Outcome Measurements in Scleroderma: Results from a Delphi Exercise

HASHIM GAZI, JANET E. POPE, PHILIP CLEMENTS, THOMAS A. MEDSGER, RICHARD W. MARTIN, PETER A. MERKEL, BASHAR KAHALEH, FRANK A. WOLLHEIM, MURRAY BARON, MARY ELLEN CSUKA, PAUL EMERY, JILL F. BELCH, SAMINA HAYAT, EDWARD V. LALLY, JOSEPH H. KORN, LÁSZLÓ CZIRJÁK, ARIANE HERRICK, ALEXANDER E. VOSKUYL, PIUS BRUEHLMANN, MURAT INANC, DANIEL E. FURST, CAROL BLACK, MICHAEL H. ELLMAN, LARRY W. MORELAND, NAOMI F. ROTHFIELD, VIVIEN HSU, MAUREEN MAYES, KEVIN M. McKOWN, THOMAS KRIEG, and JAMES R. SEIBOLD

ABSTRACT. ***Objective.*** To obtain a consensus on the minimal clinically relevant treatment effect in various scleroderma disease outcome measures to be used in future clinical trials.

Methods. A Delphi consensus building exercise using a survey was sent out to members of the Scleroderma Clinical Trials Consortium (SCTC). The 65 SCTC members were divided into 2 groups. Group 1 was informed, in a cover letter, of the usual American College of Rheumatology 20% response results in randomized trials using effective biologic treatments for rheumatoid arthritis, while Group 2 was not. The first round of the exercise presented the scleroderma experts with a survey composed of 95 questions/clinical scenarios divided into 8 categories. These included situations where the treatment group improved, or worsened, or where some outcome measures improved, while others worsened. From the responses of this first round, a mean, mode, median, and range of responses for each of the 95 questions was obtained. This information was sent out, in the second round of the Delphi exercise, only to those respondents who answered the first round. The respondent's previous answer and the mean and range from the first round were provided for each question. It gave respondents the option to change any of their initial responses. The median of their responses in the second round was used to calculate the values for the minimal clinically relevant treatment effect.

Results. Thirty-two of the 65 SCTC members returned the first round of the Delphi exercise. Twenty-eight members returned the second round. Intraclass correlation coefficients between responses to round 1 and 2 were calculated for the questions. These varied from 0.99 (excellent agreement) to 0.02 (poor agreement). The p value was under 0.09 for 9 questions and under 0.19 for 20 questions. Standard deviations (SD) were calculated and were found to be lesser for each of the questions in round 2 when compared to the SD in responses from round 1, thus indicating a movement towards a consensus by the second round. An average of 33% of the responses were changed by the respondents in the second round of the Delphi exercise to a value closer to the median/average of the first round's responses. A range in required values for the minimal clinically relevant treatment effect for Modified Rodnan skin score is 3 to 7.5 units, Health Assessment Questionnaire Disability Index (HAQ-DI) 0.2 to 0.25 units, HAQ pain 0.2 to 0.3 units, MD global (100 mm visual analog scale) 8 to 13, patient global assessment 10 to 12, and diffusing capacity (percentage predicted) 9 to 10. The scenarios were especially weighted towards overall disease modification, thus organ-specific measures, such as 6 minute walk time (which has been used in many pulmonary artery hypertension trials), forced vital capacity, and a dyspnea rating (which may be important in scleroderma lung trials), were not included in the survey.

Conclusion. Our study begins to address the current deficiency in our knowledge of appropriate values for the minimal clinically relevant treatment effect in various scleroderma disease outcome measures. A consensus could be achieved, or at least a range of minimal clinically relevant treatment effect values could be found for several outcome measurements. Of course, this consensus statement will be modified by evidence as it accrues in each consensus area. (First Release Feb 1 2007; J Rheumatol 2007;34:501–9)

Key Indexing Terms:
SYSTEMIC SCLEROSIS TREATMENT EFFECT DELPHI EXERCISE

From the Division of Rheumatology, Department of Medicine, The University of Western Ontario, London, Ontario, Canada.

Supported by the Summer Research Training Program of the Schulich School of Medicine at the University of Western Ontario, London, Ontario, Canada.

H. Gazi, BA, Medical Student, Schulich School of Medicine; J.E. Pope, MD, MPH, FRCPC, Professor of Medicine and Epidemiology and Biostatistics, Division of Rheumatology, Department of Medicine, The University of Western Ontario; P. Clements, MD, Professor of Medicine, University of California, Los Angeles, Los Angeles, California; T.A. Medsger, MD, Department of Medicine, Division of Rheumatology, University of Pittsburgh, Pittsburgh, Pennsylvania; R.W. Martin, MD, College of Human Medicine, Michigan State University, Grand Rapids, Michigan; P.A. Merkel, MD, MPH, Boston University School of Medicine, Boston, Massachusetts; B. Kahaleh, MD, Professor, Chief of

Rheumatology, Medical College of Ohio, Toledo, Ohio, USA; F.A. Wollheim, MD, PhD, Department of Rheumatology, Lund University Hospital, Lund, Sweden; M. Baron, MD, Jewish General Hospital, McGill University, Montreal, Quebec, Canada; M.E. Csuka, MD, Professor of Medicine, Department of Rheumatology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; P. Emery, BA, MA, MBChir, MRCP, MD, FRCP, Professor of Medicine, Leeds University, Academic Unit of Musculoskeletal Disease, Leeds; J.F. Belch, MD, Reader and Consultant Physician, Ninewells Hospital and Medical School, Dundee, United Kingdom; S. Hayat, MD, Department of Medicine — Rheumatology, Louisiana State University, Shreveport, Louisiana; E.V. Lally, MD, Chief, Division of Rheumatology, Brown University School of Medicine, Providence, Rhode Island; J.H. Korn, MD, Boston University School of Medicine, Boston, Massachusetts, USA; L. Czirják, MD, DSc, Department of Immunology and Rheumatology, University of Pécs, Pécs, Hungary; A. Herrick, MBChB, MD, FRCP, University of Manchester, Manchester, UK; A.E. Voskuyl, MD, PhD, VU University Medical Center, Amsterdam, The Netherlands; P. Bruehlmann, MD, Department of Rheumatology and Physical Medicine, University Hospital, Zurich, Switzerland; M. Inanc, MD, Department of Internal Medicine, Division of Rheumatology, University of Istanbul, Istanbul, Turkey; D.E. Furst, MD, University of California, Los Angeles, Los Angeles, California, USA; C. Black, MD, Centre for Rheumatology, Royal Free Hospital, London, UK; M.H. Ellman, MD, The University of Chicago, Chicago, Illinois; L.W. Moreland, MD, Division of Clinical Immunology and Rheumatology, The University of Alabama at Birmingham, Birmingham, Alabama; N.F. Rothfield, MD, Professor of Medicine, Chief, Division of Rheumatic Diseases, University of Connecticut Health Center, Farmington, Connecticut; V. Hsu, MD, Director, Scleroderma Program, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, Piscataway, New Jersey; M. Mayes, MD, MPH, Division of Rheumatology and Clinical Immunogenetics, University of Texas–Houston Medical School, Houston, Texas; K.M. McKown, MD, University of Wisconsin–Madison, Madison, Wisconsin, USA; T. Krieg, MD, Professor, Department of Dermatology, University of Cologne, Cologne, Germany; J.R. Seibold, MD, Department of Internal Medicine/Rheumatology, University of Michigan, Ann Arbor, Michigan, USA.

Address reprint requests to Prof. J.E. Pope, Rheumatology Center, St. Joseph's Health Care London, 268 Grosvenor Street, Box 5777, London, Ontario N6A 4V2, Canada. E-mail: janet.pope@sjhc.london.on.ca

Accepted for publication October 13, 2006.

Systemic sclerosis (SSc) or scleroderma is a connective tissue disease characterized by vascular abnormalities, presence of autoantibodies, and fibrosis of skin and various viscera[1,2]. Limited scleroderma involves skin distal to the elbows and/or knees, and may involve the face and neck to clavicles, while diffuse scleroderma affects skin both distal and proximal to the elbows and/or knees, and/or involvement of the trunk[3].

Patients with diffuse scleroderma may have progressive disease and develop renal crisis, cardiomyopathy, and significant pulmonary fibrosis. Both subtypes may develop pulmonary artery hypertension (PAH). Various outcome measures have been developed to monitor the progression of disease in patients with scleroderma, and the validity of many of these has been assessed[4]. These include the modified Rodnan skin score (MRSS)[5], the diffusing capacity for carbon monoxide (DLCO) and forced vital capacity for lung involvement, New York Heart Association Functional Class and 6 minute walk time for PAH, the Health Assessment Questionnaire (HAQ)[6], patient global assessment, and physician global assessment.

Clinical trials in scleroderma have been completed but most drugs have not demonstrated clinically significant benefits. This may be because these trials required too large an improvement in their outcome measures (i.e., underpowered), the outcome measures were not sensitive to change, the wrong scleroderma population was studied, or the treatment was not effective at the dose administered over the time of the trial. Our purpose was to define an acceptable value for minimal clinically relevant treatment effect in potential outcome measures. This knowledge could then be used to guide future clinical trials, especially to avoid Type I errors in scleroderma trials.

Estimating values for the minimal clinically relevant treatment effect may be influenced by many factors. These include the disease severity and potential response of the patient, whether the comparison is between 2 active drugs or in the setting of a placebo-controlled trial, the scaling properties of the outcome measures, or the personal preference of the investigator concerned[7]. The Delphi exercise is used in the development of consensus among experts in a field of study[8]. It has been used previously in setting the minimal clinically relevant treatment effect for clinical trials of antirheumatic drugs[7]. We employed a Delphi exercise in obtaining minimal clinically relevant treatment effect values for various scleroderma outcome measures.

MATERIALS AND METHODS

A survey designed to estimate values for the minimal clinically relevant treatment effect in various scleroderma clinical trial scenarios was created. This survey was sent to the 65 members of the Scleroderma Clinical Trial Consortium (SCTC) whose E-mail addresses were listed in the SCTC institution roster[9]. The SCTC is a not-for-profit organization with international membership composed of scleroderma researchers who are interested in clinical design of trials. Membership is via an institution with a dedicated scleroderma clinic and publications in scleroderma. The first 32 members listed as institutional contacts were placed in Group 1, and the remaining 33 members were placed in Group 2. The survey sent to Group 1 and Group 2 members was identical, except in a cover letter, Group 1 participants were informed of the usual American College of Rheumatology (ACR) results in rheumatoid arthritis (RA) trials associated with effective anti-tumor necrosis factor treatment[10]. We stated that excellent biologic drugs yield at most 70% of RA subjects with an ACR 20% response. The groups were separated to determine if exemplifying a comparison disease would change minimal clinically relevant treatment effect estimates.

The questionnaire asked the participants to consider a hypothetical set of patients, often with early diffuse scleroderma, who are placed in a randomized controlled trial where one group is given an active drug "X" and the other a placebo. The respondents were asked to compare these groups for clinical improvements in the following outcome measures, corresponding to the 8 sections of the survey: MRSS, HAQ Disability Index and pain score (HAQ-DI and HAQ pain), patient global assessment, physician global assessment, percentage predicted DLCO, mixed outcome scenarios (such as 2 outcomes improving and one worsening), disease prevention, and a section directly asking for the minimal clinically relevant treatment effect. The scenarios posed various combinations of either the treatment group or the placebo group improving, worsening, or not changing. In each case, the respondent was asked to estimate a value for the minimal clinically relevant treatment effect. Although respondents were asked about both improvement and worsening, there were more questions that inquired into the minimum treatment effect to define improvement, as this value would be more relevant for use in the setting of a clinical trial. These values were requested in different forms; some

questions presented the respondents with the final score in the placebo group and asked them what a minimally clinically important improvement in the outcome measure with active treatment would be. Other questions presented the results obtained in the placebo group and asked respondents to estimate values for the minimal clinically relevant treatment effect in terms of a percentage change. The section dealing with mixed outcome scenarios dealt with situations where some measures improved while others worsened, and asked respondents whether they believed that the patient had improved or worsened overall. The survey contained a total of 95 questions. Each respondent estimated the minimal clinically relevant treatment effect and answered the questions on an individual basis.

The survey was distributed as a Word-format document via E-mail and participants were given 6 weeks to respond using E-mail or facsimile transmission. From the responses to this first round of the Delphi exercise, a mean, mode, median, and range of responses for each of the 95 questions was obtained. This information was tabulated and sent, in the second round of the exercise, only to those respondents who answered the first round. The table included a personalized column showing the individual respondent's previous answers, and the mean, mode, and range of responses obtained for each question from all respondents combined. A blank column was included for the respondent to indicate his/her new response to each of the 95 questions. Thus, the respondents were given the option of changing any of their initial responses based on the answers of the other respondents in the first round. The respondents were given 8 weeks to return the second questionnaire and were sent periodic reminders. Only those respondents who returned the second round of the Delphi exercise were considered authors of the study.

The median of the responses from the second round was used to calculate the final values for the minimal clinically relevant treatment effect (or acceptable final scores). The effect of the Delphi exercise on convergence of opinion was evaluated using JMP software and descriptive statistics like median and range[11]. The JMP software was used in the calculation of p values, SD, and intraclass correlation coefficients (ICC) for comparing the answers to questions between the 2 rounds. The formula used was ICC = (MSTr – MSE)/[MSTr + (n – 1)MSE], where MSE is the error mean square, MSTr is the treatment mean square, and n is the number of measurements per entity. These values were obtained from the analysis of variance table[12]. An ICC > 0.8 was used to indicate excellent agreement and an ICC between 0.6 and 0.8 to indicate good agreement. Negative ICC were documented as 0[1]. Tests were done to determine whether the variability of results of Round 1 was statistically different from Round 2. A p value of under 0.05 implied very different results but was expected to be insensitive (i.e., only large differences in SD would be statistically significant).

RESULTS

Thirty-two of the 65 SCTC members returned the first round of the Delphi exercise (14 from Group 1 and 18 from Group 2). Twenty-eight members returned the second round (13 from Group 1 and 15 from Group 2). The median values for the minimal clinically relevant treatment effect estimates varied slightly between the 2 groups. However, these estimates for some of the questions were lower in the responses from Group 1, while for other questions they were lower in the responses obtained from Group 2. They were not consistently high or low in either group. Therefore the results of the second round of the Delphi exercise were not analyzed separately for each of the groups, but were pooled together for statistical analysis of the final responses. Thus, providing the respondents with an example of the results achievable in RA with very effective treatment did not seem to alter the results obtained from Group 2.

The responses obtained from the 28 respondents who answered both rounds were compared between the first and second rounds, and p values were obtained for all survey questions. The p value could indicate that the estimated ICC is statistically different than 0, and as noted earlier, a p value < 0.05 would indicate a statistically significant difference in responses between rounds 1 and 2. Only one of the 95 questions had a p value < 0.05.

SD were calculated and found to be smaller in round 2 than the SD in round 1 for responses to 78 of the 82 questions, implying higher consensus in the second round for the majority of questions. It remained unchanged in 2 questions and was higher in another 2 questions. Table 1 shows the SD, p values, and minimal clinically relevant treatment effect estimates obtained from responses to section 8 of the survey, which asked directly for these estimates given different baseline scores. The SD are lower in Round 2 for each question. More importantly, the median values for the minimal clinically relevant treatment effect estimate are not static for each outcome measure. Rather, as can be seen from these results, participant responses changed depending on the presenting characteristics of the patient (i.e., the baseline score). In general, the minimal clinically relevant treatment effect requirements were higher for patients presenting with more severe disease than those with milder disease.

ICC were also used to study the effect of the Delphi exercise on convergence of opinion between responses in Rounds 1 and 2. A description of selected questions asking for the minimal clinically relevant treatment effect in the first 5 sections of the survey, their responses over the 2 rounds, and a statistical analysis of these responses in terms of medians, p values, SD and ICC is included in Table 2. The ICC and p values varied from question to question. Nine questions in the table demonstrated good agreement (ICC between 0.6 to 0.8). Of the questions not included in the table, the highest ICC was 0.99 (excellent agreement). This table also reports the difference in responses, when respondents were asked for estimates of the minimum clinically significant treatment effect, between scenarios of improvement and worsening. With respect to the MRSS, the median response to one of the questions gave a minimal clinically relevant treatment effect value of at least 35% improvement from baseline, in order to consider the drug to be effective. Another pair of questions presented a patient with a baseline skin score of 22, and asked how much of a change would be required to define improvement and how much of a change would signify deterioration; the median response to both questions was 5 units. To define improvement using the HAQ-DI, the respondents required a reduction in score by at least 0.21 units, and conversely, they required an increase by at least 0.21 units in the HAQ-DI to signify worsening of the patient's condition. For both patient and physician global assessment, if 20% of patients taking placebo reported improvement, at least 37% of the patients taking active drug would need to improve for a minimal clinically relevant difference. The survey yielded a minimal clin-

Table 1. Median responses (SD) to questions asking for the minimal clinically relevant treatment effect for various scleroderma outcome measures.

Outcome Measure	Baseline Score	p Value (to measure change of opinion between the 2 rounds)	Minimal Treatment Effect Responses (n = 28) Round 1 Median Response (SD)	Round 2 Median Response (SD)
Modified Rodnan skin	26 (severe)	0.3	6 (1.9)	7.5 (1.5)
score (range 0–51)	17 (mild)	0.8	5 (1.5)	5 (1.1)
	6 (limited)	0.4	2 (1.6)	3 (1.2)
MD global assessment	32	0.8	8 (4.3)	8 (1.8)
(range 0–100)	44*	0.5	11 (4.0)	10 (2.3)
	50	0.7	13 (4.5)	13 (3.1)
Patient global assessment	30	0.7	8 (3.1)	10 (2.2)
(range 0–100)	40*	0.2	10 (4.4)	10 (3.1)
	50	0.15	13 (5.5)	12 (3.3)
ESR (0–20 normal)	26	0.5	6 (5.1)	7 (2.3)
	40*	0.7	10 (7.0)	10 (3.9)
DLCO % predicted	80	0.3	8 (6.9)	9 (5.5)
(N 80–120)	70*	0.2	10 (4.4)	10 (3.8)
	55	0.1	10 (3.6)	10 (3.2)
FVC (% predicted)	64	0.6	10 (3.5)	10 (3.2)
(N 80–120)	80*	0.3	10 (6.2)	10 (3.7)
Profibrotic cytokine	2 times normal	0.3	0.75 (0.6)	0.75 (0.47)
HAQ Disability Index	0.6	0.5	0.2 (0.08)	0.2 (0.04)
(0–3)	0.9	0.3	0.22 (0.1)	0.21 (0.04)
	1.3	0.2	0.25 (0.19)	0.25 (0.11)
HAQ pain (0–3)	0.5	0.6	0.2 (0.1)	0.2 (0.06)
	1.0	0.4	0.3 (0.17)	0.3 (0.12)
	1.125	0.3	0.3 (0.18)	0.3 (0.12)

DLCO: diffusing capacity for carbon monoxide; FVC: forced vital capacity; HAQ: Health Assessment Questionnaire. * Average presenting score.

ically relevant treatment effect value of 10% predicted for defining improvement in DLCO.

A brief summary of the range of minimal clinically relevant treatment effect estimates obtained from Round 2 for the most relevant outcome measures is presented in Table 3. There is a range of values for the minimal clinically relevant treatment effect derived from scenarios, which may suggest in some cases that a percentage change or percentage difference may be more precise. The minimal clinically relevant treatment effect value for MRSS is 3 to 7.5 units, for HAQ-DI 0.2 to 0.25, for HAQ pain 0.2 to 0.3 units, for global assessments about 10 mm on a 100 mm VAS, and for percentage predicted DLCO it is 10% change.

The change in the median value for the minimal clinically relevant treatment effect was minimal between the 2 rounds of the Delphi exercise, and for many of the questions, it did not change at all. An average of 33% of the responses were changed by the respondents in the second round of the Delphi exercise to a value closer to the median/average of the first round's responses. Individually, of the 95 questions asked, one respondent changed 60% (maximum change) of his/her previous responses, while another changed 7.5% (minimum change) of his/her answers. The median value for the minimal clinically relevant treatment effect was increased in 35%, decreased in 26%, and unchanged in 39% of the questions between Rounds 1 and 2, indicating that there are changes in both directions using the consensus technique.

Table 4 shows the responses to the section of the survey dealing with mixed outcome scenarios. As can be seen, the respondents were unsure of the definition of improvement in some of these situations, while in others there was higher consensus. From the results of responses to question asked earlier, we found the minimal clinically relevant treatment effect requirement, in similar situations, for the percentage predicted DLCO is 10% predicted, for the HAQ 0.3, and for the MRSS 5 units. In scenario 1, where the percentage predicted DLCO worsened, but the MRSS and HAQ improved (all by a measure relatively close in value to the minimal clinically relevant treatment effect), there were conflicting opinions regarding the progress of the patient. However, in scenario 2, where none of the outcome measures changed by an amount large enough to satisfy the minimal clinically relevant treatment effect requirements, 94% of respondents believed that the patient had not improved.

In order to determine if the nonrespondents differed from the respondents, all SCTC members were asked for their age

Table 2. Minimal clinically relevant treatment effect estimates and the effect of the Delphi exercise on consensus. The standard deviation decreased in round 2 in all but 4 questions, indicating an increase in consensus by the second round of the Delphi exercise.

Information Presented and Questions Asked	Median Response Round 1 (N = 32)	Median Response Round 2 (N = 28)	ICC	p	SD ROUND 1	SD ROUND 2
MRSS Initial Score of placebo and active drug group is 20. Final score in placebo: 13. Improvement minimal clinically relevant treatment effect required in active drug group:	7 units 32% from baseline 37% of patients need to achieve similar results	6 units 35% from baseline 25% of patients need to achieve similar results	0.14 0.61 0.53	0.6 0.5 0.2	5 25 18	5 21 7
MRSS Individual with baseline skin score of 22. Improvement is defined by a decrease of at least: Worsening is defined by an increase of at least:	 6 units 6 units	 5 units 5 units	 0.31 0.09	 0.3 0.3	 5 5	 1 1
HAQ Disability Index Initial Score of placebo and active drug group is 1.1. Final score in placebo: 1.1. Minimal clinically relevant treatment effect required in active drug group:	 0.21 units	 0.21 units	 0.56	 0.2	 0.01	 0.01
HAQ Disability Index Individual with high baseline at 1.875. Improvement is defined by a reduction of at least: Worsening is defined by an increase of at least:	 0.25 units 0.25 units	 0.28 units 0.28 units	 0 0	 0.09 0.09	 0.07 0.07	 0.12 0.12
HAQ Disability Index Individual with baseline HAQ-DI of 1.125. Improvement is define by a change of at least: Worsening is defined by a change of at least:	 0.25 units 0.25 units	 0.21 units 0.21 units	 0.03 0	 0.16 0.09	 0.07 0.07	 0.02 0.02
Patient Global Assessment 20% of responders on placebo reported improvement. A similar response from at least what percentage of patients would be clinically significant:	 27% of patients	 37% of patients	 0.49	 0.9	 10	 3.5
Patient Global Assessment (0-100mm): **Scenario** / **Group** / **Start** / **End** Case 1 / Placebo / 42 / 40 Active Drug / 42 / **Answer 1:** Case 2 / Placebo / 42 / 35 Active Drug / 43 / **Answer 2:** Case 3 / Placebo / 42 / 48 Active Drug / 43 / **Answer 3:**	Answer 1: 30 Answer 2: 28 Answer 3: 33	Answer 1: 33 Answer 2: 30 Answer 3: 38	0.13 0.34 0.07	0.1 0.4 0.1	7 4 12	2 0 5
Physician Global Assessment 20% of responders on placebo reported improvement. A similar response from at least what percent of patients is clinically significant:	40% of patients	37% of patients	0.63	0.2	10	3.5
Physician Global Assessment (0-100mm): **Scenario** / **Group** / **Start** / **End** Case 1 / Placebo / 49 / 50 Active Drug / 49 / **Answer 1:** Case 2 / Placebo / 50 / 56 Active Drug / 51 / **Answer 2:** Case 3 / Placebo / 50 / 45 Active Drug / 51 / **Answer 3:**	Answer 1: 40 Answer 2: 42 Answer 3: 38	Answer 1: 41 Answer 2: 46 Answer 3: 40	0.73 0.71 0.28	0.4 0.4 1	7 9 3	1 2 0
Diffusion lung capacity for carbon monoxide (DLCO) Minimal treatment increase in DLCO % predicted for active drug	12.5 % predicted	10% predicted	0.63	0.03	3.5	0
DLCO (%) Predicted: **Scenario** / **Group** / **Start** / **End** Case 1 / Placebo / 49 / 50 Active Drug / 49 / **Answer 1:** Case 2 / Placebo / 50 / 56 Active Drug / 51 / **Answer 2:** Case 3 / Placebo / 50 / 45 Active Drug / 51 / **Answer 3:**	Answer 1: 76 Answer 2: 81 Answer 3: 70	Answer 1: 77 Answer 2: 75 Answer 3: 68	0.66 0.62 0.63	0.4 0.8 0.1	1.4 1.4 0.7	0 .7 4

and number of years in practice as a rheumatologist. The mean age of the respondents was 55.5 years and the number of years in practice was 25.5, and in the nonrespondents the mean age was 56 years with 22.5 years of practice. Thus at least on simple characteristics, the respondents did not appear different from those who did not complete the survey.

Table 3. Summary of minimal clinically relevant treatment effect estimates.

Outcome Measure	Range*
MRSS (0–51)	3–7.5
HAQ Disability Index (0–3)	0.2–0.25
HAQ pain (0–3)	0.2–0.3
MD global assessment (0–100)	8–13
Patient global assessment (0–100)	10–12
DLCO % predicted (0–100%)	9–10

* Range is given as some scenarios varied with either worsening or improvement and at various levels of disease state.

DISCUSSION

The Delphi exercise is useful in the development of consensus by a group of experienced individuals who can independently express their opinions without fear of embarrassment, giving each the opportunity to change responses based on a statistical description of the responses of the rest of the group rather than individual responses of any of its members. It seemed to be effective in approaching values for the minimal clinically relevant treatment effect in outcome measurements used in scleroderma clinical trials. We have documented a first step towards actual approximation of the minimal clinically relevant treatment effect in scleroderma.

We had divided the respondents into 2 groups, as we wanted to see whether the group being more informed about RA clinical trials with effective agents would tend to have different estimates. One reason why our results did not show an appreciable difference in the minimal clinically relevant treatment effect estimated between the 2 groups may be because of our small sample size (14 in Group 1 and 18 in Group 2). Another explanation may be that we only informed the Group 1 respondents of the current minimum clinically important difference in RA once, on the first page of the survey, during the introduction; they may not have considered this information while answering the 95 questions contained in the survey. However, it may be that how a disease modifying antirheumatic drug performs in RA trials is irrelevant to expectations of drugs in scleroderma trials. The p value was chosen, along with the ICC, as a statistical descriptor for comparing the 2 rounds of the exercise as it incorporated both the means and standard deviations in comparing the 2 groups. Overall, most of the p values obtained were not statistically significant. Based also on the ICC values obtained, the implication is that there was enough of a consensus between the 2 rounds that there was not a significant difference in the way the questions were answered in the second round, but with small numbers, it was not a robust test. However, as medians did not vary much between the 2 rounds, 2 rounds were enough to reach a relatively strong consensus in opinion. This observation is also confirmed by the fact that the SD for all but 4 of the 82 questions in the survey was lower in the responses to the second round compared to the first. The strength of the convergence of opinion between the 2 rounds is further supported by our results, as almost all the questions changed in the second round were changed to a value closer to the median obtained initially.

Table 3 briefly summarizes the results that our study was aimed at obtaining. From our results, it became apparent that the minimal clinically relevant treatment effect could not be effectively expressed as a single value, but would be better expressed as a range depending on various factors related to patient presentation. These values represent the range of values for the minimal clinically relevant treatment effect that most experts would agree upon. This range is primarily the range of the minimal clinically relevant treatment effect to define improvement, based on varying degrees of severity of disease at presentation. Future trials may seek to define the range for the minimum clinically relevant treatment effect to define worsening. In the questions where we asked respondents to give an estimate for a value that would signify improvement and then an estimate for a value to signify worsening (Table 2), our results showed most respondents chose the same value, in units, to represent the minimum change, be it for improvement or for worsening.

It may be that percentage change is a more effective means of expressing the minimal clinically relevant treatment effect for variables. For instance, the minimal clinically relevant treatment effect requirement was thought to be larger for high baseline MRSS versus lower baseline skin scores. This would represent an appreciation of the variability of outcome measurements, which may be related to a percentage change as opposed to absolute units. The ICC of rater variability in skin scores has been well documented[5,13-15]. A clinically relevant change should exceed measurement error. Minimal clinically relevant treatment effect has often been studied post hoc in RA clinical trials where there is active disease at baseline with various inclusion criteria to enter a trial. Thus floor or ceiling effects and absolute change compared to percentage change have not been large issues. In scleroderma trials, characteristics are different if comparing patients with diffuse early disease to those with later disease with lower skin scores or limited scleroderma skin scores. Thus an absolute minimal clinically relevant treatment effect requirement value for all scenarios is not likely to be achieved.

While responses to most of the questions in the survey were numeric and appropriately reported as a median, the questions dealing with the mixed outcome scenarios required the respondent to answer with a yes/no/uncertain response. As can be seen in Table 4, we chose to report these responses as the percentage of respondents who answered yes/no/uncertain. A receiver-operating characteristic curve analysis may also have been used to report these responses; however, we felt that this was not necessary for the purpose of our study.

In a recent report dealing with minimally important difference (MID) in diffuse SSc for MRSS and HAQ-DI, investigators were asked to rate the change in the patient's health since

Table 4. Mixed outcome scenarios. Respondents were presented with situations where some outcome measures improved while others worsened and were then asked whether they believed the patient had improved overall.

Scenario 1

Outcome Measure	Baseline Score	Score at End of Treatment	
DLCO % predicted	68%	60%	
MRSS	18	14	
HAQ	1.2	0.9	
Did the Patient Improve?	N = 28 Percentage of "Yes" responses	N = 28 Percentage of "No" responses	N = 28 Percentage of "Don't Know" responses
Round 1	27%	43%	30%
Round 2	17%	63%	20%

Scenario 2

Outcome Measure	Baseline Score	Score at End of Treatment	
DLCO % predicted	68%	72%	
MRSS	18	16	
HAQ	1.2	1.1	
Did the Patient Improve?	N = 28 Percentage of "Yes" responses	N = 28 Percentage of "No" responses	N = 28 Percentage of "Don't Know" responses
Round 1	6%	84%	10%
Round 2	3%	94%	3%

Scenario 3

Outcome Measure	Baseline Score	Score at End of Treatment	
DLCO % predicted	65%	55%	
MRSS	23	18	
MD Global assessment	35	20	
Did the Patient Improve?	N = 28 Percentage of "Yes" responses	N = 28 Percentage of "No" responses	N = 28 Percentage of "Don't Know" responses
Round 1	23%	57%	20%
Round 2	23%	66%	11%

entering a D-penicillamine study. Patients who were rated as slightly improved were defined as minimally changed. MID estimates for the MRSS improvement ranged from 3.2 to 5.3 (0.40–0.66 effect size) and for the HAQ-DI from 0.10 to 0.14 (0.15–0.21 effect size). While the skin score results correspond with the consensus results in a general way, the HAQ-

DI results indicate that MID of the HAQ-DI in scleroderma is less than in RA, different from the consensus results[16].

Results from our survey indicated that the minimal clinically relevant treatment effect for skin score would be a 32%–35% improvement over baseline. This result is validated by previous expert opinion and a data-driven approach, which both indicate a change of 30% from baseline is clinically relevant[17,18].

One major limitation of our study was the small number of participants. The Delphi exercise requires opinions of experts within the area in which it is being employed, and there were thus only 65 individuals we could reasonably include in the survey. The large range of ICC, p values, and SD, and the 2 instances of higher SD in Round 2 may be attributed to the small sample size. Having said this, this study did not require that every member of the SCTC respond to the survey, as given the nature of a Delphi exercise, this sample size and response rate were quite appropriate. Our number of respondents was in the range of most Delphi exercises, and analysis of characteristics of respondents versus nonrespondents did not show a significant difference in either their age or the number of years they had been in practice.

Further, some of the questions in the survey were challenging to answer. The section of the survey dealing with disease prevention was especially difficult, as more than half the respondents did not answer many of the questions. This is a different clinical trial paradigm, where stabilization may be expected as benefit versus worsening.

We did not ask about some outcome measurements that other experts (such as pulmonologists) may have already deemed to be clinically relevant, such as walk time, functional class, etc. This is a potential problem, as it is not always clear that results in chronic obstructive pulmonary disease or PAH apply to all patients with SSc, as shown in the recent trial of bosentan for SSc interstitial lung disease (ILD), where the 6-minute walk distance did not appear to change in SSc ILD.

There are also intrinsic limitations to the Delphi technique, as the minimal clinically relevant treatment effect values obtained are simply a consensus of opinions of scleroderma experts. They are not based on achievable results within a randomized controlled trial. The majority opinion is not necessarily correct. Indeed, the Delphi technique is very useful as an initial step when good data are not yet available, but the results of this exercise need to be tested in actual trials and will undoubtedly change as data are accrued. Also, these estimates may not be attainable within emerging scleroderma trials unless the number of patients included in those trials is increased. For some outcome measures, it would be pertinent to pose the questions to patients rather than physicians. For example, the Medical Outcomes Study Short Form-36 Health Survey and HAQ are patient-reported outcomes. How improvement is "framed" may change the minimal clinically relevant treatment effect estimate; chemotherapy in lung cancer may double median survival, but only prolong life by 3 months. Doubling survival sounds like effective therapy, but changing median survival by 3 months does not. Clearly, future trials should try to derive the minimum clinically relevant treatment effect through prospective study and should result in data-driven, rather than consensus-derived, results. Once corroborated or changed by experimental studies, sample size calculations can take into consideration achieving at least a minimal clinically relevant treatment effect as a between-groups difference (delta) in scleroderma trials. Using a minimal clinically relevant treatment effect to assess improvement/worsening results is an intuitively attractive and clinically meaningful approach, although dichotomizing data can decrease statistical power.

The Delphi exercise does not change a case definition, so items not asked or redefined cannot be listed as part of possible future outcome measurements in scleroderma clinical trials. In addition, we do not know if the minimally important change in diffusing capacity is smaller than the within-patient test variability (i.e., the minimal between-groups difference could be less than the within-patient variability, or clinically relevant change, of 10%).

Our study has yielded estimates for the required values for the minimal clinically relevant treatment effect for various scleroderma outcome measures, keeping in perspective a variety of baseline characteristics. Using the minimal clinically relevant treatment effect estimates from the results of our study may be a more reliable option, as the data from our responses showed that unfettered estimates of individual experts may be excessively low or high. Before our study, there were no published estimates for a target minimal clinically relevant treatment effect in the various outcome measures used in scleroderma. We have addressed some of the deficiencies in this knowledge. Our results can be used as an initial guide for future clinical trials of drugs for the treatment of SSc, although they represent only the first step in developing data-derived estimates of minimally important differences.

REFERENCES

1. Pope JE. Variability of skin scores and clinical measurements in scleroderma. J Rheumatol 1995;22:1271-6.
2. Seibold J. Connective tissue diseases characterized by fibrosis: scleroderma. In: Ruddy S, Harris ED Jr, Sledge CB, editors. Textbook of rheumatology. 6th ed. Philadelphia: W.B Saunders; 2001:1133-59.
3. LeRoy EC, Black C, Fleischmajer R, et al. Scleroderma (systemic sclerosis): classification, subsets and pathogenesis. J Rheumatol 1988;15:202-4.
4. Merkel PA, Clements PJ, Reveille JD. Current status of outcome measurement development for clinical trials in systemic sclerosis. Report from OMERACT 6. J Rheumatol 2003;30:1630-47.
5. Clements P, Lachenbruch P, Seibold J, et al. Skin thickness score in systemic sclerosis: an assessment of interobserver variability in 3 independent studies. J Rheumatol 1993;20:1892-6.
6. Fries JF, Sitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
7. Bellamy N, Carette S, Ford PM, et al. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials — results of a

consensus development (Delphi) exercise. J Rheumatol 1992;19:451-7.
8. Linstone HA, Turoff M, editors. The Delphi method — techniques and applications. Reading, PA: Addison-Wesley; 1975.
9. Pope J, Ouimet JM, Krizova A. Scleroderma treatment differs between experts and general rheumatologists. Arthritis Rheum 2006;55:138-45.
10. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology preliminary definition of improvement in RA. Arthritis Rheum 1995;38:727-35.
11. SAS Institute Inc. JMP statistical software–version 4. Cary, NC: SAS Institute Inc.; 2000.
12. Bland JM, Altman DG. Statistics notes: measurement error and correlation coefficients. BMJ 1996;313:41-2.
13. Pope J, Baron M, Bellamy N, et al. The variability of skin scores and clinical measurements in scleroderma. J Rheumatol 1995;22:1271-6.
14. Clements P, Lachenbruch P, Seibold J, et al. Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. J Rheumatol 1995;22:1281-5.
15. Brennan P, Silman A, Black C, et al. Reliability of skin involvement measures in scleroderma. The UK Scleroderma Study Group. Br J Rheumatol 1992;31:457-60.
16. Khanna D, Furst DE, Hays RD, et al. Minimally important difference in diffuse systemic sclerosis — results from the D-penicillamine study. Ann Rheum Dis 2006;65:1325-9. Epub 2006 Mar 15.
17. Seibold JR, McCloskey DA. Skin involvement as a relevant outcome measure in clinical trials of systemic sclerosis. Curr Opin Rheumatol 1997;9:571-5.
18. Khanna D, Furst DE, Clements PJ, et al. Responsiveness of the SF-36 and the Health Assessment Questionnaire Disability Index in a systemic sclerosis clinical trial. J Rheumatol 2005;32:832-40.