# Patterns and Predictors of Change in Outcome Measures in Clinical Trials in Scleroderma An Individual Patient Meta-Analysis of 629 Subjects with Diffuse Scleroderma

**PA Merkel**[1], **NP Silliman**[2], **PJ Clements**[3], **CP Denton**[4], **DE Furst**[3], **MD Mayes**[5], **JE Pope**[6], **RP Polisson**[2], **JB Streisand**[2], and **JR Seibold**[7], and **the Scleroderma Clinical Trials Consortium**

[1]Boston University, Boston, MA

[2]Genzyme Corporation, Cambridge, MA

[3]University of California, Los Angeles, CA

[4]Royal Free Hospital, London, United Kingdom

[5]University of Texas, Houston, TX

[6]University of Western Ontario, London, Canada

[7]University of Connecticut, Farmington, CT

## Abstract

**Purpose**—To examine the range and responsiveness to change of clinical outcome measures and study the predictors of clinical response for patients with diffuse cutaneous systemic sclerosis (dcSSc) in the context of clinical trials.

**Methods**—Data from 629 patients with dcSSc who participated in 7 multicenter clinical therapeutic trials were combined. Trials used common outcome measures: modified Rodnan skin score (MRSS), the Health Assessment Questionnaire (HAQ), Patient Global Assessment (PtGA), pulmonary function tests (FVC, DLCO), and oral aperture (OAp).

**Results**—The combined database included 629 patients: 82% women; mean age = $46.5 \pm 11.8$ years (range 15–82) with disease duration (months): mean: $19.4 \pm 15.9$, median = 47.0, range 1.0–144.0.

Outcomes tended to improve during trials for patients with more severe disease at study entry and worsen for patients with less severe disease at entry. There were weak negative correlations between baseline values and change over 6 months for MRSS ($r = -0.17$; p<.0001), HAQ ($r = -0.15$; p= .002), and PtGA ($r = -0.44$; p<.0001). Baseline FVC and OAp did not predict change in 6 months. Baseline DLCO values were positively correlated with change in DLCO at 6 months ($r = -0.32$; p<.0001).

Disease duration was mildly negatively predictive of change in MRSS at 6 months ($r = -0.27$; p<.0001) and substantial bidirectional variation in change in MRSS and HAQ was seen over the spectrum of disease duration.

63% of patients with "early" disease (<18 months) had a decline in MRSS and 37% had an increase in MRSS. 81% of patients with late disease ( 18 months) had a decline in MRSS and 19% had an increase in MRSS. 53% of patients with early disease had a decline in HAQ and 47%

Corresponding Author: Peter A. Merkel, MD, MPH, Division of Rheumatology, University of Pennsylvania, 8th Floor Penn Tower, 3400 Spruce Street, Philadelphia, PA 19104, Tel: 215-614-4401, Fax: 215-614-4402, pmerkel@upenn.edu.

had an increase in HAQ. 51% of patients with late disease had a decline in HAQ and 49% had an increase in HAQ. Multivariate mixed models did not demonstrate that any baseline variables were strongly predictive of subsequent outcome.

These results did not differ when comparing trials of early vs. late disease or trial "completers" vs. "non-completers".

**Conclusions—**Among patients with dcSSc enrolled in clinical trials, standard outcome measures tend to improve for patients with more severe disease at study entry and worsen for patients with less severe disease at entry. Overall, MRSS scores improve during observation periods while HAQ and lung function are mostly static, although there are wide variations in individual changes in these measures. None of these variables, including disease duration, reliably identify groups of subjects whose MRSS will predictably increase or decrease in the course of a clinical trial. These findings have important implications for clinical trial design in scleroderma.

### Keywords

Scleroderma; outcome measures; clinical trials; meta-analysis

## INTRODUCTION

A number of large randomized placebo controlled studies of candidate therapies for systemic sclerosis (SSc, scleroderma) have been performed over the past 15 years. The scientific rationale for the choice of interventions has been diverse. Investigative therapies included agents with non-specific antifibrotic effects (D-penicillamine (1), recombinant human relaxin (2–3)), agents with immunosuppressant and immunomodulatory actions (interferon-$\alpha$ (4), methotrexate (5)), agents targeted to disease-specific molecular mechanisms (monoclonal antibody to TGF-$\beta$1 (6)), and an agent tested in follow-up to empiric observations (minocycline (7)). None of these therapies were demonstrated to be effective. While this failure to demonstrate therapeutic efficacy may be attributable to a lack of clear understanding of molecular and cellular pathways in scleroderma pathogenesis, it remains possible that the overall approach to trial design, including choice of patient populations and outcome measures, is insufficiently sophisticated to permit reliable detection of therapeutic effects.

Definition of the natural history of diffuse cutaneous systemic sclerosis (dcSSc) is critical in considering the design of interventional clinical trials for this disease. The population of subjects with scleroderma enrolled in clinical trials is be different from the overall population of people with scleroderma for several reasons, including: i) the sample sizes in trials are small and thus limit the range of clinical characteristics represented; ii) the duration of trials are usually less than one year which is quite short for a chronic disease in which important clinical changes occur over periods of several years; iii) research subjects often enroll in trials during periods of active, and usually progressive, skin disease, states that do not reflect the full spectrum of disease activity in the overall population of patients with scleroderma; and iv) in some instances patients with the most severe disease were treated by physicians outside the trial.

The utility and responsiveness of currently used outcome measures in treatment trials of scleroderma remains uncertain (8–11). The modified Rodnan skin score (MRSS) is a validated measure of skin disease and has become the most commonly used measure in clinical trials. The MRSS correlates with patient-derived measures of disease, physical function, and mortality(8, 12–13). However, there are several limitations to the MRSS including the high inter-observer variation, question as to the measure's sensitivity to change, and the need to take into consideration the natural history of scleroderma skin

disease. Similarly, the ability of other validated outcome measures, including the modified Health Assessment Questionnaire (HAQ)(14–18) and patient and physician global assessments, to measure substantive change in a trial setting has not been fully demonstrated for scleroderma.

The development, validation, and evaluation of usefulness of outcome measures for skin and other manifestations of scleroderma have been problematic for several related reasons. Although there has been a substantial increase in the number, size, and quality of therapeutic clinical trials in scleroderma over the past decade, most of these trials, especially those evaluating treatments for skin disease, have been "negative" studies. The absence of clearly therapy-related change makes evaluation of the sensitivity to change of measure challenging. The natural history of scleroderma is also highly variable and often includes an early period of rapid worsening of skin disease followed by either stabilization or even spontaneous improvement. Additionally, the multisystem nature of the disease makes evaluation of an overall treatment effect difficult.

To better understand the range, responsiveness, and other test characteristics of the MRSS, and other key outcomes, in the context of clinical trials, an individual patient data meta-analysis was conducted using the pooled data from 7 recently completed clinical trials of scleroderma that utilized identical outcome measures.

## METHODS

### Study Overview

This study involved combining the originally compiled data from the individual patients with SSc studied in 7 major clinical therapeutic trials of scleroderma conducted in the past 10 years.

### Data Sources and Trial Characteristics

Data from 7 multicenter therapeutic clinical trials in dcSSC that used mostly the same set of outcome measures were included in the analysis (1–7). All participating investigators had undergone at least one formal training session for key clinical measures such as skin scoring, although not all investigators participated in all 7 trials. The details of the demographics, trial designs, treatment arms are outlined in Table 1. The 7 trials enrolled a total of 635 subjects, of whom 629 were randomized and included in this analysis. These trials took place from 1996–2004 and included 6 randomized double-blind, placebo-controlled trials and 1 open-label, single-arm trial.

### Outcome Measures of Interest Common to Included Trials

In each trial the primary outcome measure was skin score but several other measures were collected and were common to the included studies.

- Modified Rodnan skin score (MRSS):
  - The standard, 17-site MRSS was used in 6 of the 7 trials. The seventh trial used a 26-site MRSS, but the 17-site MRSS was calculated for this trial for use in this analysis. Each skin site was scored on an integer scale of 0–3 for skin thickness (19).

- Disability Index of the Health Assessment Questionnaire (HAQ-DI):
  - This standard patient-completed disability index includes 20 questions in 8 functional domains and yields a composite score that is transformed into 0–3 continuous total score (20).

- Patient global assessment (PtGA), collected in 5 trials:

  – This was a visual analog scale transformed for this analysis into a 0–10 continuous score.

- Pulmonary function tests.

  Standardized testing was conducted with values reported as percent predicted, adjusted for sex and body size. The measures used in this analysis included:

  – The forced vital capacity (FVC), collected in 6 trials.

  – The diffusion capacity (DLCO), collected in 6 trials.

- Functional assessments:

  – Hand span, collected in 4 trials: The mean of three consecutive measurements of the span achieved by patients when asked to maximally open/spread their hand as measured from the distal end of digit 1 to the distal end of digit 5 (21).

  – Oral aperture, collected in 5 trials: The mean of three consecutive measurements from the lower edge of the upper lip to the upper edge of the lower lip when patients are asked maximally open their mouth (21).

### Data Compilation

When possible, data was transferred electronically as originally entered for the initial analyses by the study investigators. Some new data entry was performed for study variables from original source documents. All patient identifiers were removed prior to transfer of any data and compilation. The data was merged into SAS datasets.

### Statistical Analyses

All patients enrolled in one of the 7 trials were included in the analysis. The amount of follow-up information varied by study. All follow-up information was included in the relevant analyses. As none of the trials were able to establish a treatment effect, the study population was viewed as an observational, natural history cohort. Homogeneity of baseline demographic and disease characteristics across treatment groups within a trial, and across trials, were examined for continuous variables using analysis of variance with trial and treatment-within-trial (a nested effect) terms in the model. Similarly, homogeneity for categorical variables was evaluated using a linear model of generalized logits with trial and treatment-within-trial terms in the model. Characteristics that differed at baseline were included in models of change over time for the various study outcome measures of interest, as were effects for trial and treatment-within-trial.

For the various outcome measures of interest, descriptive summary statistics of change from baseline at the various timepoints were calculated per treatment group and trial, as well as overall. Duration of disease was calculated as the number of months between the onset of first non-Raynauds symptom and the baseline trial visit. Tests for whether the change from baseline were significant were conducted using one-sample t-tests for normally distributed changes and the Wilcoxon signed-rank test for non-normally distributed changes.

Box plots of change from baseline were used to visually examine differences between the trials. Longitudinal plots of individual patient data with an average curve generated using the Lowess function were used to examine the overall magnitude and pattern of change. Scatter plots and correlations were used to examine whether various baseline measures were predictive of change from baseline at months 6 and 12. Mixed-effects models for repeated

measures were used to examine predictors of change over time using a multivariable approach. Trial, treatment group within trial, disease duration, and the baseline value were always included in the models. In addition, to account for baseline differences, covariates that were not balanced across trials were included in models.

## RESULTS

### Summary of Outcomes Data

**Study Population—**Data was available for 629 subjects randomized and enrolled in 7 clinical trials. Key outcomes data was available for almost all subjects at the six month visit for each study and at the 12 month visit for a substantial subset of subjects. The dataset is remarkably complete for the variables presented in this analysis. Data for the main outcome measures of interest (MRSS, HAQ, PtGA, FVC, and DLCO) were available for almost all patients at the baseline visit (95%, 89%, 68%, 87%, and 92% of treated patients, respectively), and the majority of patients through 6 months of each study (80%, 70%, 54%, 60%, and 41% of treated patients, respectively). Due to variable lengths of the trials, data for the outcomes at 12 months were available for fewer patients (22%, 19%, 9%, 15%, and 19% of treated patients, respectively).

**Baseline Clinical Characteristics—**Table 1 includes summary data on the demographics of the study populations and baseline values for the major clinical outcome measures. Similarities in study subject characteristics across all 7 studies reflect the population of patients with scleroderma including a mean age of 46.5 years and a female predominance (82%). All endpoints and covariates were balanced across treatment groups *within* a study. However, there was a broad distribution of values for baseline outcomes and not all variables were balanced *across* studies. The following were balanced *across* studies: race, sex, FVC, HAQ. The following were not balanced *across* studies: age (p-value<0.001): disease duration (p-value<0.001), patient global (p-value=0.008), oral aperture (p-value<0.001), hand extension (p-value<0.001), MRSS (p-value<0.001), and DLCO (p-value<0.001).

The most striking difference in baseline variables across the studies was for disease duration. This finding was expected given the eligibility criteria for the various studies. Disease duration across studies ranged from 8.4 ± 5.1 months (mean ± standard deviation) with a range of 1–23 months for the anti-TGFβ trial to 32.0 ± 20.2 months, range 7–144 months for the Phase II study of relaxin. The total combined cohort had a mean disease duration of 19.4 ± 15.9 months with a range of 1–144 months.

The primary outcome variable for all 7 trials was the MRSS. The baseline values for this measure varied across the studies from a low mean of 21.1 ±8.0 in the D-penicillamine study to high mean of 28.3 ±10.1 for the interferon-α study. The frequency distribution of baseline MRSS for the combined cohort is displayed in Figure 1a.

The HAQ disability scores were quite similar across all 7 studies and reflect a moderately high level of mean baseline disability with means scores 1.1 ±0.71. As indicated by the relatively large coefficient of variation, 0.65 (=0.71/1.1), HAQ scores were distributed fairly widely across almost the entire 0–3 range. The frequency distribution of baseline HAQ scores for the combined cohort is displayed in Figure 2a. The correlation between baseline MRSS and baseline HAQ was 0.38 (p<0.001).

**Change in Scleroderma Outcomes During Clinical Trials—**Most outcomes improved or stabilized over the course of the clinical trials. Table 2 outlines the change in the major outcome variables during the course of clinical trials for the combined dataset.

**Change in Skin Scores:** There was a statistically significant overall improvement in total skin score during the course of the clinical trials from a mean baseline of 25.1 to a score of 22.2 at 6 months and a score of 21.7 at 12 months; the improvement trend continued through year 2 of follow-up (Figure 1b and Table 2). Nonetheless, the range of change in MRSS was wide with a near-normal distribution (Figure 1c).

**Change in HAQ Scores:** There is a slight downward trend (improvement) in HAQ disability scores over time during clinical trials, these changes did reached statistical significance at month 6 but not at month 12 (Figure 2b and Table 2). The range of change in HAQ scores was wide with a near-normal distribution (Figure 2c). The correlation between change from baseline to 6 months in MRSS versus change from baseline to 6 months HAQ was 0.31 (p<0.001).

**Change in Patient Global:** The average self-reported ratings of Patient Global Assessment did not significantly change over the course of clinical trials although, as with the other measures, there was wide person-to-person variation (Table 2).

**Change in Pulmonary Function Tests:** The trend for all pulmonary functions test parameters was for no change or minimal improvement over the course of clinical trials (Supplementary Figure 4 available and Table 2). This was true of both measures of parenchymal compliance (FVC) and gas exchange (diffusing capacity).

**Change in Physical Function Measures:** The trend for all physical functions measure was for no change or minimal improvement over the course of clinical trials (Table 2); the left hand span data was comparable to the right hand span data (not shown).

**Predictors of Change in Scleroderma Outcomes During Trials—**Change in several outcome measures from baseline to month 6 correlated with baseline of that endpoint (Table 3). Outcomes tended to improve among patients with more severe disease at study entry and worsen for patients with less severe disease at entry.

**Correlation Between Disease Duration and Outcomes—**The relationships between disease duration and outcome measures were examined in detail. The average baseline MRSS rose slightly but steadily with longer disease duration, peaking at 40 months (Figure 3a). However; there were wide ranges of worsening and improvement seen among patients at all strata of disease duration (Figure 3c). When the cohort is divided into subgroups based on "early" disease (disease duration < 18 months) and "late" disease (disease duration 18 months), it is evident that worsening of skin scores occurred more frequently among the patients with early disease than among patients with late disease (p < 0.001). Among patients with early disease, 63% had a decline in MRSS and 37% had an increase in MRSS. Among patients with late disease, 81% had a decline in MRSS and 19% had an increase in MRSS.

HAQ disability scores also rose slightly with longer disease duration, peaking at approximately 38 months (Figure 3b). However, there was no substantial difference in the proportion of patients with worsening HAQ scores between patients with early or late disease (Figure 3d): patients with early disease, 53% had a decline in HAQ and 47% had an increase in HAQ. Among patients with late disease, 51% had a decline in HAQ and 49% had an increase in HAQ.

**Multivariate Models of Predictors of Change in MRSS:** Using mixed-effects models for repeated measures and thus incorporating all observed data, the only significant predictors of

change in MRSS over time were i) disease duration (negative effect; negative correlation) and ii) baseline MRSS (negative effect; negative correlation). No other variables or measures reliably contributed to prediction of change in MRSS. Additionally, no combination criteria of baseline MRSS or disease duration selected for subjects with predictable decline or improvement.

**Effects of Specific Trial Designs on Outcomes: Targeting Early vs. Late Disease**—When data from the 3 trials with longer average disease duration at baseline were removed (2–3, 7), _none_ of the presented results changed in any significant manner including correlations, change data, and baseline predictors of later outcomes (data not shown).

**Effect of Trial Completion on Outcomes**—Compared to subjects who did not complete trials, those who completed the trials had i) shorter disease duration; ii) lower baseline MRSS and HAQ scores; and iii) higher baseline FVC and DLCO values. However: the trajectory of changes in these outcome measures was similar for completers and non-completers (data not shown).

## DISCUSSION

The current study analyzed a large set of pooled individual patient data collected during therapeutic clinical trials in scleroderma that utilized a common set of outcome measures. These data lead to several important conclusions. For patients with dcSSc enrolled in clinical trials, standard outcome measures tend to improve for patients with more severe disease at study entry and worsen for patients with less severe disease at entry, demonstrating a pattern of regression to the mean. Additionally, overall, MRSS scores improve during observation periods while HAQ, physical function measures, and lung function are mostly static, although there are wide variations in individual changes in these measures. Importantly, no variables reliably identify groups of subjects whose MRSS will predictably increase or decrease in the course of a clinical trial. The best predictor of increase in MRSS is early disease status, however this is neither a sensitive nor specific measure.

Analysis of the measure of skin disease among patients enrolled in clinical trials for SSc provides some important insights. Over the course of even a short trial, MRSS changes in most subjects, and both substantial improvement and worsening is observed. However, the average skin score among patients with scleroderma enrolled in clinical trials significantly improves from baseline values at 6 and 12 months, the most common length of therapeutic trials, with the trend continuing through month 24.

The relationship between disease duration and MRSS during clinical trials in scleroderma is intriguing. Baseline skin scores are, on average, higher for subjects with longer disease duration, up to approximately 40 months at which point the average MRSS levels off. Disease duration at baseline is only modestly negatively correlated with change in MRSS (r = −0.27). Overall, during clinical trials; patients with scleroderma and shorter disease duration tend to have lower skin scores that worsen during the course of trials while subjects with longer disease duration tend to higher skin scores that improve. However, there is a great deal of variability such that disease duration does not perform well as a means of selecting patients likely to worsen during a trial. Patients with early disease (< 18 months) are more likely to have worsening in MRSS over the first 6 months of a trial than patients with late disease ( 18 months) but there is still a substantial proportion of patients with early disease whose MRSS improves. In contrast, the portion of patients with late disease whose MRSS worsens is relatively small.

The HAQ scores within the combined dataset demonstrate that patients enrolled in clinical trials of scleroderma have substantial functional impairment at the time of randomization. However, HAQ scores for the study population do not change substantially at 6 or 12 months, although there is marked individual variation. For the first 3 years of disease, the longer the baseline disease duration, the greater the disability scores; after about 38 months disease duration, the scores level off or modestly fall. Unlike the situation with skin disease, there is no difference in the change in HAQ scores at 6 months for patients with early vs. late disease. It is noteworthy that the physical measures of function (oral aperture and hand extension) also demonstrated little change in the first 12 months of a clinical trial. These results provide support for the development of validated patient-reported scleroderma-specific function measures for use in clinical trials.

Pulmonary function among patients with scleroderma who participate in clinical trials aimed at treatment of skin disease remains fairly static (FVC) or even spontaneously improves (DLCo) through the course of the trial. This observation was seen for both patients with early and late disease at baseline. Compared to MRSS and HAQ, there was also little individual change in the pulmonary function measures across the study population. It should be emphasized that none of the trials utilized in this meta-analysis were focused on pulmonary outcomes. Recent data suggest that progressivity of pulmonary disease can be predicted by measures of extent of disease by high resolution CT (22–23). Such data were not available for the present analysis.

The findings of this study have important implications for clinical trial design in scleroderma. Both the similarities and differences in patients enrolled in clinical trials compared to patients followed in longitudinal cohorts merit consideration. For many years, the predominant approach to study design in scleroderma was that of the "prevention paradigm". With this approach, patients with "early" disease (<18 months) are preferentially recruited for clinical trials with the expectation that if untreated, such subjects' disease manifestations, especially skin thickening, will likely progress during the course of a short trial. However, the data presented in this study questions the efficiency and utility of recruiting patients with "early" disease in order to enrich a study population and indicates that the prevention paradigm is problematic. Many patients with early disease enrolled in clinical trials have improvement or stabilization of major clinical outcomes. Additionally, the trial-level data is consistent with the patient-level data: disease expression and progression (change) data from trials focusing on "early" patients was not different from the data from the "late" trials. Thus, while the overall pattern of change in MRSS in clinical trials is consistent what has been reported in longitudinal cohorts, the amount of bidirectional change seen and tendency towards regression to the mean all lead to the conclusion that even if patients with "early" disease are preferentially selected, a substantial portion of the study population will have improvement in MRSS.

This study failed to find either single variables or combinations of clinical variables that reliably predict which patients will have progressive disease during a clinical trial in scleroderma. While baseline values of most major endpoints correlate with change in that outcome, these predictors are insufficient to select subjects. These data also highlight the need for placebo-controlled trials given the trend towards improvement overall, even in early disease, of most outcomes of interest. The finding that outcomes were similar among trial "completers" vs 'non-completers" implies that patients who drop out of trials do so for reasons not systematically related to outcomes of interest; for example, people whose disease progressed more rapidly weren't more likely to drop out of the studies. These data also provide important information for future sample size (power) calculations by supplying the descriptive baseline parameters including means and standard deviations for several major outcomes in a large group of patients actually enrolled in trials.

This study has several important strengths. The sample size of 629 treated patients is the largest group of patients with scleroderma enrolled in therapeutic clinical trials ever studied. The patients were all evaluated at academic centers in several countries, each center with expertise and experience in the evaluation and management of dcSSC. Each of the trials included in this analysis assessed the major clinical domains of interest using the same set of validated outcome measures and did so within the context of rigid protocols and the same set schedules. The combined study population encompasses a broad range of patients in terms of important demographics and disease-related variables including age, disease duration, and baseline values of the key outcome measures of interest. Furthermore, the ability to combine individual level patient data is a substantially more powerful study design than traditional meta-analysis that only utilizes published summary data. Additionally, the dataset was quite complete through six months of follow up, with data at 12 months available for a smaller subset of patients.

There are also some limitations to this study to consider. That all the clinical trials included in this study were "negative" somewhat limits the ability to measure change in the outcomes; however, this same characteristic also allows for combined use of data from both the placebo and experimental arms of the trials. Moreover, even with more than 600 patients, some clinical subsets of dcSSC and additional potentially important variables were not evaluable. Nonetheless, the results in this analysis are broadly applicable to the type of patients routinely considered for inclusion in trials of new therapies for scleroderma.

While these conclusions challenge some of the previously applied design principles, the data should help guide researchers to develop more sophisticated studies of promising new agents for scleroderma. Consideration might be given to additional methods of patient selection beyond baseline MRSS and disease duration. Factors of potential interest include use of scleroderma-specific autoantibodies which have prognostic value. Similarly, there are multiple biomarkers under development that attempt to provide both quantification of disease activity and prognostic information using tests related to the underlying pathophysiology of scleroderma. Work on a composite outcome measure for scleroderma may obviate some of the limitations of organ-specific measures currently in use (24).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

## References

1. Clements PJ, Furst DE, Wong WK, Mayes M, White B, Wigley F, et al. High-dose versus low-dose D-penicillamine in early diffuse systemic sclerosis: analysis of a two-year, double-blind, randomized, controlled clinical trial. Arthritis Rheum. 1999 Jun; 42(6):1194–203. [PubMed: 10366112]

2. Khanna D, Clements PJ, Furst DE, Korn JH, Ellman M, Rothfield N, et al. Recombinant human relaxin in the treatment of systemic sclerosis with diffuse cutaneous involvement: a randomized, double-blind, placebo-controlled trial. Arthritis Rheum. 2009 Apr; 60(4):1102–11. [PubMed: 19333948]

3. Seibold JR, Korn JH, Simms R, Clements PJ, Moreland LW, Mayes MD, et al. Recombinant human relaxin in the treatment of scleroderma. A randomized, double-blind, placebo-controlled trial. Ann Intern Med. 2000 Jun 6; 132(11):871–9. [PubMed: 10836913]

4. Black CM, Silman AJ, Herrick AI, Denton CP, Wilson H, Newman J, et al. Interferon-alpha does not improve outcome at one year in patients with diffuse cutaneous scleroderma: results of a randomized, double-blind, placebo-controlled trial. Arthritis Rheum. 1999 Feb; 42(2):299–305. [PubMed: 10025924]

5. Pope JE, Bellamy N, Seibold JR, Baron M, Ellman M, Carette S, et al. A randomized, controlled trial of methotrexate versus placebo in early diffuse scleroderma. Arthritis Rheum. 2001 Jun; 44(6): 1351–8. [PubMed: 11407694]

6. Denton CP, Merkel PA, Furst DE, Khanna D, Emery P, Hsu VM, et al. Recombinant human anti-transforming growth factor beta1 antibody therapy in systemic sclerosis: a multicenter, randomized, placebo-controlled phase I/II trial of CAT-192. Arthritis Rheum. 2007 Jan; 56(1):323–33. [PubMed: 17195236]

7. Mayes MD, O'Donnell D, Rothfield NF, Csuka ME. Minocycline is not effective in systemic sclerosis: results of an open-label multicenter trial. Arthritis Rheum. 2004 Feb; 50(2):553–7. [PubMed: 14872498]

8. 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 Jul; 30(7):1630–47. [PubMed: 12858472]

9. Furst DE, Khanna D, Mattucci-Cerinic M, Silman AJ, Merkel PA, Foeldvari I. Scleroderma-- developing measures of response. J Rheumatol. 2005 Dec; 32(12):2477–80. [PubMed: 16331791]

10. Furst D, Khanna D, Matucci-Cerinic M, Clements P, Steen V, Pope J, et al. Systemic sclerosis - continuing progress in developing clinical measures of response. J Rheumatol. 2007 May; 34(5): 1194–200. [PubMed: 17477486]

11. Khanna D, Merkel PA. Outcome measures in systemic sclerosis: an update on instruments and current research. Curr Rheumatol Rep. 2007 May; 9(2):151–7. [PubMed: 17502046]

12. Clements PJ, Hurwitz EL, Wong WK, Seibold JR, Mayes M, White B, et al. Skin thickness score as a predictor and correlate of outcome in systemic sclerosis: high-dose versus low-dose penicillamine trial. Arthritis Rheum. 2000 Nov; 43(11):2445–54. [PubMed: 11083267]

13. Steen VD, Medsger TA Jr. Improvement in skin thickening in systemic sclerosis associated with improved survival. Arthritis Rheum. 2001 Dec; 44(12):2828–35. [PubMed: 11762943]

14. Poole JL, Steen VD. The use of the Health Assessment Questionnaire (HAQ) to determine physical disability in systemic sclerosis. Arthritis Care Res. 1991 Mar; 4(1):27–31. [PubMed: 11188583]

15. 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 Nov; 40(11):1984–91. [PubMed: 9365087]

16. Clements PJ, Wong WK, Hurwitz EL, Furst DE, Mayes M, White B, 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 Mar; 44(3):653–61. [PubMed: 11263780]

17. Merkel PA, Herlyn K, Martin RW, Anderson JJ, Mayes MD, Bell P, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum. 2002 Sep; 46(9):2410–20. [PubMed: 12355489]

18. Khanna D, Furst DE, Clements PJ, Park GS, Hays RD, Yoon J, et al. Responsiveness of the SF-36 and the Health Assessment Questionnaire Disability Index in a systemic sclerosis clinical trial. J Rheumatol. 2005 May; 32(5):832–40. [PubMed: 15868618]

19. Clements P, Lachenbruch P, Siebold 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 Jul; 22(7):1281–5. [PubMed: 7562759]

20. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. J Rheumatol. 1982 Sep-Oct;9(5):789–93. [PubMed: 7175852]

21. Furst DE, Clements PJ, Harris R, Ross M, Levy J, Paulus HE. Measurement of clinical change in progressive systemic sclerosis: a 1 year double-blind placebo-controlled trial of N-acetylcysteine. Ann Rheum Dis. 1979 Aug; 38(4):356–61. [PubMed: 386962]

22. Tashkin DP, Elashoff R, Clements PJ, Goldin J, Roth MD, Furst DE, et al. Cyclophosphamide versus placebo in scleroderma lung disease. N Engl J Med. 2006 Jun 22; 354(25):2655–66. [PubMed: 16790698]

23. Goh NS, Desai SR, Veeraraghavan S, Hansell DM, Copley SJ, Maher TM, et al. Interstitial lung disease in systemic sclerosis: a simple staging system. Am J Respir Crit Care Med. 2008 Jun 1; 177(11):1248–54. [PubMed: 18369202]

24. Khanna D, Lovell DJ, Giannini E, Clements PJ, Merkel PA, Seibold JR, et al. Development of a provisional core set of response measures for clinical trials of systemic sclerosis. Ann Rheum Dis. 2008 May; 67(5):703–9. [PubMed: 17893248]

## Appendix

Will include compilation of contributing authors/investigators to individual trials. Nature of list depends on preference for in-text or on-line appendix.

**1a. MRSS at Baseline (all subjects)**

**1b. Change in Individual Skin Scores Over Time**

The average line is produced by calculating the average value at each timepoint, and smoothing this curve using a Lowess Function

**1c. Distribution of Change in MRSS From Baseline to 6 Months**
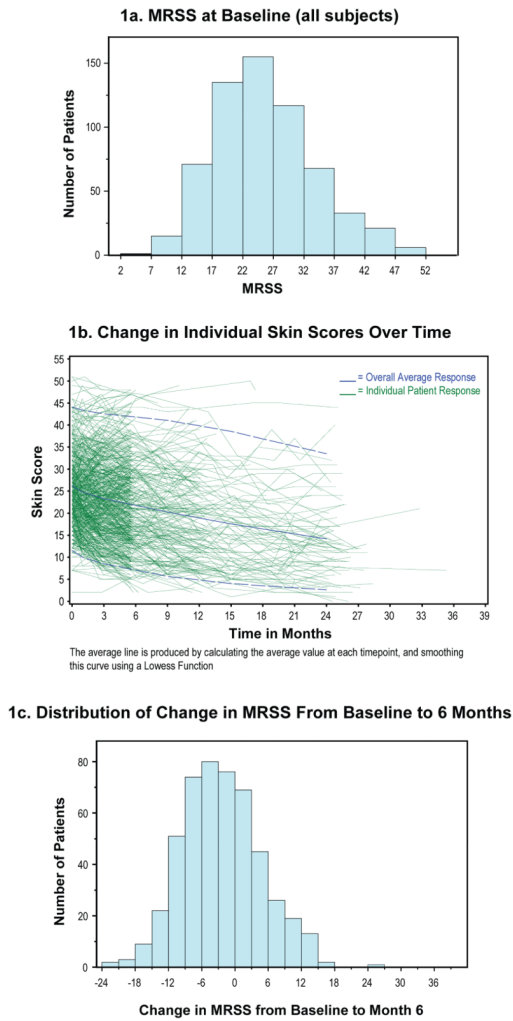
**Figure 1.**
Skin Scores at Baseline and Change During Clinical Trials
1a. MRSS at Baseline (all subjects)
1b. Change in Individual Skin Scores Over Time
1c. Distribution of Change in MRSS From Baseline to 6 Months

**2a. HAQ at Baseline (all subjects)**

**2b. Change in Individual HAQ Scores Over Time**

The average line is produced by calculating the average value at each timepoint, and smoothing this curve using a Lowess Function

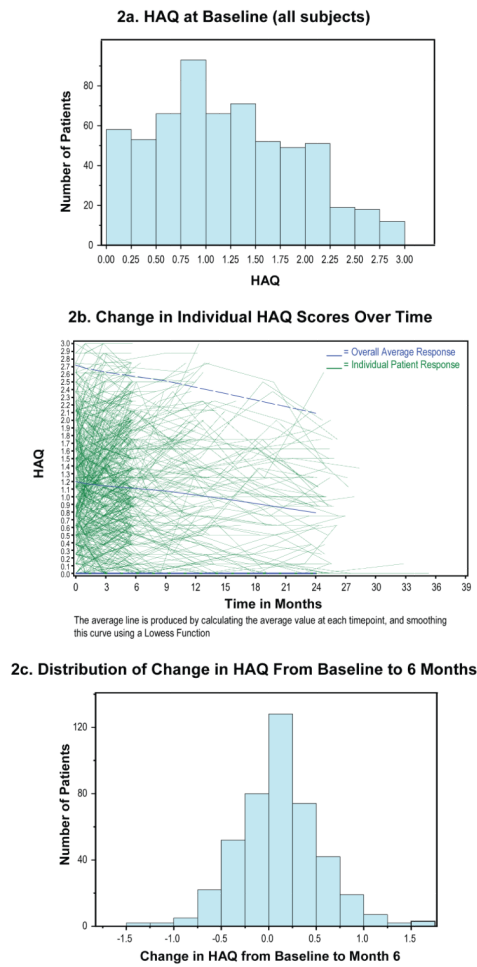**2c. Distribution of Change in HAQ From Baseline to 6 Months**

**Figure 2.**
HAQ Scores at Baseline and Change During Clinical Trials
2a. HAQ at Baseline (all subjects)
2b. Change in Individual HAQ Scores Over Time
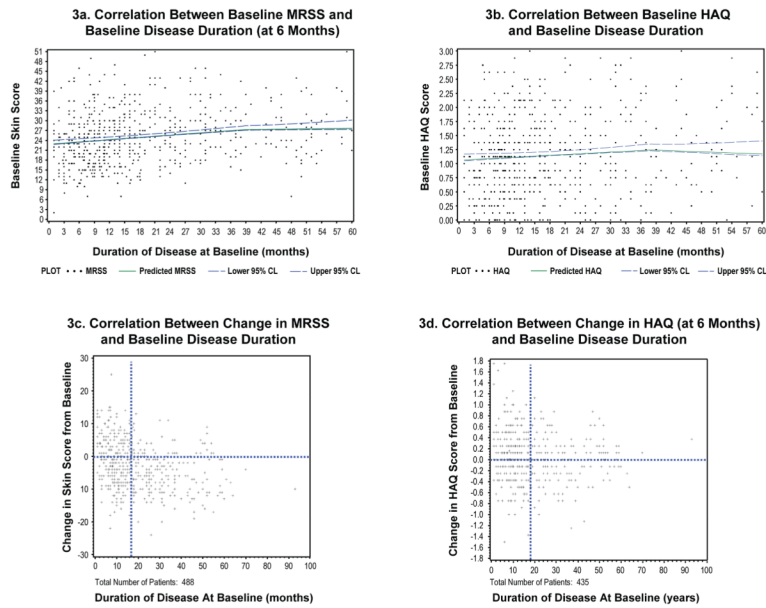2c. Distribution of Change in HAQ From Baseline to 6 Months

**3a. Correlation Between Baseline MRSS and Baseline Disease Duration (at 6 Months)**

**3b. Correlation Between Baseline HAQ and Baseline Disease Duration**

**3c. Correlation Between Change in MRSS and Baseline Disease Duration**

**3d. Correlation Between Change in HAQ (at 6 Months) and Baseline Disease Duration**

**Figure 3.**

Relationships Between MRSS or HAQ and Disease Duration

3a. Correlation Between Baseline MRSS and Baseline Disease Duration (at 6 Months)

3b. Correlation Between Baseline HAQ and Baseline Disease Duration

3c. Correlation Between Change in MRSS and Baseline Disease Duration

3d. Correlation Between Change in HAQ (at 6 Months) and Baseline Disease Duration

**Table 1**

Summary of Clinical Trials Included in Individual Patient Meta-Analysis

| Study Drug, Reference Year | Number of Patients Randomized/ Enrolled | Age (years) Mean ± SD (range) | Number of Females (%) | Disease Duration (months) Mean ± SD (range) n | Baseline MRSS Mean ± SD (range) n | Baseline HAQ Score Mean ± SD (range) n | Baseline FVC Mean ± SD (range) n | Baseline DLCO Mean ± SD (range) n |
|---|---|---|---|---|---|---|---|---|
| Alpha interferon(4) 1999 | 29 | 49.1 ± 10.3 (27–66) | 23 (79%) | 18.6 ± 12.3 (4–47) 29 | 28.3 ± 10.1 (15–51) 29 | 1.3 ± 0.86 (0–3.0) 27 | 87.5 ± 20.9 (42–138) 26 | 71.3 ± 21.9 (35–134) 27 |
| D-Penicillamine(1) 1999 | 134 | 43.7 ± 12.4 (19–74) | 104 (78%) | 9.5 ± 4.1 (2–19) 134 | 21.1 ± 8.0 (7–49) 132 | 1.0 ± 0.67 (0–2.8) 134 | 83.6 ± 16.9 (36–144) 133 | 75.3 ± 18.4 (32–128) 131 |
| Relaxin (Phase II)(3) 2000 | 72 | 44.6 ± 11.7 (15–70) | 61 (85%) | 32.0 ± 20.2 (7–144) 69 | 27.2 ± 7.1 (15–46) 69 | 1.1 ± 0.67) (0–2.9) 69 | 82.9 ± 18.6 (39–129) 68 | 67.6 ± 16.1 (38–100) 68 |
| Methotrexate(5) 2001 | 71 | 47.9 ± 13.7 (18–77) | 59 (83%) | 9.8 ± 10.2 (1–48) 66 | 23.3 ± 9.7 (2–44) 67 | 1.1 ± 0.70 (0–3.0) 66 | 81.8 ± 16.2 (46–115) 33 | 74.1 ± 19.4 (30–118) 67 |
| Minocycline(7) 2004 | 39 | 53.2 ± 13.4 (26–82) | 28 (72%) | 23.3 ± 16.3 (5–61) 38 | 22.8 ± 8.5 (11–48) 39 | 1.1 ± 0.73 (0–2.5) 28 | NA | NA |
| Anti-TGFbeta(6) 2007 | 45 | 49.1 ± 11.3 (24–76) | 35 (78%) | 8.4 ± 5.1 (1–23) 44 | 22.2 ± 5.5 (11–38) 45 | 1.0 ± 0.73 (0–2.6) 45 | 91.2 ± 14.1 (60–125) 45 | 76.4 ± 18.8 (37–117) 45 |
| Relaxin (Phase III)(2) 2008 | 239 | 46.4 ± 10.3 (20–71) | 205 (86%) | 25.6 ± 15.8 (1–93) 231 | 27.8 ± 6.9 (16–51) 232 | 1.2 ± 0.73 (0–2.9) 231 | 84.5 ± 15.6 (42–130) 229 | 60.3 ± 16.1 (24–118) 206 |
| **TOTAL** | **629** | **46.5 ± 11.8 (15–82)** | **515 (82%)** | **19.4 ± 15.9 (1–144) 611** | **25.1 ± 8.2 (2–51) 613** | **1.1 ± 0.71 (0–3.0) 600** | **84.6 ± 16.6 (36–144) 483** | **67.9 ± 18.7 (24–128) 479** |

**Table 2**

Baseline and Change Data for Major Outcomes Measures

| Outcome Variable | Baseline (Mean ± SD) | Mean change in 6 months (Mean ± SD) | Mean change in 12 months (Mean ± SD) |
|---|---|---|---|
| Modified Rodnan Skin Score (total) | 25.1 ± 8.2 N = 613 | −2.9 ± 7.2* N = 492 | −3.4 ± 8.4[†] N = 129 |
| Health Assessment Questionnaire | 1.1 ± 0.71 N = 600 | 0.05 ± 0.45* N = 438 | 0.01 ± 0.51 N = 117 |
| Patient Global Assessment | 48.6 ± 24.2 N = 439 | −1.8 ± 22.7 N = 341 | −3.6 ± 21.8 N = 58 |
| Forced Vital Capacity (% predicted) | 84.6 ± 16.6 N = 483 | −1.7 ± 9.0* N = 379 | 1.3 ± 13.2 N = 87 |
| Diffusion Capacity (% predicted) | 67.9 ± 18.7 N = 479 | −3.4 ± 11.5* N = 259 | −1.6 ± 14.9 N = 115 |
| Oral Aperture | 43.2 ± 9.9 N = 540 | 0.0 ± 6.2 N = 423 | 0.0 ± 6.2 N = 109 |
| Right Hand Extension. | 153.1± 47.5 N = 540 | −1.7 ± 21.3 N = 425 | −3.6 ± 19.3 N = 109 |

SD = standard deviation; N = number of patients with evaluable data for specific outcome.

*
p<.05 vs baseline

[†]
p<.05 vs baseline

**Table 3**

Correlations Between Baseline Outcome Measure Value/Baseline Disease Duration and Change in Same Measure at 6 Months

| Outcome measure | Correlation between baseline outcome measure and change in outcome at 6 months | p-value | Correlation between baseline disease duration and change in outcome at 6 months | p-value |
|---|---|---|---|---|
| MRSS | r = −0.17 | p < 0.001 | r = −0.27 | p < 0.001 |
| HAQ | r = −0.15 | p = 0.002 | r = −0.05 | NS |
| Patient Global Assessment | r = −0.44 | p < 0.001 | r = −0.05 | NS |
| FVC | r = −0.22 | p < 0.001 | r = −0.09 | NS |
| DLCo | r = −0.32 | p < 0.001 | r = −0.02 | NS |
| Oral Aperture | r = −0.28 | p < 0.001 | r = 0.06 | NS |
| Right Hand Extension | r = −0.36 | p < 0.001 | r = 0.12 | p = 0.01 |