/day, relaxin 10 µg/kg/day, or placebo in a 2:1:2 ratio. The study was conducted in US academic scleroderma clinical research centers. All subjects had SSc as defined by the American College of Rheumatology criteria[22], with diffuse disease defined as the presence of thickening proximal as well as distal to the elbows and knees inclusive of the trunk and face[23].

Subjects 18–70 years old were recruited to reflect a population of patients with stable diffuse SSc (disease duration ≤ 5 years, modified Rodnan skin score not changed by more than 5 units during screening and baseline visit). The goal of the trial was to attempt to reverse established fibrosis in the skin. The other inclusion criterion was moderately severe skin disease with a skin score ≥ 20 (or ≥ 16 if there was truncal skin involvement)[23]. The major exclusion criteria included the presence of significant renal insufficiency as defined by serum creatinine ≥ 2.0 mg/dl, significant pulmonary disease [defined as percentage forced vital capacity (% FVC) < 50% of predicted and/or CO diffusion capacity (DLCO) < 40% of predicted], uncontrolled congestive heart failure, uncontrolled hypertension (defined as systolic > 160 mm Hg and diastolic > 100 mm Hg), use of > 10 mg prednisone/day, and current pregnancy. Concomitant use of agents reported to be effective therapies of SSc (e.g., D-penicillamine and immunosuppressives) were not permitted. The primary outcome measure of the trial was the change in modified Rodnan skin score. An improvement of 30% in the skin score was considered clinically meaningful based on consensus by SSc experts[24].

Two hundred thirty-nine patients with diffuse SSc were enrolled, with 136 patients receiving relaxin and 103 patients receiving the placebo. One hundred ninety-six patients (115 in the relaxin group, 81 placebo) completed the 24 week trial and make up the group used in this analysis. For purposes of the study, subject data were analyzed irrespective of treatment assignment. As part of the study, each patient completed the following procedures at baseline and at 24 weeks: history and physical examination, skin score, SF-36, HAQ-DI, pulmonary function tests, and routine laboratory investigations. The joints were assessed at the metacarpophalangeal joint (scored as a single joint unit), wrist, and knee bilaterally for presence or absence of swelling and tenderness.

Generic HRQOL measure. The SF-36 is a generic measure of HRQOL (i.e., the concepts are not specific for any age, disease, or treatment group)[10,25]. The SF-36 includes a health transition item and assesses 4 physical health domains as well as 4 mental health domains. The SF-36 scales can be summarized into physical component summary (PCS) and mental component summary (MCS) scores. For clarity, the PCS will be referred to as “physical score” and the MCS as “mental score.” These summary scores are standardized to responses from the US general population, with a mean score fixed at 50 and standard deviation at 10[25]. Each SF-36 scale is scored 0–100, a higher score representing better health.

Musculoskeletal-targeted measure. The HAQ-DI is a self-administered arthritis-targeted measure intended for assessing functional ability in arthritis[11]. The HAQ-DI is a self-administered 20 question instrument that assesses a patient’s level of functional ability and includes questions of fine movements of the upper extremity, locomotor activities of the lower extremity, and activities that involve both upper and lower extremities. In the original HAQ-DI, an additional grade of difficulty was added in patients using assistive/adaptive devices (such as canes or walker). In the current study, the patient responses were not modified for patient use of assistive/adaptive devices. The standard HAQ-DI is determined by summing the highest item score in each of the 8 domains and dividing the sum by 8, which results in a score from 0 (no disability) to 3 (severe disability).

The physician global assessment was assessed at each visit on a 10 cm visual analog scale (VAS) asking the physician to rate the subject’s scleroderma and how it affects him/her today. The patient global assessment was also assessed at each visit on 10 cm VAS asking the patient to rate the severity of his/her disease today. The patient global assessment was administered as part of the scleroderma HAQ[19]. The global assessment is scored 0–100, where a higher score indicates worse overall disease.

The modified Rodnan skin score (referred to here as “skin score”) is a

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

clinical measure of the extent and severity of skin thickening[26,27]. Skin thickening in SSc, in addition to serving as the primary basis for disease classification, is viewed as a clinical surrogate of disease progression[28]. The skin thickening is assessed in 17 body areas: fingers, hands, forearms, arms, feet, legs, and thighs (bilaterally) and face, chest, and abdomen (singly). Each area is scored 0–3, 0 representing normal skin and 3 being severe thickening [range 0 (no thickening) to 51 (severe thickening in all 17 areas)]. Scores > 20 are considered to reflect moderately severe skin thickening.

External criteria for change. To assess if the SF-36 and HAQ-DI were responsive to a meaningful change in the patient's clinical status, we developed 4 different external criteria for classifying subjects as unchanged or as having a clinically meaningful change between Weeks 0 and 24. The 4 different criteria were based on the literature in other arthritides[29] or expert consensus agreement[24]. The cutoffs for these external criteria were based on calculations of the area under the receiver operating characteristic (ROC) curve[30]. The area under the ROC curve can be used to quantify discriminatory ability of the classification rule, in which an index of 0 means absolutely no discrimination, 0.5 means no better than guessing, and an index of 1.0 identifies changed group perfectly. Indices ≥ 0.7 are considered to have acceptable discrimination[31]. A sensitivity analysis with 5% increments was conducted for all 4 external measures for the change (both improvement and worsening) in 3 measures: SF-36 physical and mental scores and HAQ-DI. The cutoff corresponding to the largest area under the ROC curve was selected for the external measures. For example, the cutpoint of 20% in the skin score was tested to see if it might be clinically important. To do this using the ROC approach, one graphs the tradeoff between the sensitivity and specificity of the skin score with a cutpoint of > 20%. This graph has an area under the curve (AUC) summarizing sensitivity and specificity of the skin score using a 20% change versus each of physical score, mental score, and HAQ-DI. The same is done for a skin score change of > 25%, > 30%, > 35%, etc. For each cutpoint the AUC of the ROC curve is calculated. After all skin score differences of interest are modeled, graphed, and AUC are calculated, the AUC are compared. The skin score cutoff with the largest AUC was deemed the best to use for defining a clinically meaningful change in skin score.

If there were 2 cutoffs with similar AUC, then the one comparable to other arthritides was chosen. Interestingly, the cutoffs in clinical measures based on the largest area under ROC curve were comparable to clinically meaningful cutoffs based on the literature in other arthritides[29] or expert consensus agreement[24]. The 4 criteria were (1) change ≥ 30% in the skin score (the primary endpoint of the clinical trial); (2) change ≥ 20% in self-reported patient global assessment (on 0–100 mm VAS)[29]; (3) change ≥ 20% in the self-reported physician-rated global assessment (on 0–100 mm VAS)[29]; and (4) change ≥ 15% in % FVC predicted.

The groups were divided into the "changed" group (with improvement or worsening ≥ 30% in skin score and ≥ 15% in the % FVC predicted) and "unchanged" group (< 30% change in skin score and < 15% change in % FVC predicted). A parallel classification was made based on the patient and physician global assessment groups. For the analysis of responsiveness, direction of change (improvement versus worsening) is distinguished.

Analysis plan. Baseline descriptive statistics for the SF-36 and HAQ-DI scores were calculated including mean score, standard deviation, minimum and maximum score, and percentage at the floor and ceiling. Floor and ceiling effects are the percentages of respondents scoring at the lowest and highest possible scale level, respectively. These effects can influence responsiveness, as they may limit a change of score over time.

Reliability. Reliability was measured by internal consistency, which refers to the extent that different items in an instrument are measuring the same underlying construct of interest. Internal consistency for multi-item scales was estimated using Cronbach's alpha[32]. An alpha ≥ 0.70 is considered satisfactory for group comparisons.

Responsiveness indices. Responsiveness to change was evaluated using the effect size (ES), standardized response mean (SRM_{ch}), and responsiveness statistic (RS)[8]. These indices are ratios of observed change to a measure of variance (also known as signal to noise). For all 3 indices, the numerator is the mean change from the baseline to Week 24 in the changed group and the denominators are the standard deviation at baseline (ES), the standard deviation of change for the changed group (SRM_{ch}), and the standard deviation of change for people who are deemed not to change (RS). To evaluate the magnitude of response in the "unchanged" group, the SRM_{unch} was calculated by dividing mean change from the baseline to week 24 in the unchanged group by the standard deviation of change for the unchanged group.

Cohen's rule-of-thumb for interpreting effect size is that a value of 0.20–0.49 represents a small change, 0.50–0.79 a medium change, and ≥ 0.80 a large change[33-35]. Area under the ROC, as described above, was also calculated.

To evaluate whether the magnitude of responsiveness indices for each external measure between the SF-36 and HAQ-DI were statistically different, an effect size index was calculated for each subject as the "individual change" divided by the standard deviation of the group at baseline. The standard deviation around the individual estimates provided the estimate of the standard error of the effect size. Student's t test was used to calculate the significance of the mean change from baseline to Week 24 in the changed group for the SF-36 and HAQ-DI. A p value < 0.05 was considered statistically significant.

RESULTS

The 24 week trial failed to show an improvement in the primary objective: change in the skin score in the relaxin group (change in skin score –4.03 ± 7.27) versus placebo (change in skin score –3.70 ± 7.11; p = 0.68; negative score indicates an improvement in the skin score at Week 24 compared to baseline)[36-39].

Table 1 describes the baseline characteristics of the study group (n = 196). Patients had severe skin involvement (skin score of 27.3 ± 6.9)[40], marked compromise in their physical health as measured by physical score (1.7 SD below the US general population), and moderate functional disability, with a HAQ-DI of 1.18 ± 0.71[4]. When the baseline demographic data were compared between 196 patients who completed the 24 week study and 36 patients who did not, the noncompleters had a statistically significantly higher skin score (30.2 ± 6.4 vs 27.3 ± 6.9), physician global assessment (57.6 ± 17.0 vs 49.4 ± 20.5), and HAQ-DI (1.6 ± 0.8 vs 1.18 ± 0.7) (p < 0.05); and lower SF-36 physical (29.1 ± 10.4 vs 33.8 ± 10.6) and mental scores (46.1 ± 11 vs 49.8 ± 9.6) of the SF-36 (p < 0.05 for all).

Floor effects were more common in the HAQ-DI (4.5%) than in the SF-36 physical (0%) and mental (0%) scores. There was no ceiling effect for any of the 3 scores.

Reliability

Internal consistency measured by Cronbach's α was adequate for both the SF-36 scales (ranging from 0.76 to 0.93) and the HAQ-DI (range 0.70–0.90).

Change in the SF-36 physical score, mental score, and HAQ-DI in relation to the 4 external measures (Table 2). In the patients who worsened ≥ 30% in their skin score, the change in the HAQ-DI was statistically different (worse) compared to patients with < 30% worsening of the skin

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

Table 1. Baseline information on 196 patients completing the study.

Baseline Variables	Patient Population	Range
Age, yrs*	47.2 ± 10.3	20.0–69.7
Females, %	85.2	
Race, %		
Caucasian	74.0	
African American	13.3	
Hispanic	11.2	
Asian	0.5	
SSc duration**, yrs*	2.20 ± 1.37	0.07–7.78
SF-36 physical score (0–100)*	33.8 ± 10.6	11.6–57.8
SF-36 mental score (0–100)*	49.8 ± 9.6	17.1 –68.3
Patient global assessment (visual analog scale)*	50.7 ± 23.4	0–100
Physician global assessment (visual analog scale)*	49.4 ± 20.5	5–98
HAQ Disability Index (0–3)*	1.18 ± 0.71	0–2.875
Modified Rodnan skin score*	27.3 ± 6.9	16–51
FVC % predicted*	84.9 ± 15.7	42–130
DLCO % predicted*	69.3 ± 21.2	36–147
Swollen joints (out of 6)*	0.91 ± 1.5	0–6
Tender joint count (out of 6)	1.22 ± 1.9	0–6
Right hand extension, mm*	162.34 ± 34.17	12–250
Cutaneous ulcer present, %	25.2	

Physical (Physical Component Summary) and mental (Mental Component Summary) scores have means of 50 and standard deviations of 10 in the US general population, with higher scores indicating better health. HAQ Disability Index is from 0 to 3 with higher score indicating worse functional disability. Patient and physician global assessment is from 0 to 100, with higher score indicating worse overall patient assessment. Modified Rodnan skin score is from 0 to 51, with higher score indicating worse thickening of the skin. * Mean ± SD. ** Duration defined as first non-Raynaud's phenomenon manifestation.

Table 2. Change in the SF-36 physical score, mental score, and HAQ-DI from baseline to 24 weeks in the 4 external measures. A negative score for the physical score and mental score denotes worsening of generic HRQOL; a negative score for the HAQ-DI denotes improvement in functional abilities.

External Measures	Change in Physical Score	p	Change in Mental Score	p	Change in HAQ-DI	p
≥ 30% Worsening in skin score (n = 10)	–2.5 ± 6.2	0.4	1.37 ± 15.6	0.83	0.35 ± 0.51	0.014
< 30% Worsening in skin score (n = 186)	–0.03 ± 8.2		0.24 ± 8.4		0.03 ± 0.40	
≥ 30% Improvement in skin score (n = 72)	–0.65 ± 9.4	0.5	–0.02 ± 8.5	0.7	0.03 ± 0.41	0.76
< 30% Improvement in skin score (n = 124)	1.33 ± 7.3		0.49 ± 9.1		0.05 ± 0.40	
≥ 15% Worsening in % FVC (n = 17)	–2.2 ± 10.2	0.3	–2.84 ± 8.9	0.14	0.28 ± 0.5	0.01
< 15% Worsening in % FVC (n = 176)	0.06 ± 7.9		0.37 ± 8.6		0.02 ± 0.39	
≥ 15% Improvement in % FVC (n = 7)	–0.8 ± 3.8	0.8	–0.4 ± 8.7	0.8	–0.16 ± 0.11	0.04
< 15% Improvement in % FVC (n = 186)	–0.11 ± 8.2		0.12 ± 8.7		0.04 ± 0.41	
≥ 20% Worsening in patient global (n = 72)	–3.95 ± 8.9	< 0.0001	–1.25 ± 8.1	0.06	0.21 ± 0.39	< 0.0001
< 20% Worsening in patient global (n = 124)	2.05 ± 6.8		1.2 ± 9.1		–0.05 ± 0.37	
≥ 20% Improvement in patient global (n = 58)	5.9 ± 5.6	< 0.0001	3.13 ± 8.7	0.003	–0.21 ± 0.34	< 0.0001
< 20% Improvement in patient global (n = 138)	–2.7 ± 7.7		–0.9 ± 8.7		0.15 ± 0.37	
≥ 20% Worsening in physician global (n = 52)	–0.57 ± 6.6	0.5	–1.73 ± 9.14	0.07	0.06 ± 0.4	0.5
< 20% Worsening physician global (n = 136)	0.21 ± 8.2		0.83 ± 8.6		0.02 ± 0.4	
≥ 20% Improvement in physician global (n = 73)	0.26 ± 8.07	0.71	1.91 ± 8.6	0.03	–0.02 ± 0.4	0.15
< 20% Improvement in physician global (n = 115)	–0.19 ± 7.8		–0.01 ± 8.7		0.06 ± 0.4	

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

score (0.35 ± 0.51 vs 0.03 ± 0.40; p = 0.014; Table 2). In comparison, the physical and mental scores were not statistically different in the 2 groups. A similar pattern was seen in the other SSc-specific measure, % FVC predicted, where change (for both worsening and improvement) in HAQ-DI was statistically different from the "unchanged" groups (p < 0.05). All 3 measures (physical score, mental score, and HAQ-DI) separated the ≥ 20% "changed" patient global assessment groups compared to the "unchanged" groups (p < 0.05 for all, except change in the mental score for 20% "improvement" in patient global assessment, p = 0.06). For patients with ≥ 20% improvement in global assessment, the improvement in the HAQ-DI was 0.21 ± 0.34, analogous to the minimal clinically important difference reported in the RA literature[41]. For change ≥ 20% in the physician global assessment of disease activity, only the mental score showed a significant trend for both improvement (p = 0.03) and worsening (p = 0.07) compared to the "unchanged" group.

Responsiveness

The next 2 sections give the results of the responsiveness indices divided into the disease-specific measures (skin score and % FVC predicted) and the global assessments (patient and physician). The data in this section and in Table 3 relate only to total HAQ-DI score and the SF-36 physical and mental scores.

Disease-specific measures

Responsiveness to skin score. An increase (worsening; n = 10) of the skin score at Week 24 was associated with a medium to large magnitude of responsiveness (as determined by the 3 responsiveness indices) to the worsening of skin score in the HAQ-DI (Table 3). The magnitude of responsiveness for the physical score was small (as determined by the 3 responsiveness indices). The SF-36 mental score was not responsive to worsening in skin score (Figure 1, left side, and Table 3). Effect size estimates for the HAQ-DI (0.52 ± 0.76, mean ± SD), physical score (0.25 ± 0.64), and mental score (0.12 ± 1.42) did not differ statistically (p > 0.05).

In patients with an improvement in skin score (n = 72), the magnitude of responsiveness for the physical score, mental score, and HAQ-DI score was negligible (Figure 1, right side; Table 3). Effect size estimates for the HAQ-DI (0.05 ± 0.67, mean ± SD), physical score (0.07 ± 0.98), and mental score (0.0 ± 0.9) did not differ statistically (p > 0.05).

Responsiveness to % FVC predicted. Using % FVC predicted as measure of response (≥ 15% of change), 17 patients worsened, while 7 patients improved (Table 3). When using the worsening of % FVC as a measure, the total HAQ-DI (0.37 ± 0.67) had a numerically larger magnitude of the responsiveness (Figure 1, left side; Table 3) than the physical score (0.24 ± 1.12) and the mental score (0.27 ± 0.84); however, the effect size estimates for the HAQ-DI, physical score, and mental score did not differ statistically (p > 0.05).

When using improvement of the % FVC as a measure, the total HAQ-DI (0.22 ± 0.16) had numerically but not statistically a larger magnitude of the responsiveness than the physical score (0.06 ± 0.29) and the mental score (0.04 ± 0.75; p > 0.05) (Figure 1, right side; Table 3).

Global assessment

Responsiveness to patient global assessment. Using ≥ 20% worsening in patient global assessment (n = 72) as the measure for change, the magnitude of responsiveness for the physical score was larger than for the HAQ-DI total score in

Table 3. Responsiveness of the SF-36 physical and mental scores and the HAQ-DI to change (worsening and improvement) in the 4 external criteria.

	Worsening of the External Measures															
	≥ 30% Worsening in Skin Score (n = 10)				≥ 15% Worsening in % FVC (n = 17)				≥ 20% Worsening in Patient Global Assessment (n = 72)				≥ 20% Worsening in Physician Global Assessment (n = 52)			
Instrument	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC
Physical score	0.25	0.39	0.33	0.59	0.24	0.22	0.27	0.57	0.36	0.45	0.67	0.71	0.05	0.09	0.07	0.53
Mental score	0.12	0.09	0.16	0.46	0.27	0.32	0.33	0.61	0.15	0.15	0.14	0.57	0.20	0.19	0.20	0.55
HAQ-DI	0.52	0.69	0.92	0.69	0.37	0.55	0.30	0.68	0.33	0.54	0.60	0.66	0.08	0.16	0.15	0.52
	Improvement of the External Measures															
	≥ 30% Improvement in Skin Score (n = 72)				≥ 15% Improvement in % FVC (n = 7)				≥ 20% Improvement in Patient Global Assessment (n = 58)				≥ 20% Improvement in Physician Global Assessment (n = 73)			
	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC	ES	SRM_{ch}	RS	ROC
Physical score	0.07	0.07	0.09	0.49	0.06	0.21	0.10	0.57	0.59	1.05	1.00	0.82	0.03	0.03	0.03	0.54
Mental score	0.00	0.00	0.00	0.52	0.04	0.05	0.05	0.51	0.34	0.36	0.34	0.63	0.19	0.22	0.23	0.59
HAQ-DI	0.05	0.08	0.09	0.55	0.22	1.33	0.41	0.65	0.30	0.62	0.60	0.76	0.03	0.05	0.05	0.56

ES (effect size) = D/SD at baseline; SRM_{ch} (standardized response mean) = D/SD of D the "changed" group; RS (responsiveness statistics) = D/SD of D of unchanged, where D = raw score change in the "changed" group; SD: baseline standard deviation; ROC, area under ROC curve; physical score: Physical Component Summary; mental score: Mental Component Summary; HAQ-DI: Health Assessment Questionnaire Disability Index; Skin score: modified Rodnan skin score; FVC: forced vital capacity.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

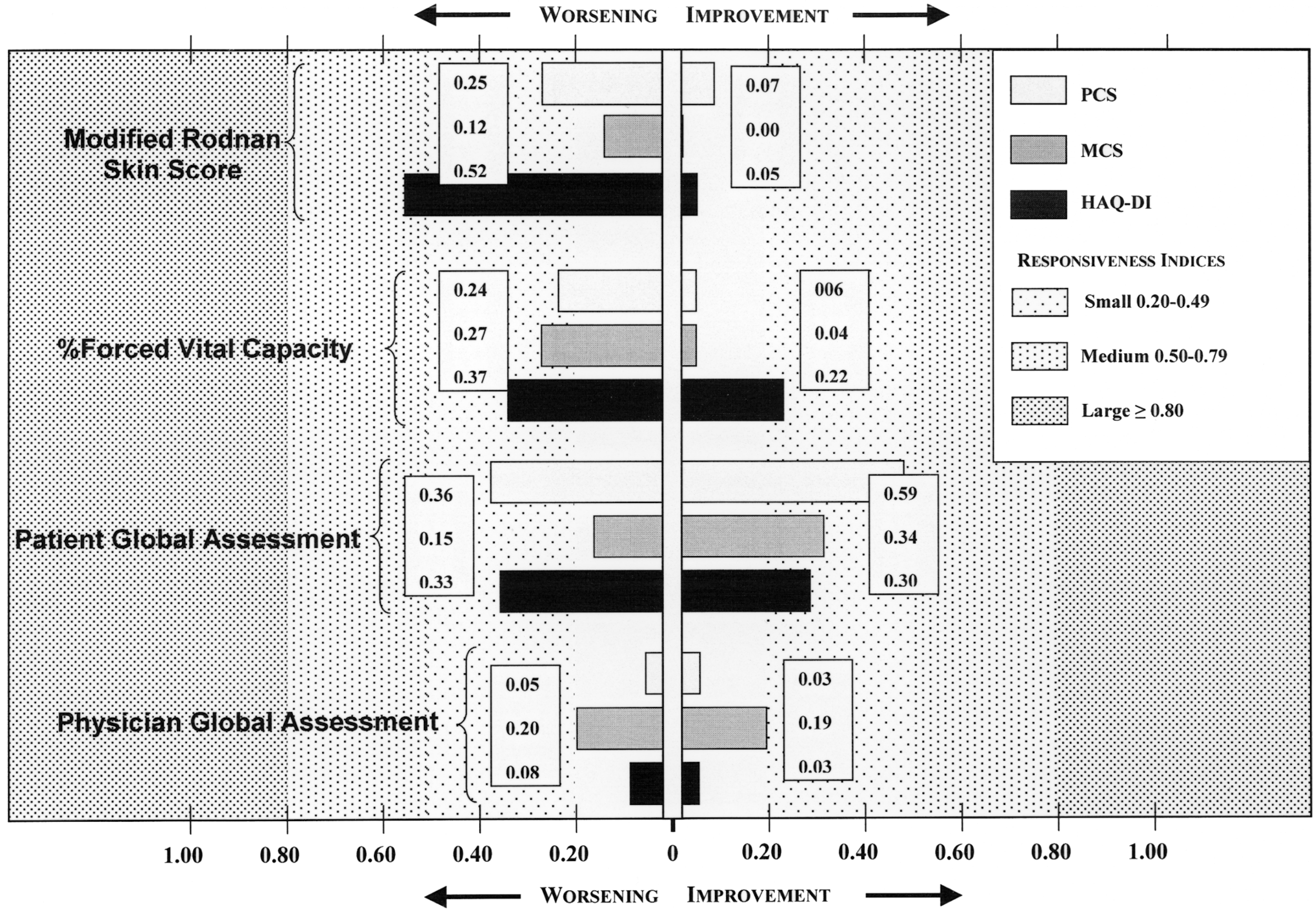

Figure 1. Effect size responsiveness index of the SF-36 physical and mental scores and the HAQ-DI to change (worsening and improvement) in the 4 external criteria.

2 out of 3 indices (Figure 1, left side; Table 3). The mental score was not responsive to change for this subgroup analysis. Effect size estimates for the HAQ-DI (0.33 ± 0.62), physical score (0.36 ± 0.81), and mental score (0.15 ± 0.99) did not differ statistically ($p > 0.05$).

For patients who improved ≥ 20% on their global assessment (n = 58; Table 3), the magnitude of responsiveness tended to be greater for the physical score (as determined by the 3 responsiveness indices) and comparable for the mental score and HAQ-DI (Figure 1, right side; Table 3). The magnitude of responsiveness index as assessed by the individual effect size was statistically different between the physical score (0.59 ± 0.56) and the HAQ-DI (0.30 ± 0.49; $p = 0.003$).

Responsiveness to physician global assessment. For worsening (n = 52) and improvement (n = 73) of the physician global assessment, only mental score showed a small magnitude of responsiveness, as assessed by the 3 responsiveness indices (Figure 1; Table 3). For both worsening and improvement, effect size estimates did not differ significantly between the HAQ-DI, physical score, and mental score ($p > 0.05$).

Relationship of the "changed" group and the "unchanged" group. The responsiveness indices presented so far do not take into account the change over time in the "unchanged" group. For a HRQOL measure to be valid, the change in the "changed" group should be of larger magnitude than the "unchanged" group, in addition to the change being in the same direction as the clinical measure. Table 4 presents a comparison of the SRM magnitude and direction of change in the "changed" and "unchanged" groups. The SRM_{ch} was generally of larger magnitude than SRM_{unch}, the exception being the physical and mental scores in the improvement of the skin score and % FVC predicted.

DISCUSSION

Scleroderma is a chronic disease associated with physical impairment and work disability[4]. The ability of an HRQOL or a functional instrument to detect clinically important change is crucial to their usefulness in determining the effectiveness of drug trials and other therapies on HRQOL[9]. The magnitude of responsiveness, as measured by these indices, is useful as a measure of treatment efficacy and in

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

Table 4. Comparison of the standardized response mean (SRM) in the "changed" and "unchanged" groups to change (worsening and improvement) in the 4 external criteria.

Instrument	≥ 30% Worsening in Skin Score (n = 10)		≥ 15% Worsening in % FVC (n = 17)		≥ 20% Worsening in Patient Global Assessment (n = 72)		≥ 20% Worsening in Physician Global Assessment (n = 52)	
	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$
Physical score	0.39	0.003	0.22	–0.007	0.45	–0.36	0.09	–0.02
Mental score	–0.09	–0.03	0.32	–0.04	0.15	–0.13	0.19	–0.1
HAQ-DI	0.69	0.08	0.55	0.05	0.54	–0.13	0.16	–0.05
	≥ 30% Improvement in Skin Score (n = 72)		≥ 15% Improvement in % FVC (n = 7)		≥ 20% Improvement in Patient Global Assessment (n = 58)		≥ 20% Improvement in Physician Global Assessment (n = 73)	
	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$	SRM$_{ch}$	SRM$_{unch}$
Physical score	–0.07	0.18	–0.21	–0.01	1.05	–0.35	0.03	–0.02
Mental score	–0.002	0.05	–0.05	0.01	0.36	–0.1	0.22	–0.0001
HAQ-DI	–0.08	–0.12	1.33	–0.1	0.62	–0.4	0.05	–0.15

SRM$_{ch}$: D/SD of D the "changed" group, SRM$_{unch}$: D′/SD of D′ the "unchanged" group. D: mean change from baseline to Week 24 in the changed group, D′: mean change from the baseline to Week 24 in the unchanged group. Positive number indicates that SF-36 and the HAQ-DI changed in the same direction (in agreement) as the changes in clinical measures, and negative number indicates that SF-36 and the HAQ-DI changed in the opposite direction (in disagreement).

estimating sample size for future study design[34]. Although a variety of approaches have been proposed to assess responsiveness of an instrument, no consensus has been reached on which is the best[1,34,42,43].

Both the SF-36 (generic) and the HAQ-DI (musculoskeletal-targeted) changed in the same direction as the changes in clinical measures in the "changed" group. This is in contrast to the "unchanged group," where the responsiveness indices had either a smaller size of change compared to the "changed" group or change in the opposite direction than the clinical measures. The SF-36 had a larger magnitude of response to the patient and physician global assessment compared to the HAQ-DI, whereas the HAQ-DI had a larger magnitude of responsiveness in clinical measures (i.e., change in skin score and % FVC predicted) than the SF-36. The area under the ROC curve in Table 3 addresses the ability of the instruments to reveal both change and no change in the external criteria. Although not formally tested, the area under the ROC curve appeared to be comparable to the magnitude of responsiveness indices. A medium magnitude of response (≥ 0.5) in one of the indices was associated with achieving the acceptable 0.7 level of discrimination for the ROC curve analysis. The SF-36 is a generic measure of HRQOL[10,25] that allows comparisons of the relative burden of different diseases and benefits of therapies across normal and various disease cohorts. Another advantage of a generic HRQOL is its ability to measure an unexpected clinical event during a clinical trial, which might be missed by clinical assessment and condition-specific measures. The advantages of the HAQ-DI are (1) the ability to quantify functional impairment associated with SSc, especially to quantify hand dysfunction; and (2) inclusion of unique scales, such as the grip, arising, hygiene, and eating scales of the HAQ-DI, not captured by the SF-36. Being a musculoskeletal-targeted instrument, HAQ-DI had a larger magnitude of response in worsening of the skin score compared to the SF-36 physical and mental scores. On the other hand, for physician global assessment, only the mental score of the SF-36 had small responsiveness indices, suggesting that in this study the physician global assessment was predominantly associated with mental health.

Our study has several strengths. The study included a large sample size of patients with early diffuse SSc and had little missing data at the end of the study (Week 24). Also, the randomized, placebo controlled design is an ideal method to study outcome measure performances.

There were certain limitations of the study. First, since the relaxin was not effective, this analysis cannot test whether the SF-36 or the HAQ-DI are able to discriminate between efficacious or inefficacious drugs. However, given the multiple domains of health measured by these 2 tools, such discriminant capabilities are expected. Additionally, an effective study drug would have resulted in more subjects with improvement, and thus may have revealed more information about the responsiveness of the instruments. The SF-36 and HAQ-DI, nevertheless, can discern change in patient disease status. Second, this clinical trial evaluated only diffuse cutaneous SSc, and the results are not applicable to limited cutaneous SSc. Additional study is required of HRQOL among individuals with limited SSc, although the HAQ-DI has already been validated in limited disease[17]. A continuing randomized-controlled trial in patients with both diffuse and limited SSc with active alveolitis will help answer this question. Third, this trial enrolled only subjects with a modified Rodnan skin score ≥ 20, making the results less generalizable to patients with diffuse disease and skin scores < 20.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

Fourth, we studied patients for only 6 months. It may well be that analysis of longer duration trials would be more or less sensitive for disease improvement.

This study provides support for the reliability and validity of the SF-36 and HAQ-DI in diffuse SSc over 6 months. The SF-36 had a larger magnitude of responsiveness in detecting the changes in both patient and physician global assessment compared to the HAQ-DI, while the HAQ-DI had a larger magnitude of responsiveness in detecting changes in skin and pulmonary disease associated with the SSc than the SF-36. The 2 instruments were complementary to each other in providing valuable information in this study population. The data support inclusion of both the SF-36 and HAQ-DI as outcome measures in future clinical trials of diffuse SSc.

REFERENCES

1. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622-9.
2. Medsger TA Jr. Systemic sclerosis: clinical aspects. In: Koopman W, editor. Arthritis and allied conditions. Baltimore: Williams and Wilkins; 1997:1433-64.
3. Clements PJ, Wong WK, Hurwitz EL, et al. Correlates of the Disability Index of the Health Assessment Questionnaire: a measure of functional impairment in systemic sclerosis. Arthritis Rheum 1999;42:2372-80.
4. Clements PJ, Wong WK, Hurwitz EL, et al. The Disability Index of the Health Assessment Questionnaire is a predictor and correlate of outcome in the high-dose versus low-dose penicillamine in systemic sclerosis trial. Arthritis Rheum 2001;44:653-61.
5. Reveille JD, Fischbach M, McNearney T, et al. Systemic sclerosis in 3 US ethnic groups: a comparison of clinical, sociodemographic, serologic, and immunogenetic determinants. Semin Arthritis Rheum 2001;30:332-46.
6. Khanna D, Clements PJ, Furst DE, et al. Correlation of the degree of dyspnea with health related quality of life, functional abilities, patient global assessment and decreased diffusing capacity in systemic sclerosis patients with active alveolitis: result from scleroderma lung study. Arthritis Rheum 2005;52:592-600.
7. Cossutta R, Zeni S, Soldi A, Colombelli P, Belotti MA, Fantini F. Evaluation of quality of life in patients with systemic sclerosis by administering the SF-36 questionnaire [Italian]. Reumatismo 2002;54:122-7.
8. Hays RD, Hadorn D. Responsiveness to change: an aspect of validity, not a separate dimension. Qual Life Res 1992;1:73-5.
9. Salaffi F, Stancati A, Carotti M. Responsiveness of health status measures and utility-based methods in patients with rheumatoid arthritis. Clin Rheumatol 2002;21:478-87.
10. 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.
11. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
12. Bell MJ, Bombardier C, Tugwell P. Measurement of functional status, quality of life, and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591-601.
13. Rood MJ, Borggreve SE, Huizinga TW. Sensitivity to change of the MOS SF-36 quality of life assessment questionnaire in patients with systemic lupus erythematosus taking immunosuppressive therapy. J Rheumatol 2000;27:2057-9.
14. Davies GM, Watson DJ, Bellamy N. Comparison of the responsiveness and relative effect size of the Western Ontario and McMaster Universities Osteoarthritis Index and the short-form Medical Outcomes Study survey in a randomized, clinical trial of osteoarthritis patients. Arthritis Care Res 1999;12:172-9.
15. Hawley DJ, Wolfe F. Sensitivity to change of the Health Assessment Questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. Arthritis Care Res 1992;5:130-6.
16. Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF. The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis. J Rheumatol 1988;15:1480-8.
17. Merkel PA, Herlyn K, Martin RW, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410-20.
18. Poole J, Steen V. The use of the Health Assessment Questionnaire to determine physical disability in systemic sclerosis. Arthritis Care Res 1991;4:27-31.
19. 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.
20. Del Rosso A, Boldrini M, D'Agostino D, et al. Health-related quality of life in systemic sclerosis as measured by the Short Form 36: relationship with clinical and biologic markers. Arthritis Rheum 2004;51:475-81.
21. Georges C, Chassany O, Mouthon L, et al. Quality of life assessment with the MOS-SF36 in patients with systemic sclerosis [French]. Rev Med Interne 2004;25:16-21.
22. Preliminary criteria for the classification of systemic sclerosis (scleroderma). Subcommittee for Scleroderma Criteria of the American Rheumatism Association Diagnostic and Therapeutic Criteria Committee. Arthritis Rheum 1980;23:581-90.
23. Clements PJ, Hurwitz EL, Wong WK, 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.
24. Seibold JR, McCloskey DA. Skin involvement as a relevant outcome measure in clinical trials of systemic sclerosis. Curr Opin Rheumatol 1997;9:571-5.
25. Ware JE, Kosinki M, Keller S. SF-36 physical and mental health summary scales: a user's manual. Boston: Health Institute, New England Medical Center; 1994.
26. Clements P, Lachenbruch P, Siebold J, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol 1995;22:1281-5.
27. Merkel PA, Clements PJ, Reveille JD, Suarez-Almazor ME, Valentini G, Furst DE. Current status of outcome measure development for clinical trials in systemic sclerosis. Report from OMERACT 6. J Rheumatol 2003;30:1630-47.
28. Black CM. Measurement of skin involvement in scleroderma. J Rheumatol 1995;22:1217-9.
29. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
30. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chron Dis 1986;39:897-906.
31. Hosmer D, Lemeshow S. Applied logistic regression. 2nd ed. New York: John Wiley & Sons; 2000.
32. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrica 1951;16:297-334.
33. Cohen J. A power primer. Psychol Bull 1992;112:155-9.
34. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org

1990;28:632-42.

35. Kim S, Hays RD, Birbeck GL, Vickrey BG. Responsiveness of the quality of life in epilepsy inventory (QOLIE-89) in an antiepileptic drug trial. Qual Life Res 2003;12:147-55.

36. Seibold J, Clements P, Korn JH, et al. U.S. phase III trial of relaxin in diffuse scleroderma [abstract]. J Rheumatol 2001;28 Suppl 63:55.

37. Seibold J. Relaxin: Lessons and limitations. Curr Rheumatol Rep 2002;4:275-6.

38. Erikson M, Unemore E. Relaxin clinical trials in systemic sclerosis. In: Tregear GW, Ivell I, Bathgate RA, Wade JD, editors. Relaxin. Dordrecht: Kluwer Academic Publishers; 2000: 373-81.

39. Seibold JR, Korn JH, Simms R, et al. Recombinant human relaxin in the treatment of scleroderma. A randomized, double-blind, placebo-controlled trial. Ann Intern Med 2000;132:871-9.

40. Medsger TA Jr, Silman AJ, Steen VD, et al. A disease severity scale for systemic sclerosis: development and testing. J Rheumatol 1999;26:2159-67.

41. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557-60.

42. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171-8.

43. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27 Suppl:S178-89.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2005. All rights reserved.

Downloaded on April 10, 2024 from www.jrheum.org