# Disability in Systemic Sclerosis — A Longitudinal Observational Study

MIREILLE SCHNITZER, MARIE HUDSON, MURRAY BARON, Canadian Scleroderma Research Group, and RUSSELL STEELE

***ABSTRACT. Objective.*** To assess disability in systemic sclerosis (SSc) longitudinally and to identify disease-specific determinants, after accounting for informative patient dropout.

***Methods.*** We performed a multicenter, longitudinal study of 745 patients with SSc followed in the Canadian Scleroderma Research Group registry. Disability was assessed using the Health Assessment Questionnaire (HAQ). Longitudinal changes in disability were modeled using statistical approaches accounting for various levels of patient dropout.

***Results.*** In all the models, disability in SSc worsened over time. The magnitude of the worsening was small when patient dropout was assumed to be completely at random (increase in the HAQ of 0.022, 95% CI 0.002–0.042, per year). After accounting for different levels of informative patient dropout, the increase in the HAQ ranged from 0.039 (95% CI 0.018–0.061) per year to 0.071 (95% CI 0.048–0.094) per year. Thus, using the most conservative of these estimates, this was equivalent to an increase in the HAQ of 0.12 over 3 years. The disease correlates found to be most closely associated with disability were diffuse disease and breathing problems.

***Conclusion.*** Our study provides strong evidence that SSc causes increased disability over time, with breathing problems and disease type being the strongest predictors of disability. Statistical modeling accounting for informative patient dropout is necessary to properly assess the outcomes of patients followed longitudinally. (First Release Dec 15 2010; J Rheumatol 2011;38:685–92; doi:10.3899/jrheum.100635)

*Key Indexing Terms:*
SYSTEMIC SCLEROSIS DISABILITY HEALTH ASSESSMENT QUESTIONNAIRE

Systemic sclerosis (SSc) is a connective tissue disorder characterized by vascular damage and collagen deposition in the skin and internal organs, resulting in a wide variety of signs and symptoms. The limited and diffuse subsets of scleroderma are defined in terms of skin involvement, ranging in extent from the distal extremities only to widespread involvement including the chest and trunk[1]. Progression of the disease is highly variable among patients and over time leads to a wide spectrum of organ damage, including pulmonary fibrosis, gastroesophageal reflux, malabsorption, renal failure, arthritis, digital ulcers, and hand contractures[2,3].

The Health Assessment Questionnaire-Disability Index (HAQ-DI)[4] is a self-report of physical function originally developed to measure function in arthritis[5,6]. Nevertheless, it has also been validated as a measure of disability in SSc[7,8,9] and used in other rheumatic diseases[10,11,12,13] and in general epidemiologic research[14].

Disability in patients with SSc, measured using the HAQ, has been shown to be considerable and to be associated with skin and other disease-related manifestations[7]. However, little is known on how disability progresses over time in SSc. Longitudinal analysis of disability in SSc could yield important prognostic information as well as indicate general trends in how disability and specific disease-related manifestations are related in SSc.

A major potential confounder in longitudinal analysis of cohort data is informative patient dropout[15]. This occurs whenever patients stop participating in a study because of reasons associated with disease, such as worsening SSc, or death. The remaining cohort may appear to be doing better over time simply because the sicker patients are dropping out. Such biased results may be especially misleading when studying a disabling or fatal disease such as SSc, where informative dropout is of particular concern. Statistical models accounting for informative patient dropout exist and their implementation is relatively straightforward using standard software.

Our purpose was to assess disability, measured using the HAQ, longitudinally in patients with SSc and to identify disease-specific determinants, while taking into account potentially informative patient dropout.

*From the Jewish General Hospital and McGill University, Montreal, Quebec, Canada.*

*Funded in part by the Canadian Institutes of Health Research, the Scleroderma Society of Canada, and educational grants from Actelion Pharmaceuticals and Pfizer Inc. Dr. Hudson is a New Investigator funded by the Canadian Institutes of Health Research.*

*M. Schnitzer, MSc; M. Hudson, MD, MPH; M. Baron, MD; R. Steele, PhD, Jewish General Hospital and McGill University.*

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

*Accepted for publication November 8, 2010.*

## MATERIALS AND METHODS

*Study subjects*. Our study sample consisted of patients enrolled in the Canadian Scleroderma Research Group registry with baseline visits between September 2004 and September 2008. Patients in the registry are recruited from 15 centers across Canada. They must have a diagnosis of SSc made by the referring rheumatologist, be over 18 years of age, and be fluent in English or French. Patients in the registry undergo a yearly standardized evaluation including a detailed medical history, physical evaluation, and laboratory investigations. In particular, demographic and disease-related information is recorded, including age, sex, duration of the disease since the onset of the first non-Raynaud's symptoms, and disease subset (limited or diffuse). At each yearly visit, patients complete a HAQ and several other self-report questionnaires.

For our study, only patients who had complete baseline demographic information between September 2004 and September 2008 were included. This sample accounted for 91% of the total number of registered patients. The remainder of the patients had incomplete baseline demographic information because of the early loss of some study centers (and thus unrelated to disease status). Patients with and without followup visits were included. Those without followup visits contributed to the data for visit 1 and those with followup visits to the longitudinal data as well.

*Outcome measure*. The HAQ is a self-administered questionnaire intended to assess functional ability in patients with arthritis[4,5,6]. It includes 20 questions in 8 categories (dressing and grooming, standing, eating, walking, hygiene, reach, grip, and performing activities). The use of assistive aids and devices to help with function is also recorded. The patient is asked to rate his/her difficulty over the past week in performing the specific tasks in each category on a scale of 0 to 3 (without difficulty = 0, with some difficulty = 1, with much difficulty or with assistance = 2, unable = 3). The most abnormal score in each of the 8 categories becomes the score for that category. The final score for the HAQ is obtained by adding the scores for all 8 categories and dividing by 8. It ranges from 0 (no disability) to 3 (severe disability).

*Covariates*. The covariates (or explanatory variables) of interest included both baseline and time-dependent (longitudinal) variables. The baseline characteristics of interest were age, sex, disease duration, and disease subset. The time-dependent variables included study visit number (i.e., the longitudinal time covariate) and the 5 SSc disease-specific questions developed to be administered along with the HAQ[7]. These include questions on the severity of Raynaud's phenomenon, digital ulcers, gastrointestinal symptoms, breathing problems, and pain in the past week. Each question is anchored by the descriptors "does not interfere" and "very severe limitation" and is scored separately. Unlike the 15-cm visual analog scales used originally with these questions, the assessments in our study were made using numerical rating scales ranging from 0 to 10. Because pain may act as a mediator of the effect of the 4 other disease symptoms on disability, pain was excluded as a covariate in the primary analyses (although it was still used in certain models as predictive of the unobserved data). Secondary analyses were also conducted to determine whether the presence of pain in the model modified the results.

*Study visits*. Patients in the registry are scheduled to return yearly after their initial (baseline) visit. Patients may intermittently miss yearly study visits or drop out of the registry entirely (because of death or other reasons). It is also possible that patients fail to complete their evaluations during an attended study visit. We attempted to collect information on the reasons for partially incomplete or unattended visits, specifically whether the missing data were due to SSc.

In practice, patient visits do not always occur on their originally scheduled date. Subsequent attended study visits were numbered 2, 3, and 4, provided that they fell within 6 months of the originally scheduled date. We used visit numbers rather than the actual time since baseline in order to avoid measurement error confounding. Patients may delay scheduled appointments if they are particularly sick, which would increase the intervals between visits. However, there may be other reasons for variation in the timing of appointments (e.g., physician schedules, patient lack of proximity to the site, and traveling conditions). Rather than having a more precise but potentially misleading unit of time (time since baseline), we felt that a more coarse measure (visit number) would be both more conservative and more robust. If 2 visits fell within the same 1-year interval around the expected appointment date, the first was dropped and the latter was used in this analysis. If an unscheduled followup visit occurred within 6 months of the baseline visit, it too was dropped. However, only 8 of the 745 subjects in our study had any visits dropped and it is therefore unlikely that the dropped visits would have influenced the results significantly.

*Missing data*. The treatment of missing data depends on how we assume the "missing-ness" is related to the observed and unobserved data. Data are called missing completely at random (MCAR) if their absence is assumed independent of any data, observed or otherwise (for example, patient dropout because of financial issues or participating physician withdrawal from our study was considered MCAR). If missing values in a dataset can be explained completely by additional observed variables, then they are called missing at random (MAR). However, if the missingness cannot be explained by observed variables, it is called not missing at random (NMAR)[16]. The statistical models used in our study each rely on one of these 3 assumptions. The interpretation of the results is therefore dependent on the appropriateness of the missingness assumption.

Prior to undertaking the statistical analysis, each missing appointment was recorded where it should have occurred, and entered as an empty visit in the dataset. An appointment was considered to be missing if one did not occur within 6 months of the scheduled date (the scheduled date being exactly 1, 2, 3, or 4 years after the baseline visit). For patients who left our study, missed appointments were also considered to have occurred every year after their final visit, up until a year and 3 months before the end date of our study (September 2008).

Thereafter, using additional information from the registry database, some of the missing or incomplete visits were coded as being MCAR if they were found to have nothing to do with the disease. Otherwise, all visit and item missingness was considered potentially informative (either MAR or NMAR, depending on the chosen model). If the response (i.e., HAQ) was missing from any patient visit, and that visit was coded as MCAR, then that individual visit was not used in our study (even if there was additional covariate information available). Henceforth, "missing value" refers only to informative missingness, and not to the visits that were removed.

*Statistical analysis*. Descriptive statistics were used to summarize the baseline characteristics of the sample, and the extent of the missing values. In order to determine the overall trend of the HAQ over time, 2 types of models were fit: the complete-case mixed model and the Bayesian selection model.

*Complete-case mixed model*. This type of model incorporated all subjects with complete information (defined here as having no missing visits or HAQ scores) into an unbalanced mixed model using the R software package "nlme." This standard type of analysis makes a simplifying assumption that all of the data missing from the dataset is MCAR[17]. Hence, it does not account for informative patient dropout. Visit number was included as a covariate in the analysis to model the longitudinal aspect of the disease and to look for trends in time. The covariates were selected using a backward and forward stepwise model-building approach initially with only the baseline variables. For the results presented here, we assumed an independent error structure within groups, although other dependent error structures, which all resulted in inferior Akaike's Information Criterion and Bayesian Information Criteria values, were also considered.

*Bayesian selection model*. Because the mixed model does not account for the potentially informative nature of our patient dropout, a fully Bayesian selection model with a truncated mixed-model likelihood for the response was used (i.e., missing HAQ values were restricted to the interval 0–3). The missingness of data (i.e., selection) probabilities were modeled as linear in the baseline covariates and the response (following the Bernoulli model[18]). The value of the coefficient for the response in the missingness model, ω, was predetermined. Set to zero, the model utilizes a MAR assumption, and when chosen to be nonzero, the missingness mechanism is NMAR. Two nonzero values of ω were used (0.695 and 1.39). A value of ω = 0.695 corresponds to an assumed 15% increased risk of a missing visit for each increase in HAQ of 0.2. The less conservative value of ω = 1.39 corresponds to an assumed 15% increased risk of missingness for each increase in HAQ of 0.1. Larger values of ω correspond to increasingly informative NMAR dropout because of disability. Measurements of interest were drawn with Markov-chain Monte Carlo using the programs OpenBUGS and the R package BRugs. After a 10,000-draw burn-in period, 3 chains were run for 100,000 samples and thinned down to 10,000 samples by taking every tenth sample. Diagnostic plots were used to verify the convergence and level of interdependence of these draws.

To ascertain the association between disability and other time-varying SSc-related problems, the 2 classes of models (mixed and Bayesian selection) were refit. In these models, the pool of covariates included both the baseline and time-dependent variables (with the exception of the pain variable). In the Bayesian selection models, the time-dependent variables were hierarchically modeled (providing a model for the missing values).

In the secondary analysis with pain, the mixed and Bayesian selection models (from the analysis investigating the association between disability and other SSc-related problems) were again refit with pain as an added covariate.

The results of the analyses were presented using standardized coefficients, which enables a unit-free comparison of the magnitude of association between the HAQ-DI score and the covariates in terms of standard deviations. Raw (unstandardized) covariate estimates were also reported for the HAQ in order to demonstrate the magnitude of the change in disability score.

*Ethical considerations.* Ethics committee approval for our study was obtained at each site and each patient provided written informed consent to participate.

## RESULTS

*Characteristics of the cohort and exploratory summaries.* Our study included 745 patients. Mean age (SD) of the patients was 55 (± 12) years; 86% were women; mean duration of disease measured from the onset of the first non-Raynaud's disease manifestation of SSc was 11 (± 9) years; 59% of patients had limited and 41% diffuse SSc. At baseline, mean HAQ score was 0.84 (± 0.70), indicating overall mild to moderate disability in our cohort.

Figure 1 displays the flow of patients over time. Patients were classified at each timepoint as having (1) complete visits (and therefore progressed to the subsequent visit); (2) incomplete visits (and considered MCAR, such as the next appointment was not yet due, the patient left the registry for reasons unrelated to the disease, or incomplete data entry); (3) skips (if the patient was unobserved for an intermediate visit); and (4) dropout (the patient left the registry for rea-

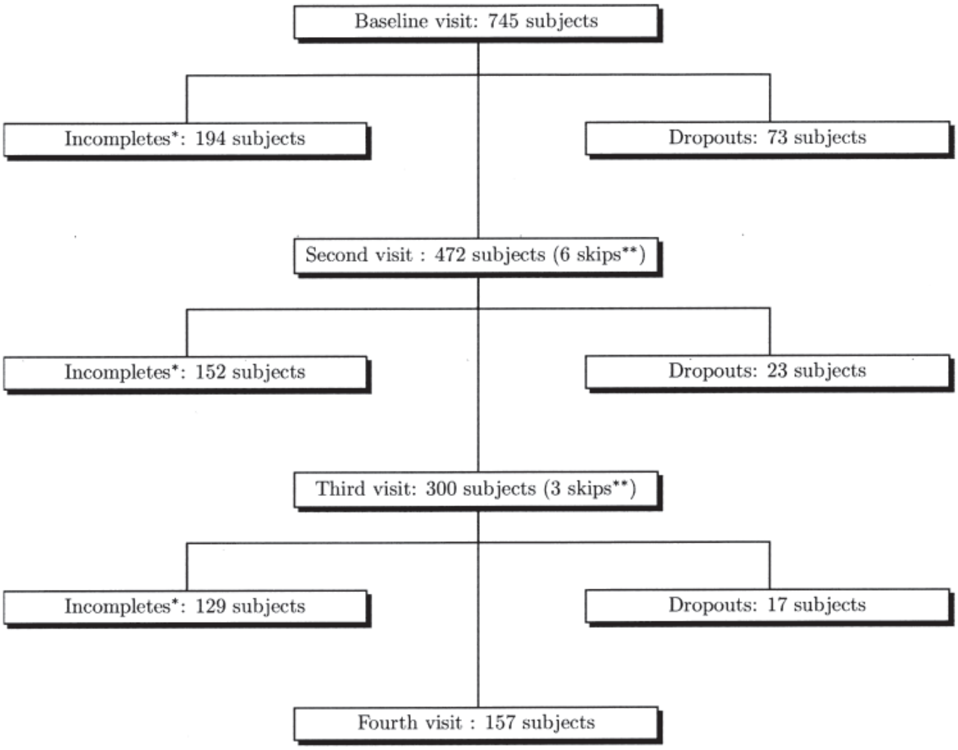
Baseline visit: 745 subjects
Incompletes*: 194 subjects
Dropouts: 73 subjects
Second visit : 472 subjects (6 skips**)
Incompletes*: 152 subjects
Dropouts: 23 subjects
Third visit: 300 subjects (3 skips**)
Incompletes*: 129 subjects
Dropouts: 17 subjects
Fourth visit : 157 subjects

*Figure 1.* Observed patient visits (center column), patient dropouts after each visit (right column, noncumulative), and patients whose participation was incomplete (left column). *Incomplete refers to subjects who did not proceed to their next appointment because of conditions that we consider "completely at random" (e.g., next appointment has not yet occurred, or leaving project for reasons unrelated to the disease). ** Some patients were unobserved for intermediate visits. Some of these skipped visits were deemed "missing completely at random" and hence not included in the final subset.

sons that are not MCAR, such as death due to SSc). In all, there were 1655 attended visits and 207 missed visits considered potentially informative. Thus, the total number of visits (both empty and attended) included in this analysis was 1862.

Mean scores for the HAQ and the disease-specific questions were computed for each visit number, along with the proportion of missing data for that variable at that visit (Table 1). For instance, 3 patients out of 745 (0.4%) did not have a HAQ score for their baseline visit, and 80 of the 445 (18%) who attended (or informatively missed) their second appointment did not have a HAQ score recorded for that second visit. At the fourth visit, the mean HAQ-DI score decreased from previous visits (0.81 from 0.85), indicating improvement for those patients with complete data. The mean scores for the severity of Raynaud's (baseline 3.01), finger ulcers (baseline 2.10), gastrointestinal problems (baseline 1.92), and pain (baseline 3.66) fluctuated over the 4 visits. The mean score for breathing problems improved gradually from baseline (2.08) to the fourth visit (1.97). However, interpreting these changes as evidence of trends for patients with SSc would be misleading because these changes ignore the informative process of dropout. Specifically, if sick patients were to leave the registry before their subsequent visits, we would in fact observe less disability on average in the cohort for visits 3 and 4.

*Table 1.* HAQ and disease-specific symptom scores, by visit. The percentage missing refers to the proportion of patients unobserved for that variable during their visit, or unobserved for the entire visit (for reasons that might be related to their disease).

| Scores | Mean (SD) | No. Available | % Missing |
|---|---|---|---|
| HAQ | | | |
| Baseline | 0.84 (0.70) | 742 | 0.4 |
| Visit 2 | 0.85 (0.69) | 445 | 18 |
| Visit 3 | 0.85 (0.72) | 281 | 24 |
| Visit 4 | 0.81 (0.71) | 155 | 23 |
| Raynaud's | | | |
| Baseline | 3.01 (2.98) | 743 | 0.3 |
| Visit 2 | 2.97 (2.95) | 446 | 18 |
| Visit 3 | 3.04 (3.01) | 282 | 24 |
| Visit 4 | 3.04 (3.07) | 155 | 23 |
| Finger ulcers | | | |
| Baseline | 2.10 (3.03) | 739 | 0.8 |
| Visit 2 | 2.00 (2.92) | 442 | 19 |
| Visit 3 | 2.31 (3.13) | 282 | 24 |
| Visit 4 | 2.26 (3.09) | 155 | 23 |
| Gastrointestinal problems | | | |
| Baseline | 1.92 (2.68) | 738 | 0.9 |
| Visit 2 | 1.71 (2.50) | 444 | 19 |
| Visit 3 | 1.90 (2.74) | 281 | 24 |
| Visit 4 | 1.71 (2.40) | 155 | 23 |
| Breathing problems | | | |
| Baseline | 2.08 (2.57) | 739 | 0.8 |
| Visit 2 | 2.03 (2.63) | 445 | 18 |
| Visit 3 | 2.01 (2.71) | 281 | 24 |
| Visit 4 | 1.97 (2.58) | 155 | 23 |
| Pain | | | |
| Baseline | 3.71 (2.79) | 743 | 0.3 |
| Visit 2 | 3.57 (2.73) | 445 | 18 |
| Visit 3 | 3.53 (2.87) | 281 | 24 |
| Visit 4 | 3.45 (2.77) | 155 | 24 |

HAQ: Health Assessment Questionnaire.

*HAQ over time.* The complete-case mixed model required the removal of all subjects with any missed visits or missing HAQ values. Therefore, all patients who dropped out, and all those who skipped visits for reasons that were not MCAR, were removed. Patients with fewer than 4 visits but who had only MCAR missingness were included. Thus, for this analysis, we had a sample size of 585 subjects with 1360 visits in total.

The results of the mixed model for complete subjects only are presented in Table 2 (first column). Age at baseline (SE 0.11, 95% CI 0.03–0.18) and diffuse disease (SE 0.23, 95% CI 0.16–0.31) were both associated with a higher HAQ, indicating that older patients and those with diffuse disease generally have higher disability. The association between visit number and HAQ was also positive (standardized coefficient 0.031, 95% CI 0.003–0.060), indicating worsening disability over time. The raw (unstandardized) change in HAQ score was 0.022 (95% CI 0.002–0.042), which means that the average increase in HAQ score would be 0.022 per year [i.e., 0.07 (3 × 0.022) over 3 years].

Of note, disease duration and sex were not found to be significant on their own or with the other variables included in the final model (data not shown). Moreover, model fit was significantly improved when including a random intercept and random visit number (so that each patient was modeled as having their own intercept and slope over visits).

With $\omega$ set to zero, the selection model in Table 2 (second column) assumes a MAR mechanism, thereby accounting for potentially informative patient dropout. The coefficients for the baseline covariates were almost identical to the estimates produced by the mixed model. However, the unstandardized estimate for visit number increased to 0.039 (a 77% increase from the estimate obtained using the mixed model). This would imply an increase in HAQ of 0.12 points (3 × 0.039) over 3 years.

The selection model with $\omega = 0.695$ assumes an NMAR mechanism and that an individual increase in HAQ of 1 would double the probability of a missing HAQ score (Table 3). This model produced an estimate of 0.055 for the unstandardized visit number (i.e., increase in HAQ of 0.17 over 3 years), increasing the slope estimate obtained for the selection MAR model by over 40%. The selection model with $\omega = 1.39$ also assumed an NMAR mechanism but that an increase of HAQ score by 0.5 would double the probability of a missing value. This less conservative assumption produced an estimated slope of 0.071 (i.e., increase in HAQ of 0.21 over 3 years), which was again a large increase from the previous estimates. The coefficient estimates of the

*Table 2.* Results of fixed-effects models to identify changes in disability over time in SSc. Unless indicated, results are standardized using the values obtained at baseline visits.

| Variable | HAQ Over Time: Standardized Fixed-effect Estimates<br>Mixed Model<br>Estimate (95% CI) | <br>Selection Model ω = 0<br>Estimate (95% CI) |
|---|---|---|
| Age | 0.11 (0.03, 0.18) | 0.10 (0.02, 0.17) |
| Diffuse disease | 0.23 (0.16, 0.31) | 0.26 (0.19, 0.33) |
| Visit number | 0.031 (0.003, 0.060) | 0.056 (0.025, 0.087) |
| Unstandardized visit no. | 0.022 (0.002, 0.042) | 0.039 (0.018, 0.061) |
| Variable | HAQ-related Symptoms Over Time: Standardized Fixed-effect Estimates<br>Mixed Model<br>Estimate (95% CI) | <br>Selection Model ω = 0<br>Estimate (95% CI) |
| Age | 0.12 (0.05, 0.18) | 0.11 (0.05, 0.17) |
| Diffuse disease | 0.23 (0.16, 0.29) | 0.24 (0.18, 0.30) |
| Raynaud's | 0.06 (0.02, 0.10) | 0.07 (0.03, 0.11) |
| Finger ulcers | 0.09 (0.05, 0.13) | 0.09 (0.05, 0.13) |
| Intestinal problems | 0.07 (0.03, 0.10) | 0.06 (0.03, 0.10) |
| Breathing problems | 0.23 (0.19, 0.28) | 0.21 (0.17, 0.25) |
| Visit number | 0.025 (–0.001, 0.051) | 0.049 (0.018, 0.079) |
| Unstandardized visit no. | 0.018 (–0.001, 0.036) | 0.034 (0.013, 0.056) |

HAQ: Health Assessment Questionnaire.

*Table 3.* Results of models assuming data not missing at random. Unless indicated, results are standardized using the values obtained at baseline visits.

| Variable | Models with Baseline Covariates<br>Selection Model<br>ω = 0<br>Estimate (95% CI) | <br>Selection Model<br>ω = 0.695<br>Estimate (95% CI) | <br>Selection Model<br>ω = 1.39<br>Estimate (95% CI) |
|---|---|---|---|
| Visit number | 0.056 (0.025, 0.087) | 0.078 (0.046, 0.110) | 0.100 (0.068, 0.134) |
| Unstandardized visit no. | 0.039 (0.018, 0.061) | 0.055 (0.033, 0.077) | 0.071 (0.048, 0.094) |
| Variable | Models with Time-dependent Covariates<br>Selection Model<br>ω = 0<br>Estimate (95% CI) | <br>Selection Model<br>ω = 0.695<br>Estimate (95% CI) | <br>Selection Model<br>ω = 1.39<br>Estimate (95% CI) |
| Visit number | 0.049 (0.018, 0.079) | 0.069 (0.039, 0.100) | 0.090 (0.058, 0.121) |
| Unstandardized visit no. | 0.034 (0.013, 0.056) | 0.049 (0.027, 0.070) | 0.063 (0.041, 0.085) |

baseline covariates in these models were essentially the same as in the previous model.

There was a large degree of heterogeneity in the slopes for the subjects in the selection models as well. The standard deviation of the slopes was actually 5 times as large as the magnitude of the mean slope. This suggested that, although there is overall worsening in the HAQ over time, the trajectories of individual patients might be quite heterogeneous.

*Symptoms associated with disability.* An investigation of the associations between specific disease manifestations and the HAQ was undertaken (Table 2). Using a mixed model for the complete cases, age, disease subset, and time-dependent covariates for Raynaud's, finger ulcers, gastrointestinal problems, and breathing problems were all found to be significantly associated with the HAQ. In particular, diffuse disease (SE 0.23) and breathing problems (SE 0.23) were most strongly associated with a contemporaneous increase in the HAQ. In this model, visit number was not significant. Thus, the variability of the HAQ over time was explained by the other covariates.

The results for the MAR selection model (Table 2) were almost identical to those of the mixed model. In the MAR selection model, though, the estimate of the coefficient for visit number was much larger and significant (SE 0.049, 95% CI 0.018–0.079).

The estimates for the covariates of the NMAR selection model with ω = 0.695 (Table 3) were very similar to the results of the MAR selection model (ω = 0), with the most notable difference being a 44% increase in the estimate of the coefficient for visit number (i.e., an absolute increase in unstandardized visit number slope of 0.015). The final selection model with $\omega = 1.39$ produced few changes in the estimation of the strengths of associations with HAQ. The most closely linked disease-symptom covariates remained diffuse disease and breathing problems. Again, a large increase in the visit number coefficient was observed, although with a 0.014-point absolute increase in the unstandardized coefficient (Table 3). The SD of the random effects (data not shown) indicated a substantial amount of variability of the individual patient slopes, again indicating a wide variety of trajectories among subjects.

*Secondary analyses.* Because the causal relationship between pain and function is complex, pain was not included in the main analysis. However, we performed a secondary analysis including pain as a time-dependent covariate, in addition to severity of Raynaud's, finger ulcers, gastrointestinal problems, and breathing problems. The results were similar to those from the models without pain (data not shown). Visit number was consistently found to be significant over and above the variables included in the model.

The Bayesian NMAR analyses described above used a truncated linear mixed model. An alternative non-normal model for the response was investigated. The HAQ can be viewed as a binomial variable of 24 trials, as the HAQ multiplied by 8 is the raw score out of 24. Hence, the Bayesian selection models were modified to include a logistic model of interest for the raw score of the HAQ. This logistic model returned an almost identical interpretation to the truncated linear mixed model (data not shown), confirming our original model choice and the results of our analysis.

## DISCUSSION

In this large, multicenter cohort study, we found strong evidence that disability worsens over time in SSc. Although the magnitude of worsening varied according to the statistical models used, there was strong agreement of worsening between models. Among the models accounting for informative patient dropout, the most conservative slope estimate indicated an overall increase of 0.12 in HAQ over 3 years. However, if one assumes that the probability of a missing visit depends on the current level of patient function, we obtained estimates of increase in HAQ over 3 years of up to 0.21. The magnitude of this finding is greater than that observed in other rheumatic diseases. For example, a 3-year increase in HAQ between 0.03 and 0.036 was observed in 2 studies of patients with rheumatoid arthritis[10,19].

The strongest determinants of disability over time in SSc were diffuse disease and the severity of breathing problems. This was a consistent finding of every model. These results further justify current recommendations that emphasize the management of skin and cardiorespiratory disease in SSc[20]. In addition, the fact that the estimate for visit number remained significant after also accounting for the severity of Raynaud's, gastrointestinal symptoms, and finger ulcers suggests that some component of disability remains unexplained by those covariates. There are numerous other possible sources of disability in SSc, in particular joint pain and contractures, fatigue, and depression, that remain to be studied.

In a single-center study, Steen and Medsger were the first to investigate changes in the HAQ over time in SSc[7]. In their study, they were able to identify important predictors of the HAQ, including skin and respiratory disease, and to show that the HAQ was an excellent predictor of survival in SSc. However, the authors acknowledged that their data had limitations, in particular that measurements of the HAQ and clinical variables were not always simultaneous, that HAQ scores were collected at variable intervals, and that only 42% of their patients completed more than 2 HAQ questionnaires. Hence, the precise relationship between the HAQ and time could not be fully studied in their analysis. The authors called for additional "prospective studies with the HAQ administered at regular intervals" to further clarify issues related to change in disability over time. Our analysis is therefore highly complementary to that study. In addition, our study has important additional strengths. First, while Steen and Medsger used very simple modeling of the patient HAQ scores (looking only at first and last HAQ values), our analysis used all the available data to examine change. Second, the original article[7] did not account for the potential for heterogeneity between patients with respect to change over time, while our proposed models did. Third, Steen and Medsger did report some improvements in the HAQ in certain subgroups of patients over time, in particular in patients treated with D-penicillamine[7]. However, a key part of our analysis was to address the possibility that patient dropout might affect measured change over time. Without accounting for patient dropout, there is a risk that only healthier patients or patients who tolerate a medication remain in a cohort, and measured HAQ values may appear to improve because of this selection bias. Therefore, we believe that our multicenter study of 745 patients with SSc followed with yearly HAQ scores up to 4 years using advanced statistical modeling provides robust estimates of changes in the HAQ over time that have never been previously published.

Several randomized clinical trials (RCT) have used the HAQ as outcome measure. However, patients in RCT are fundamentally different from the general population of patients with SSc and therefore data on the HAQ from the RCT are not easily comparable to the data presented here. For example, in the Scleroderma Lung Study[21,22], all the patients had early disease with lung involvement. Such a sample of patients does not provide an opportunity to summarize the broad patterns of functional decline in patients with SSc overall or the inherent heterogeneity in such a sample. Moreover, in RCT there is an intervention. Again, in the Scleroderma Lung Study, HAQ scores fell by 0.07 in the cyclophosphamide-treated group, while they increased by 0.11 in the placebo group, a difference that was highly statistically significant (p = 0.0001). However, our study reports on the natural history of the HAQ, with no specific intervention, and shows a progressive decline. Finally, RCT are usually short (typically 1 year or less). Our study, on the other hand, reflects clinical progression of a large heterogeneous cohort of patients with SSc managed with standard of care over several years (with 300 patients achieving 3 years of visits and over 150 patients with 4 visits). Therefore, our study better reflects the natural history of the disease and is generalizable to a larger spectrum of patients with SSc than RCT. In that sense, our study is also an important contribution to the current knowledge concerning the HAQ in SSc over time.

The minimum clinically important difference (MCID) is the smallest change in the score of a health measure that patients can perceive[23]. MCID may be different depending on the direction of change, either improvement or worsening[24]. The MCID for the HAQ in SSc has been suggested to be up to 0.14[25,26]. Our conservative finding that the HAQ increases by 0.12 over 3 years suggests that this change is in the range of a clinically meaningful change for patients with SSc.

Our findings also have implications in so far as the design of future studies is concerned. Experts have recommended the HAQ as an outcome measure for SSc[27,28,29]. However, if it is to be used as a primary outcome, temporal changes inherent in the disease also need to be considered in calculating sample sizes.

Modeling patient dropout has become an extremely important consideration in longitudinal data analyses. The importance of considering models other than the traditional ones that assume that data are MCAR is clearly shown in our study. The complete-subject data significantly underestimated the extent of worsening disability revealed by the models that accounted for informative patient dropout.

Our models that assume a relationship between missingness and current function are best viewed as sensitivity analyses. In other words, the results of these models can be seen as establishing a spectrum of estimates for the change in disability over time under increasingly less conservative assumptions. Therefore, the conclusion derived from these models is that there is evidence of worsening disability in SSc over time, and the decline may actually be much steeper than a change of HAQ of 0.12 over 3 years identified in the model that makes the implausible assumption that current HAQ does not affect the probability of a missing visit.

Other methods also exist that would have allowed adjustment for the missing data. Multiple imputation methods have become quite popular and are less computationally demanding than Bayesian selection models. However, multiple imputation methods are less easily implemented under NMAR assumptions. In other analyses (not reported here), we found that multiple imputation using the PAN software[30] yielded results similar to those obtained using the mixed model for complete cases. PAN does not restrict imputed HAQ scores to be between 0 and 3, which is likely the reason that it gave different results to our Bayesian selection model with $\omega = 0$. Other methods using inverse probability weighting have also become popular because of their double robustness to model specification[31]. However, these doubly robust methods, despite their theoretical advantages, are not without their drawbacks. They can be computationally challenging to implement[32] and are prone to erratic behavior that must sometimes be addressed through ad hoc means[33].

The main limitation of our findings is that, although mean disability worsened considerably in our cohort over time, we found evidence of heterogeneity among individual patients. Indeed, SSc is highly heterogeneous in terms of manifestation and progression and the overall estimates we found do not clearly identify individual patient trajectories. Thus, caution must be exercised in extrapolating group data to individual patients. On the other hand, the strength of our study lies in its large sample and detailed data analysis accounting for informative patient dropout. We believe that our analysis provides robust estimates of worsening of disability in SSc over time.

This is the largest longitudinal study of disability in SSc to date. Although the finding that HAQ scores worsen significantly over time is intuitive, our study provides robust data estimating the magnitude of this change. Our data provide important prognostic information and has ramifications for future studies using the HAQ as outcome measure.

## APPENDIX

*List of study collaborators.* Canadian Scleroderma Research Group Investigators: J.E. Pope, London, Ontario; J. Markland, Saskatoon, Saskatchewan; D.A. Masetto, Sherbrooke, Quebec; E. Sutton, Halifax, Nova Scotia; N.A. Khalidi, Hamilton, Ontario; D. Robinson, Winnipeg, Manitoba; N. Jones, Edmonton, Alberta; E. Kaminska, Hamilton, Ontario; P. Docherty, Moncton, New Brunswick; C.D. Smith, Ottawa, Ontario; J-P. Mathieu, Montreal, Quebec; S. Ligier, Montreal, Quebec; and S. Mittoo, Winnipeg, Manitoba.

## REFERENCES

1. Leroy EC, Black C, Fleischmajer R, Jablonska S, Krieg T, Medsger TA, et al. Scleroderma (systemic sclerosis) — classification, subsets and pathogenesis. J Rheumatol 1988;15:202-5.
2. Siebold J. Scleroderma. In: Harris E, Budd R, Firestein G, Genovese M, Sergent J, Ruddy S, et al, editors. Kelley's textbook of rheumatology. 7th ed. Amsterdam: Elsevier; 2005.
3. Wigley F, Hummers L. Clinical features of systemic sclerosis. In: Hochberg M, Silman A, Smolen J, Weinblatt M, Weisman M, editors. Rheumatology. 3rd ed. Amsterdam: Elsevier; 2006.
4. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167-78.
5. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
6. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales.

J Rheumatol 1982;9:789-93.
7. 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.
8. 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;32:832-40.
9. Poole JL, Williams CA, Bloch DA, Hollak B, Spitz P. Concurrent validity of the Health Assessment Questionnaire Disability Index in Scleroderma. Arthritis Care Res 1995;8:189-93.
10. Sokka T, Kautiainen H, Hannonen P, Pincus T. Changes in Health Assessment Questionnaire disability scores over five years in patients with rheumatoid arthritis compared with the general population. Arthritis Rheum 2006;54:3113-8.
11. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? Arthritis Rheum 2007;56:840-9.
12. Ward MM. Predictors of the progression of functional disability in patients with ankylosing spondylitis. J Rheumatol 2002;29:1420-5.
13. Wiles N, Dunn G, Barrett E, Silman A, Symmons D. Associations between demographic and disease-related variables and disability over the first five years of inflammatory polyarthritis: a longitudinal analysis using generalized estimating equations. J Clin Epidemiol 2000;53:988-96.
14. Chakravarty EF, Hubert HB, Lingala VB, Fries JF. Reduced disability and mortality among aging runners: a 21-year longitudinal study. Arch Intern Med 2008;168:1638-46.
15. Shen S, Beunckens C, Mallinckrodt C, Molenberghs G. A local influence sensitivity analysis for incomplete longitudinal depression data. J Biopharm Stat 2006;16:365-84.
16. Schafer JL. Analysis of incomplete multivariate data. London: CRC Press; 1997.
17. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.
18. Boscardin WJ, Yan X, Wong WK. A reanalysis of a longitudinal scleroderma clinical trial using non-ignorable missingness models. J Stat Plan Inference 2007;137:3848-58.
19. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. Arthritis Rheum 2000;43:2751-61.
20. Kowal-Bielecka O, Landewe R, Avouac J, Chwiesko S, Miniati I, Czirjak L, et al. EULAR recommendations for the treatment of systemic sclerosis: a report from the EULAR Scleroderma Trials and Research group (EUSTAR). Ann Rheum Dis 2009;68:620-8.
21. Khanna D. Health-related quality of life: a primer with a focus on scleroderma. Sclero Care Res 2006;3:3-13.
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;354:2655-66.
23. Jaeschke R, Singer J, Guyatt G. Health-status measurement — ascertaining the minimal clinically important difference. Clin Res 1989;37:A315.
24. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18:419-23.
25. Khanna D, Furst DE, Hays RD, Park GS, Wong WK, Seibold JR, et al. Minimally important difference in diffuse systemic sclerosis: results from the D-penicillamine study. Ann Rheum Dis 2006;65:1325-9.
26. Sekhon S, Pope J, Baron M. The minimally important difference in clinical practice for patient-centered outcomes including Health Assessment Questionnaire, fatigue, pain, sleep, global visual analog scale, and SF-36 in scleroderma. J Rheumatol 2010;37:591-8.
27. Distler O, Behrens F, Pittrow D, Huscher D, Denton CP, Foeldvari I, et al. Defining appropriate outcome measures in pulmonary arterial hypertension related to systemic sclerosis: a Delphi consensus study with cluster analysis. Arthritis Rheum 2008;59:867-75.
28. 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;34:1194-200.
29. 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;67:703-9.
30. Schafer JL. Imputation of missing covariates under a multivariate linear mixed model. University Park, PA: Department of Statistics, Pennsylvania State University; 1997. [Internet. Accessed Nov 10, 2010.] Available from: http://www.stat.psu.edu/reports/1997/tr9704.pdf
31. Gonzalez-Alvaro I, Descalzo MA, Carmona L. Trends towards an improved disease state in rheumatoid arthritis over time: influence of new therapies and changes in management approach: analysis of the EMECAR cohort. Arthritis Res Ther 2008;10:R138.
32. Carpenter JR, Kenward MG, Vansteelandt S. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J Roy Statist Soc: Series A (Statistics in Society) 2006;169:571-84.
33. Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 2007;22:523-39.