Published in final edited form as:
*Arthritis Rheum*. 2008 May 15; 59(5): . doi:10.1002/art.23564.

# Validity, Reliability, and Feasibility of Durometer Measurements of Scleroderma Skin Disease in a Multicenter Treatment Trial

**PETER A. MERKEL, MD, MPH**[1], **NANCY P. SILLIMAN, PhD**[2], **CHRISTOPHER P. DENTON, MD, FRCP**[3], **DANIEL E. FURST, MD**[4], **DINESH KHANNA, MD, MSc**[4], **PAUL EMERY, MD, FRCP**[5], **VIVIEN M. HSU, MD**[6], **JAMES B. STREISAND, MD**[2], **RICHARD P. POLISSON, MD, MHS**[2], **ANITA ÅKESSON, MD, PhD**[7], **JOHN COPPOCK, MB, FRCP**[8], **FRANK van den HOOGEN, MD, PhD**[9], **ARIANE HERRICK, MD, FRCP**[10], **MAUREEN D. MAYES, MD, MPH**[11], **DOUGLAS VEALE, MD, FRCPI**[12], **JAMES R. SEIBOLD, MD**[6], **CAROL M. BLACK, MD, FRCP, DBE**[3], **JOSEPH H. KORN**[†], **CAT-192 RESEARCH GROUP**, and **SCLERODERMA CLINICAL TRIALS CONSORTIUM**

[1]Boston University, Boston, Massachusetts [2]Genzyme, Cam bridge, Massachusetts [3]Royal Free Hospital, London, UK [4]University of California, Los Angeles [5]University of Leeds, Leeds, UK [6]University of Medicine and Dentistry of New Jersey, New Brunswick [7]Lund University Hospital, Lund, Sweden [8]University Hospital, Coventry, UK [9]Radboud University Medical Center, Nijmegen, The Netherlands [10]University of Manchester, Hope Hospital, Salford, UK [11]University of Texas Medical School, Houston [12]Saint Vincent's University Hospital, Dublin, Ireland

## Abstract

**Objective—**To determine the validity, reliability, and feasibility of durometer measurements of skin hardness as an outcome measure in clinical trials of scleroderma.

**Methods—**Skin hardness was measured during a multicenter treatment trial for scleroderma using handheld digital durometers with a continuous scale. Skin thickness was measured by modified Rodnan skin score (MRSS). Other outcome data collected included the Scleroderma Health Assessment Questionnaire. In a reliability exercise in advance of the trial, 9 investigators examined the same 5 scleroderma patients by MRSS and durometry.

**Results—**Forty-three patients with early diffuse cutaneous systemic sclerosis were studied at 11 international centers (mean age 49 years [range 24–76], median disease duration 6.4 months



Address correspondence to Peter A. Merkel, MD, MPH, Boston University School of Medicine, 715 Albany Street, E533, Boston, MA 02118. pmerkel@bu.edu..

Dr. Silliman owns stock and/or stock options in Genzyme. Dr. Denton has received consultant fees, speaking fees, and honoraria (less than $10,000 each) from Actelion and Encysive. Dr. Furst has received consultant fees (less than $10,000 each) from Genzyme, Actelion, and Gilead. Drs. Streisand and Polisson own stock and stock options in Genzyme. Dr. Mayes has received consultant fees (less than $10,000) from Novartis and honoraria (less than $10,000 each) from Actelion and Encysive. Dr. Veale has received consultant fees, speaking fees, and honoraria (less than $10,000 each) from Schering-Plough, Wyeth, and Glaxo-SmithKline.

[†]Dr. Korn is deceased.

**AUTHOR CONTRIBUTIONS** Dr. Merkel 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 design.** Merkel, Silliman, Denton, Furst, Streisand, Polisson, van den Hoogen, Mayes, Veale, Seibold, Black, Korn.

**Acquisition of data.** Merkel, Denton, Furst, Khanna, Emery, Hsu, Streisand, Polisson, Åkesson, Coppock, van den Hoogen, Herrick, Mayes, Veale, Seibold, Black, Korn.

**Analysis and interpretation of data.** Merkel, Silliman, Denton, Furst, Streisand, Polisson, van den Hoogen, Mayes, Veale, Sei-bold.

**Manuscript preparation.** Merkel, Silliman, Denton, Furst, Khanna, Emery, Hsu, Streisand, Polisson, van den Hoogen, Mayes, Veale, Seibold.

**Statistical analysis.** Merkel, Silliman.

[range 0.3–23], and median baseline MRSS 22 [range 11–38]). The reliability of durometer measurements was excellent, with high interobserver intraclass correlation coefficients (ICCs) (0.82–0.92), and each result was greater than the corresponding skin site ICCs for MRSS (0.54–0.85). Baseline durometer scores correlated well with MRSS (r = 0.69, $P < 0.0001$), patient self-assessments of skin disease (r = 0.69, $P < 0.0001$), and Health Assessment Questionnaire (HAQ) disability scores (r = 0.34, $P = 0.03$). Change in durometer scores correlated with change in MRSS (r = 0.70, $P < 0.0001$), change in patient self-assessments of skin disease (r = 0.52, $P = 0.003$), and change in HAQ disability scores (r = 0.42, $P = 0.017$). The effect size was greater for durometry than for MRSS or patient self-assessment.

**Conclusion—**Durometer measurements of skin hardness in patients with scleroderma are reliable, simple, accurate, demonstrate good sensitivity to change compared with traditional skin scoring, and reflect patients' self-assessments of their disease. Durometer measurements are valid, objective, and scalable, and should be considered for use as a complementary outcome measure to skin scoring in clinical trials of scleroderma.

## INTRODUCTION

Manual skin scoring is a validated and widely applied method of measuring the extent and severity of skin disease in patients with scleroderma, and the modified Rodnan skin score (MRSS) has become the standard primary outcome measure of skin involvement in clinical trials in scleroderma (1–3). However, additional methods to assess sclerodermatous skin that are more precise, objective, and reproducible would be valuable for use in clinical trials of scleroderma. Additionally, although skin thickness, the property measured by the MRSS, is an important aspect of scleroderma skin disease, hardness and elasticity are separate but partially related characteristics that may also have important meanings in the disease process and merit quantification.

Durometers are handheld devices used in manufacturing to measure the hardness of materials with internationally standardized durometer units. Previously published work from a single center demonstrated that durometer measurements are reliable and valid measures of skin disease in patients with scleroderma and correlate well with MRSS (4), ultrasound-measured skin thickness (4), skin hyalinized collagen content (5), and skin myofibroblast score (5). Full validation of durometers as an outcome tool in scleroderma still requires demonstration of feasibility and reliability in the setting of a clinical trial with multiple investigators, sensitivity to change, and correlation with other disease measures.

This study was designed to determine the validity, reliability, sensitivity to change, and feasibility of durometer measurements of skin hardness as an outcome measure for scleroderma when compared with skin scoring and patient self-assessment tools in the setting of a therapeutic clinical trial.

## MATERIALS AND METHODS

### Study setting and subjects

The durometer was tested during the conduct of a multicenter, randomized, double-blind, placebo-controlled treatment trial of early diffuse cutaneous systemic sclerosis (dcSSc; scleroderma). Details of the therapeutic agent, study design, study subjects, and primary results of the trial have been previously reported (6). The trial was conducted at 11 scleroderma centers in the US and Europe.

### Investigator training and reliability exercise

All investigators were trained in durometer and MRSS methodologies using real patients with dcSSc. To ensure uniformity of technique, investigators and their scores were compared with a single standard investigator experienced in the use of the durometer and performing skin scoring in clinical trials of scleroderma. Durometry training took <15 minutes.

After training, a reliability exercise was conducted. Five patients with dcSSc were evaluated; investigators performed durometer measurements and skin scoring on each subject and then repeated the measurements on 2 subjects.

### Skin thickness scores

MRSS were calculated using the standard 0, 1, 2, 3 integer scale at 17 body sites (fingers, hands, forearms, arms, face, chest, abdomen, thighs, legs, feet). The overall thickness was assessed by an investigator at each body site. Easily identified landmark sites were also established for the 6 body sites where durometer measurements were taken: forearms, thighs, and legs/calves. The landmark site for forearms was the dorsal aspect of forearms, midway between the radial head/styloid and lateral epicondyle; the landmark site for thighs was the dorsal aspects of the thigh, midway between the midpoint of femoral crease and superior pole of the patella; and the landmark site for legs/calves was midway between the midpoint of superior aspect of calcaneus and the midpoint of inferior aspect of popliteal fossa. The overall measurements were used to compare the 17-site MRSS with a durometer. Landmark measurements were used to compare a 6-site MRSS with a durometer, the 6 sites being the forearms, thighs, and calves.

### Durometer skin hardness scores

Durometer measurements were made using a handheld digital durometer (Rex Gauge Type OO; Rex Gauge Durometer, Buffalo Grove, IL) with a continuous scale measured in standard durometer units (4). Three consecutive durometer measurements were taken at each of the 6 landmark sites described above (forearms, thighs, and legs/calves); the sum of the means of the 3 measurements at each site was used for analyses. These sites were chosen because they were not over bony prominences and were easily standardized.

### Other outcome measures

All subjects completed the Scleroderma Health Assessment Questionnaire, which includes the Health Assessment Questionnaire (HAQ) disability scale and visual analog scales (VAS), including 1 for patient self-assessment of skin disease (3,7–9).

### Statistical analysis

Reliability was measured by intra-class correlation coefficients (ICCs). Correlations were measured by Pearson's method; the associated *P* values were calculated using the F test from a simple linear regression.

To compare the repeat assessments for the durometer and whether they differed, a repeated measures random-effects model was used (10). Specifically, the model was individual durometer score = intercept + slope × order of measurement, where order of measurement (1st measurement, 2nd measurement, 3rd measurement) was a random effect (i.e., the trend/slope was allowed to vary by patient).

To determine the most important predictors of change in durometer scores at month 6, backwards stepwise linear regression was used (11). The initial model included all

covariates (treatment group, investigator site in US or Europe, sex, baseline age, body mass index, disease duration, MRSS, durometer score, patient self-assessment of skin disease, and HAQ disability index [DI]), and then the least significant term was removed and the model was rerun until the remaining covariates were all significant at the 0.05 level.

The sensitivities to change of MRSS, durometry, HAQ, and patient self-assessment of skin disease were compared using both effect sizes and standardized response means (SRMs). Effect size was defined as the mean change from baseline to month 6 for each measure divided by the baseline SD of that measure (12,13). SRM was defined as the mean change from baseline to month 6 divided by the SD of the change from baseline to month 6 (14).

Two-sided tests of significance with an alpha level of 0.05 were used for all calculations. Analyses were performed using S-PLUS, version 6.2 (Insightful, Seattle, WA).

## RESULTS

Data from 43 subjects (33 women, 10 men) were included in this analysis. The subjects had a mean age of 49 years (range 24–76 years), median disease duration of 6.4 months (range 0.3–23 months), median baseline MRSS of 22 (range 11–38), median baseline durometer score of 218 (range 127–330), median baseline HAQ DI of 1.1 (range 0–2.6), and median baseline patient self-assessment VAS of skin disease of 32 (range 5–91).

Durometer measurements were highly reliable and reproducible at all sites (Table 1 and Figure 1). Interrater reliability for durometry ranged from 0.819–0.913 at the 3 body areas and was 0.919 for the 6-site total. The interrater reliability was higher for durometer measurements than for MRSS at each of the skin sites for which both measures were recorded and for the 6-site total (ICC 0.919 versus 0.844).

The 3 repeated durometer measurements at individual skin sites were generally comparable (Figure 1). When using repeated measures random-effects models, a significant trend toward lower values with repeated measurements was detected at 4 of the 6 skin sites; however, none of these average differences were considered clinically relevant (~1 point on an average score of 35–40).

Durometer scores (continuous) within a given individual site's skin score (integer) ranged widely, indicating durometry may provide a greater dynamic range than MRSS (Figure 2). For example, right forearm durometer scores were 34.7–35.0 for an MRSS of 0, 22.2–46.5 for an MRSS of 1, 24.4–64.1 for an MRSS of 2, and 30.5–63.8 for an MRSS of 3.

Durometer measurements were significantly correlated with MRSS at each skin site: thighs ($r = 0.51$, $P < 0.0001$), forearms ($r = 0.48$, $P < 0.0001$), and calves ($r = 0.32$, $P = 0.003$). The correlation between total 6-site durometry scores and the total (17-site) MRSS was high ($r = 0.69$, $P < 0.0001$) (Figure 3A). The correlation between 6-site durometry and the 6-site MRSS was also significant, although not quite as strong ($r = 0.52$, $P = 0.0006$). Durometer scores (6-site total) also correlated with other outcome measures, including patient self-assessment of skin disease ($r = 0.69$, $P < 0.0001$) (Figure 3B) and HAQ disability scores ($r = 0.34$, $P = 0.03$) (Figure 3C).

Change in durometer scores was highly correlated with change in MRSS ($r = 0.70$, $P < 0.0001$) (Figure 4A). This sensitivity to change in durometry versus MRSS was observed both in patients with improving skin disease and patients with worsening skin disease. Change in durometer scores was also highly correlated with change in the 6-site MRSS ($r = 0.70$, $P < 0.0001$). Still significant, but not quite as strong of a relationship as with MRSS, change in durometer scores was correlated with change in patient self-assessment of skin

disease (r = 0.52, *P* = 0.003) (Figure 4B) and change in HAQ disability scores (r = 0.42, *P* = 0.017) (Figure 4C).

Backwards stepwise regression was used to find the most important predictors of change in durometer score at month 6. The initial model included treatment group, investigator site in US versus Europe, sex, and baseline age, body mass index, disease duration, MRSS, durometer score, patient self-assessment of skin disease, and HAQ DI. The final model included disease duration at baseline (*P* = 0.004) and sex (*P* = 0.04). Over 6 months, durometer scores tended to increase for patients with shorter disease duration and decrease for patients with longer disease duration (r = −0.47, *P* = 0.006), similar to findings for MRSS in this study (r = −0.54, *P* = 0.001). Durometer scores increased slightly for women (median score 2.6, n = 25) and decreased for men (median score −19.9, n = 7).

The effect sizes and SRMs for durometry, MRSS, HAQ, and patient self-assessment of skin disease are shown in Table 2. These data demonstrate that durometry is more sensitive to change in skin disease than either MRSS or patient self-assessments, and of similar sensitivity to change as the HAQ disability scale. Furthermore, the finding that the SRMs are nearly identical to the effect sizes indicates that the modest changes seen in these measures are fairly easy to detect.

## DISCUSSION

This study further demonstrates that durometer measurements of skin hardness in patients with scleroderma are reliable, simple, accurate, sensitive to change, and should be considered for use as an outcome measure in clinical trials of dcSSc.

Durometry compared well with skin scoring, with good correlation to both site-specific scoring and total MRSS. Durometry may offer an increased range of values for skin assessment in patients with scleroderma compared with semiquantitative MRSS; specifically, limiting skin scoring to 4 integer values reduces the ability of the measure to record either improvement or worsening that may be clinically important. Patients and investigators often report that they perceive changes in skin disease occurring at specific body sites even when the MRSS for that site does not change. Therefore, the seemingly wider range of durometer scores and their scalability imply a capacity to detect small or moderate changes in skin disease; increased precision in skin assessment may be particularly important in clinical trials of investigational agents.

The demonstration in this study of sensitivity to change of durometry is a critical aspect of validating this tool for use in clinical trials. Change in durometer readings correlated well with change in the MRSS, and durometer measurements also correlated well with patient ratings of skin disease and the HAQ DI. Furthermore, the effect size and SRM, measures of relative magnitude and ease of detecting change, were substantially greater for durometry than for MRSS or patient ratings of skin disease. Therefore, this study considerably extends the available data on the validity of durometer measurements as an outcome measure in the study of dcSSc. This study confirms the reliability, face validity, content validity, and discriminant validity of durometry (4,5,15,16). Additionally, to our knowledge this study demonstrated for the first time the utility of durometry in the setting of a multicenter clinical trial; prior work in use of the durometer in dcSSc was all conducted at single centers (4,5,15,16).

This study further demonstrates that durometry is highly reliable and easy to learn. The lower rate of inter-rater variation compared with skin scoring indicates that durometry provides a means for objective measurement of skin disease in patients with scleroderma

even when different investigators examine study subjects and different trial visits. Investigators found the durometer easy to use after only minimal training. Additionally, durometry provides a good means of assessing changes in skin disease at a single skin site.

Although durometer scores correlated well with skin scores and patient-derived measures of skin disease and disability, the differences between durometry and these measures are notable and imply that durometry is capturing similar, but not exactly the same, domains of disease as these other measures. Therefore, in addition to its apparent increased precision, durometry may provide additional information on the disease course in scleroderma. In summary, durometer measurements are valid, objective, and scalable, and should be considered for use as a complementary outcome measure to skin scoring and functional assessments in clinical trials of scleroderma.

## Acknowledgments

The authors thank their many coinvestigators at each site, the site study coordinators and other site personnel, the Genzyme clinical trial team, and most importantly our volunteer subjects.

**ROLE OF THE STUDY SPONSOR** Personnel from Genzyme Corporation were directly involved in the study design, data collection, data analysis, and the editing of the manuscript. Genzyme Corporation approved the content of the manuscript and agreed to submit the manuscript for publication.

Supported by Genzyme, the Scleroderma Foundation, and the National Center for Research Resources (NIH) General Clinical Research Centers program at Boston University (grant M01-RR0-00533). Dr. Merkel's work was supported by a Mid-Career Clinical Investigator Award (National Institute of Arthritis and Musculoskeletal and Skin Diseases, grant K24-AR2224-01A1). Dr. Khanna's work was supported by the Scleroderma Foundation and an NIH BIRCWH Award (HD051953).

## REFERENCES

1. Clements PJ, Lachenbruch PA, Ng SC, Simmons M, Sterz M, Furst DE. Skin score: a semiquantitative measure of cutaneous involvement that improves prediction of prognosis in systemic sclerosis. Arthritis Rheum. 1990; 33:1256–63. [PubMed: 2390128]
2. 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. [PubMed: 7562759]
3. 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. [PubMed: 12858472]
4. Kissin EY, Schiller AM, Gelbard RB, Anderson JJ, Falanga V, Simms RW, et al. Durometry for the assessment of skin disease in systemic sclerosis. Arthritis Rheum. 2006; 55:603–9. [PubMed: 16874783]
5. Kissin EY, Merkel PA, Lafyatis R. Myofibroblasts and hyalinized collagen as markers of skin disease in systemic sclerosis. Arthritis Rheum. 2006; 54:3655–60. [PubMed: 17075814]
6. Denton CP, Merkel PA, Furst DE, Khanna D, Emery P, Hsu VM, et al. on behalf of the CAT-192 Study Group; Scleroderma Clinical Trials Consortium. Recombinant human anti–transforming growth factor β1 antibody therapy in systemic sclerosis: a multicenter, randomized, placebo-controlled phase I/II trial of CAT-192. Arthritis Rheum. 2007; 56:323–33. [PubMed: 17195236]
7. Poole JL, Steen VD. The use of the Health Assessment Questionnaire (HAQ) to determine physical disability in systemic schlerosis. Arthritis Care Res. 1991; 4:27–31. [PubMed: 11188583]
8. 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. [PubMed: 9365087]
9. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Scleroderma Clinical Trials Consortium. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum. 2002; 46:2410–20. [PubMed: 12355489]

10. Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied longitudinal analysis. Wiley; New York: 2004.
11. Rawlings, JO. Applied regression analysis: a research tool. Wadsworth; Belmont (CA): 1988.
12. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989; 27(3 Suppl):S178–89. [PubMed: 2646488]
13. 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. [PubMed: 1457487]
14. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum. 1985; 28:542–7. [PubMed: 4004963]
15. Falanga V, Bucalo B. Use of a durometer to assess skin hardness. J Am Acad Dermatol. 1993; 29:47–51. [PubMed: 8315077]
16. Aghassi D, Monoson T, Braverman I. Reproducible measurements to quantify cutaneous involvement in scleroderma. Arch Dermatol. 1995; 131:1160–6. [PubMed: 7574833]

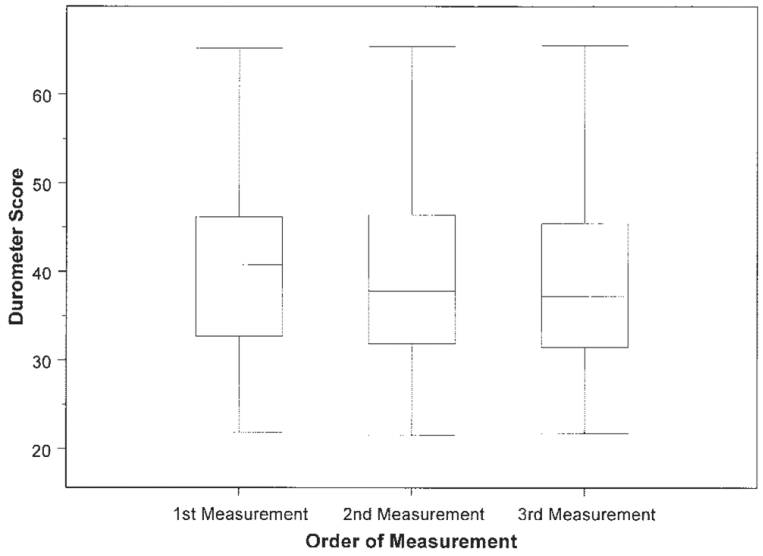


**Figure 1.**

Repeated durometer measurements at the right forearm taken at the baseline visit demonstrate no clinically meaningful variation among the 3 repeated measurements. Box plots represent the 25th and 75th percentiles, lines inside the boxes represent the median, and whiskers represent the minimum and maximum values.

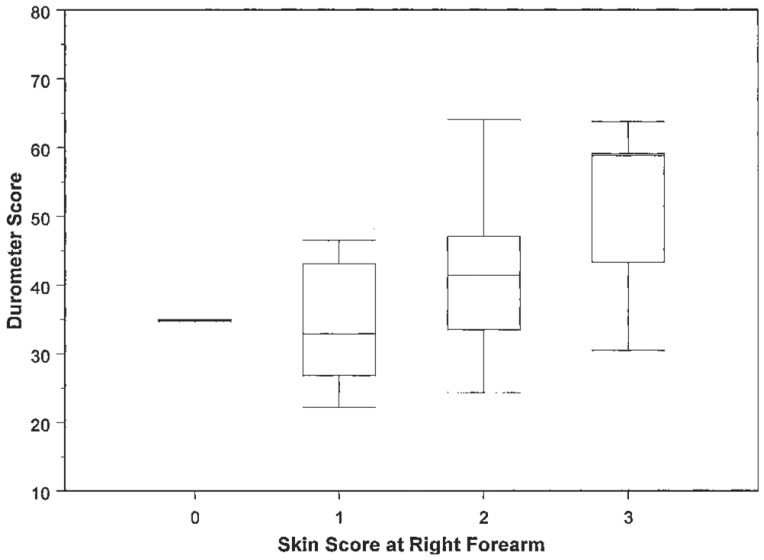


**Figure 2.**
Durometer scores within a given skin score ranged widely, indicating durometry may provide a greater dynamic range than the modified Rodnan skin score at baseline. Box plots represent the 25th and 75th percentiles, lines inside the boxes represent the median, and whiskers represent the minimum and maximum values.

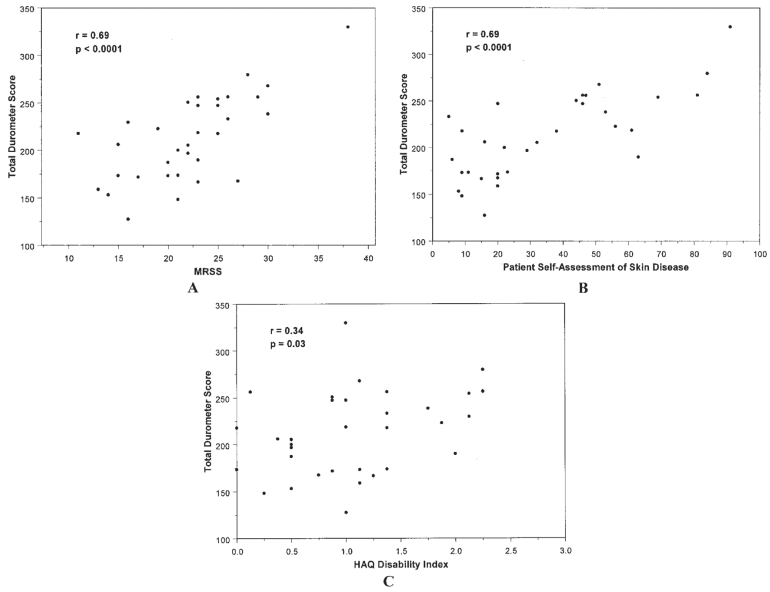


**Figure 3.**

**A**, Correlation between the total (17-site) modified Rodnan skin score (MRSS) and the 6-site total durometer score for subjects at baseline trial visit. **B**, Correlation between total 6-site durometer scores and patients' self-assessment of their skin disease measured by a visual analog scale (scored 0–100) at baseline trial visit. **C**, Correlation between total 6-site durometer scores and patients' disability index of the Health Assessment Questionnaire (HAQ) at baseline trial visit, which is included in the Scleroderma Health Assessment Questionnaire.

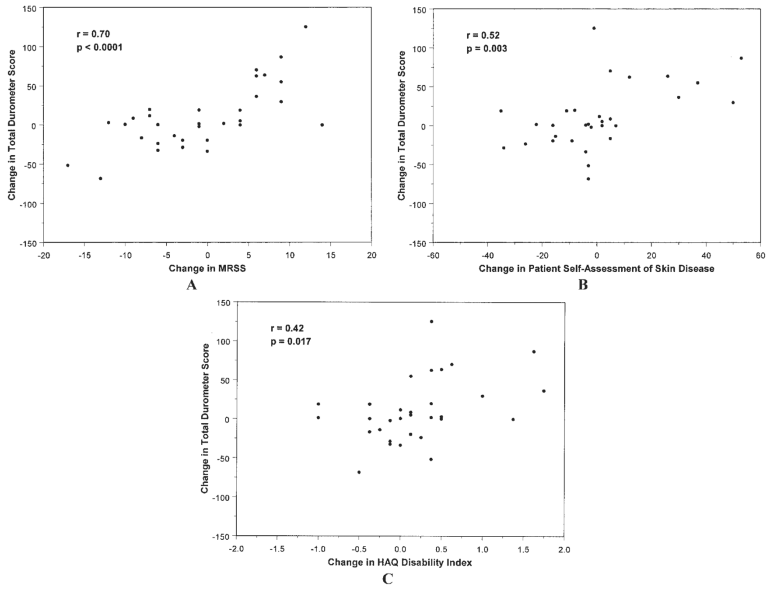


**Figure 4.**

**A**, Correlation between the change in the total 6-site durometer score and changes in the total 17-site modified Rodnan skin score (MRSS) during the treatment phase of the trial. **B**, Correlation between change at month 6 in total 6-site durometer scores and change at month 6 in patients' self-assessment of their skin disease as measured by a visual analog scale (scored 0–100). **C**, Correlation between total 6-site durometer scores and the patients' disability index of the Health Assessment Questionnaire (HAQ), which is included in the Scleroderma Health Assessment Questionnaire. Both endpoints are measured as the change in score from baseline to month 6.

**Table 1**

Reliability of durometer measurements*

| | Durometer score | MRSS | Durometer score vs. MRSS† | *P* |
|---|---|---|---|---|
| Forearms | 0.894 | 0.821 | 0.75 | < 0.0001 |
| Thighs | 0.819 | 0.542 | 0.38 | < 0.0005 |
| Calves | 0.913 | 0.845 | 0.79 | < 0.0001 |
| 6-site total | 0.919 | 0.844 | 0.80 | < 0.001 |

*Values are the interobserver intraclass correlation unless indicated otherwise. MRSS = modified Rodnan skin score.

†By Spearman's correlation.

## Table 2

Effect sizes and SRMs of durometer score, skin score, HAQ disability scores, and patient self-assessment of skin disease*

| Measurement | Baseline | Change from baseline to month 6 | Effect size | SRM |
|---|---|---|---|---|
| Durometer score (total) | 212.70 ± 44.71 | 9.39 ± 40.88 | 0.21 | 0.23 |
| MRSS | 22.18 ± 5.63 | −0.35 ± 7.83 | −0.06 | −0.04 |
| HAQ disability score | 1.06 ± 0.64 | 0.19 ± 0.61 | 0.30 | 0.31 |
| Patient self-assessment of skin disease | 34.61 ± 24.61 | 1.41 ± 22.17 | 0.06 | 0.06 |

*Values are mean ± SD unless otherwise indicated. SRM = standardized response mean; HAQ = Health Assessment Questionnaire; MRSS = modified Rodnan skin score.