A comparison of the measurement properties and estimation of minimal important differences of the EQ-5D and SF-6D utility measures in patients with systemic sclerosis

L. Kwakkenbos[1,2,3], J. Fransen[4], M.C. Vonk[4], E.S. Becker[5], M. Jeurissen[1], F.H.J. van den Hoogen[1], C.H.M. van den Ende[1]

[1]Department of Rheumatology, Sint Maartenskliniek, Nijmegen, The Netherlands;
[2]Department of Psychiatry, McGill University, Montreal, Canada;
[3]Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada;
[4]Department of Rheumatology, Radboud University Nijmegen Medical Centre The Netherlands;
[5]Behavioural Science Institute, Clinical Psychology, Radboud University Nijmegen, Nijmegen, The Netherlands.

Linda Kwakkenbos, PhD
Jaap Fransen, PhD
Madelon C. Vonk, MD PhD
Eni S. Becker, PhD
Maurice Jeurissen, MD PhD
Frank H.J. van den Hoogen, MD PhD
Cornelia H.M. van den Ende, PhD

Please address correspondence and reprint requests to:
Linda Kwakkenbos, PhD,
Jewish General Hospital,
4333 Cote Ste Catherine Road,
Montreal,
Quebec H3T 1E4, Canada.
E-mail: kwakkenbosL@gmail.com

Received on June 25, 2012; accepted in revised form on April 19, 2013.

Clin Exp Rheumatol 2013; 31 (Suppl. 76): S50-S56.

Key words: scleroderma, systemic, quality of life, psychometrics, outcome assessment (health care), economics

Competing interests: none declared.

ABSTRACT

Objective. *To compare measurement properties of the EQ-5D and SF-6D utility measures, to assess the association and agreement between these measures and to estimate minimal important differences (MID) in patients with systemic sclerosis (SSc).*

Methods. *Both measures were assessed twice in an observational prospective design over a 12-month period (n=211). Spearman's rank correlation between the EQ-5D and SF-6D was calculated at baseline. Agreement was assessed using Lin's concordance coefficient (LCC) and a Bland-Altman plot. MIDs were estimated using three anchors; the global rating of change item (SF-36) and changes on the Health Assessment Questionnaire-Disability Index (HAQ-DI) of ≥0.14 and ≥0.22.*

Results. *At baseline, the mean EQ-5D and SF-6D scores were 0.64 (SD=0.25) and 0.65 (SD=0.11), respectively. The correlation between EQ-D and SF-6D scores was r=0.74. Agreement was moderate (LCC=0.49), and the Bland-Altman plot showed a mean difference of 0.003 but wide limits of agreement (-0.38 to 0.39) and a structural bias for lower scores. The mean MID estimate for the EQ-5D was 0.08 for the improved subgroup, and -0.13 for the deteriorated subgroup. For the SF-6D, the MID estimate was 0.05 for the improved and -0.04 for the deteriorated subgroup.*

Conclusion. *Although there was a marked correlation between the measures, the moderate agreement implies that EQ-5D and SF-6D scores cannot be used interchangeably. The MID estimates we provided can be used to calculate sample sizes for clinical trials involving SSc patients, and in interpreting the relevance and importance of treatment effects.*

Introduction

Self-administered utility measures are increasingly used in economic evaluations of treatments and policy-making (*e.g.* 1, 2). Utility measures like the EQ-5D (EuroQol 5D) (3, 4) and SF-6D (5) (derived from the Medical Outcomes Trust Short Form-36; SF-36) (6) are designed to assess value of health in a single summary measure, with a value of 0 for death and 1 for perfect health. Utility measures cover different domains of health-related quality of life that might be influenced by (chronic) diseases, including pain, physical limitations and mental health.

The EQ-5D and SF-6D are frequently used and considered as well-established measures of utility across a diversity of chronic diseases, including rheumatic diseases (7, 8). Although both utility measures seem promising and are often used interchangeably, it is not clear whether this is justified given the differences in scoring and reports of low agreement between the two measures, especially for the lower and higher ranges of utility scores (9–11). Possible EQ-5D scores range from -.59 to 1.00 (negative scores reflect patient belief that health status is valued worse than death), while the SF-6D ranges from 0.29 to 1.00.

Because of the low agreement between measures, differences in calculations of cost-effectiveness outcomes have been reported (12). This decreases the comparability of (cost-) effectiveness studies tremendously and potentially leads to different resource allocation decisions. In previous studies, levels of agreement between the EQ-5D and SF-6D differed wildly across diseases and patient samples (11). Thus, the level of agreement should preferably be examined in a patient sample in which the

instruments will be applied as an outcome measure.

Improvement of health related quality of life is one of the most important goals of pharmacological as well as non-pharmacological interventions in many chronic diseases, particularly when there is no cure, as in systemic sclerosis (SSc, scleroderma). SSc is a rare disease of connective tissue, characterised by abnormalities of the vascular and immune system (13). Fibrosis leads to thickening of the skin and might lead to significant organ dysfunction in the internal organs as well. Usually, two clinical subtypes are recognised based on the extent of skin involvement: limited SSc, with skin involvement distal to the elbows and knees, with or without face involvement and diffuse SSc, with skin involvement proximal to the elbows and knees and the trunk (14).

SSc has an impact on physical as well as mental health related quality of life (15). Currently, there is no treatment available to cure SSc, but an increasing number of interventions are being developed and tested (*e.g.* 16-18). To improve health care, accurate measurement of outcomes is crucial. Studies assessing the psychometric properties of patient-reported outcome measures usually include properties such as construct validity, discriminative ability and responsiveness. In addition, in the evaluation and interpretation of the effectiveness of treatments, it is important to determine the clinical significance of change scores, and which changes in outcomes could be interpreted as meaningful improvements to patients and clinicians. The minimal important difference (MID) could be defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's (health care) management" (19). For a particular instrument, MIDs may vary by population and context, for instance the baseline from which the patient starts, and whether they are improving or deteriorating (20). Thus, MIDs for utility measures, notably the EQ-5D and SF-6D, should be estimated for the population in which the measures are used, for instance, among patients with SSc.

The SF-6D has been shown to be a valid measure of utility in a sample of patients with diffuse SSc (21), whereas the EQ-5D has not yet been tested in SSc. Therefore, the purpose of the present study was to compare measurement properties of the EQ-5D and SF-6D utility measures, to assess the association and agreement between these measures and to estimate minimal important differences (MID), effect sizes (ES) and standardised response means (SRM) in SSc patients. Patients with limited as well as diffuse disease were included in our sample.

Materials and methods

Design

Measurement properties of the EQ-5D and SF-6D were assessed using prospective data collection in the cohort study "Psychological factors in scleroderma" in patients with SSc. Patients completed both questionnaires at inclusion (baseline) and after 12 months follow-up. Details of this study were described elsewhere (22). The study was approved by the local medical ethics board (CMO Arnhem-Nijmegen 2008/109).

Patients and procedures

Patients with SSc of the departments of Rheumatology of the Sint Maartenskliniek or Radboud University Nijmegen Medical Center, both in Nijmegen, the Netherlands, were included between June 2008 and August 2009. All patients had a diagnosis of SSc according to the preliminary ACR classification criteria (23). Exclusion criteria for participation in the cohort were a life expectancy of less than a year, acute serious complications, severe psychiatric comorbidity (*e.g.* severe substance abuse, psychosis or dementia), other serious comorbidities (*e.g.* cancer) and insufficient competence in the Dutch language.

For the current analyses, patients were included if they completed either the EQ-5D or SF-6D at baseline. For 4 out of the 215 patients who completed the baseline measures, both the EQ-5D and SF-6D could not be calculated due to missing values (n=211). In addition, 4 patients had missing values for the EQ-5D and 12 patients for the SF-6D. Of the 211 patients who were included at baseline, 163 (77.3%) also completed the follow-up measure. Drop-out reasons were: death (n=8), illness (n=3), logistic reasons (n=11), unknown (n=26). Additionally, 9 patients had missing values preventing the calculation of both the EQ-5D and SF-6D scores at follow-up (n=154). Of the patients with follow-up, 6 patients had no EQ-5D score and 7 patients had no SF-6D score.

Assessments

i. Demographics and disease characteristics

Age, sex, marital status and current employment status were assessed by questionnaire. In addition, at baseline the attending rheumatologist assessed disease duration (time since onset of first non-Raynaud symptom), SSc limited or diffuse subtype (14), and the modified Rodnan skin score (mRSS), which is a rating of skin involvement ranging from 0 (no involvement) to 3 (severe thickening) in 17 body areas (total score range 0–51) (24).

The SF-36, EQ-5D, and Scleroderma modified disability index of the Health Assessment Questionnaire (SHAQ) were self-assessed by the patient at home using paper versions.

ii. EQ-5D (EuroQol 5D)

The EQ-5D (3, 4) is a 5-item standardised health-related quality of life questionnaire, measuring 5 dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). The items are rated from 1 (no problems) to 3 (extreme problems). The EQ-5D ranges from -0.59 to 1.00, with 1.00 indicating full health and 0 representing death. Since negative EQ-5D scores are possible, these indicate health status valued worse than death.

iii. Medical Outcomes Trust Short Form-36 (SF-36)

The SF-36 (6) measures eight domains of health status using 36 items; physical functioning (10 items), role-physical (4 items), bodily pain (2 items),

general health (5 items), vitality (4 items), social functioning (2 items), role-emotional (3 items) and mental health (5 items). In addition, one item assesses global rating of change (GRoC) on a 5-point scale from "Much better now than one year ago" (1) to "Much worse now than one year ago" (5). For the calculation of the SF-6D, 11 items are used covering six domains (5); social functioning, bodily pain, mental health, physical functioning, role-limitation and vitality. The SF-6D ranges from .29 to 1.00, with 1.00 indicating full health.

iv. Scleroderma Health Assessment Questionnaire (SHAQ)

The SHAQ (25) consists of the disability index of the Health Assessment Questionnaire (HAQ) (26) and six Visual Analogue Scales (VAS) measuring perceived severity of pain, digital ulcers, Raynaud's phenomenon, lung involvement, gastrointestinal problems and patient global assessment (PGA). A VAS was added to measure fatigue, since this is perceived as an important symptom of SSc (27). The HAQ- Disability Index score consists of 20 items, measuring 8 dimensions of functioning (dressing and grooming, ability to get up, eating, walking, personal hygiene, reach, grip strength and activities). The score of each dimension ranges from 0 (best function) to 3 (worst function), and the mean of these scores can be calculated as an indicator of overall physical functioning (HAQ-DI) with higher scores pointing to worse functioning. The HAQ was originally developed for use in Rheumatoid Arthritis (26) but has also demonstrated to be reliable and valid in patients with SSc (25).

Statistical analyses

Descriptive statistics were provided as mean and standard deviation (SD) or median (P25-P75) for continuous variables and percentages for categorical variables. For the EQ-5D and SF-6D, descriptive statistics (mean, median, SD, minimum, maximum and frequencies) were calculated at baseline and follow-up.

The COSMIN recommendations (28) were followed in the assessment of measurement properties, except for the recommendation on responsiveness. The COSMIN recommendation includes hypotheses testing and considers traditionally accepted responsiveness measures such as standardised response mean (SRM) and effect size (ES) inappropriate. However, since the SRM and ES are used to compare which outcome measure detects changes over time more accurately, it can be considered appropriate for this purpose, especially in the absence of a gold standard or external criterion (29). All patients had completed the EQ-5D and/or SF-6D measures at baseline. However, some patients had missing data at follow-up. Missing values for the EQ-5D and SF-6D at follow-up were imputed using baseline EQ-5D and SF-6D scores, as well as baseline HAQ-DI score and VAS scores for the different domains included in the SHAQ, and available demographic and disease parameters. The assumption of values missing at random was tested with regression models examining the association of baseline characteristics and missingness of the EQ-5D and SF-6D at follow-up, respectively, as well as *t*-tests comparing baseline characteristics between patients with and without missing values at follow-up (results not shown). Multiple imputation by chained equations was used to create 20 datasets, using 15 cycles for each dataset (30). Results of these 20 imputed datasets were combined following Ruben's rules (31).

Floor and ceiling effects were assessed by calculating the percentage of patients scoring the lowest and highest possible score, respectively. The association between the EQ-5D and SF-6D at baseline was assessed using Spearman's rank correlation coefficient, and agreement between measures was assessed using Lin's concordance coefficient (32) and a Bland-Altman plot (33). Post-hoc, the agreement between utility measures was further explored for EQ-5D scores <0.45 .

The construct validity of the EQ-5D and SF-6D was assessed by calculating the correlation of both measures with HAQ-DI scores, VAS scores and skin score (mRSS). *A priori,* it was hypothesised that both measures would have 1) at least a moderate correlation ($r>0.40$) with HAQ-DI scores and with PGA, 2) low correlations ($r\leq0.40$) with the 6 VAS scores, and 3) no correlation ($0.00\leq r\leq0.20$) with skin score (21). Correlations were interpreted following Franzblau (34): 0.00-.20 indicating no correlation, 0.21-0.40 indicating a low correlation, 0.41-0.60 indicating a moderate correlation, 0.61-0.80 indicating a marked correlation, and 0.81-1.00 indicating a high correlation. The utility measures were considered valid if ≥75% of our *a priori* hypotheses were confirmed (35).

The discriminative ability among baseline HAQ-DI and PGA categories was assessed for both measures. HAQ-DI scores were classified by convenience in three categories: no-to-mild disability (0.00-1.00), moderate disability (1.01-2.00) and severe disability (2.01-3.00). For PGA, a score of 0.0-33.0 was conveniently categorised as mild disease activity, 33.1-66.0 as moderate disease activity and 66.1-1.00 as severe disease activity. Differences between the HAQ-DI and PGA categories were assessed for the EQ-5D and SF-6D using MANOVA. In case of a significant result, post-hoc analyses were conducted between the three categories. A *p*-value <0.05 was considered statistically significant.

The Minimal Important Difference (MID) was defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's (health care) management" (19). To calculate MIDs for the EQ-5D and SF-6D, three different anchors were used: 1) the global rating of change item (GRoC) of the SF-36, 2) the HAQ-DI minimal important difference (0.22 points) for patients with rheumatoid arthritis (36), and 3) the upper bound of a published MID of the HAQ-DI for patients with diffuse SSc (0.14 points) (37). For the GRoC, patients who scored 2 (somewhat better) or 4 (somewhat worse) on the item "Compared to a year ago, how would you rate your health now?" of the SF-

36 were considered the minimally improved and deteriorated subgroups, respectively. Patients with a score of 3 (about the same) were considered the unchanged group. For the HAQ-DI groups, minimally improved and deteriorated subgroups were categorised as patients with a decrease or increase of at least 0.22 (or 0.14 points for the third anchor) on the HAQ-DI, respectively. Patients with a change score <|0.22| (or <| 0.14| for the third anchor) were considered the unchanged group.

To assess responsiveness, a distribution-based approach was used. For the abovementioned anchors, responsiveness was calculated using Effect Size (ES = $D/SD_{baseline}$) (38) and Standardised Response Mean (SRM=$D/SD_{difference}$) (39), in which D is the difference between baseline and follow-up measures, and $SD_{baseline}$ and $SD_{difference}$ are the standard deviations of the baseline measures and difference scores, respectively. Effect sizes were interpreted following Cohen's criteria (38): small (0.2), moderate (0.5) or large (0.8). ES and SRMs of the improved and deteriorated groups were compared with those of the unchanged groups. All statistical analyses were conducted using Stata/IC 10.1 software (StataCorp LP, College Station, TX).

Results

Sample characteristics

Demographics and disease characteristics were displayed in Table I. The majority of patients were female and middle-aged. Most patients were living together or married. One-third of the patients were employed at time of the study, and a large minority received higher education. Time since the onset of the first non-Raynaud's symptom ranged from 2 months to 52 years.

Distribution and agreement

At baseline, the mean EQ-5D score was 0.64 (SD = 0.25, range = -0.18–1.00). The mean SF-6D score was 0.65 (SD = 0.11, range = 0.38–1.00). Mean follow-up scores for the EQ-5D and SF-6D were 0.59 (SD = 0.31) and 0.64 (SD = 0.14), respectively. Neither the EQ-5D nor the SF-6D showed a floor effect (both 0.0%) or a ceiling ef-

Table I. Baseline patient demographic and disease characteristics.

Variables	Values (n=211)
Female (%)	143 (67.8)
Mean age, years (SD)	56.4 (12.0)
Median time since onset first Non-Raynaud symptom (P25-P75)	7.4 (3.5-12.3)[a]
Higher education (%>12 years)	86 (41.6)[b]
Currently working (%)	69 (32.7)
Married or living as married (%)	159 (75.4)
Limited disease (%)	154 (74.4)
Mean modified Rodnan Skin Score (SD; range)	6.3 (6.0; 0-37)
Mean HAQ-DI score (SD; range)	1.04 (0.73; 0.0-3.0)[c]
Mean EQ-5D score (SD; range)	0.64 (0.25; -0.18-1.00)[d]
Mean SF-6D score (SD; range)	0.65 (0.11; 0.38-1.00)[e]

Due to missing values: [a]n=204, [b]n=207, [c]n=210, [d]n=207, [e]n=199

Table II. Correlations [95% CI] of SF-6D an EQ-5D with physical functioning measures and disease-related visual analogue scales at baseline.

	SF-6D	Hypothesis confirmed	EQ-5D	Hypothesis confirmed
HAQ-DI	-0.63 [-0.71, -0.54]	Yes	-0.63 [-0.71, -0.54]	Yes
PGA	-0.58 [-0.66, -0.48]	Yes	-0.62 [-0.70, -0.53]	Yes
Pain VAS	-0.57 [-0.65, -0.46]	No	-0.55 [-0.64, -0.45]	No
Ulcer VAS	-0.36 [-0.48, -0.23]	Yes	-0.30 [-0.42, -0.17]	Yes
Raynaud's VAS	-0.32 [-0.44, -0.19]	Yes	-0.28 [-0.40, -0.15]	Yes
Lung VAS	-0.45 [-0.56, -0.33]	No	-0.38 [-0.49, -0.26]	Yes
GI VAS	-0.31 [-0.43, -0.18]	Yes	-0.32 [-0.44, -0.19]	Yes
Fatigue VAS	-0.63 [-0.71, -0.54]	Yes	-0.55 [-0.64, -0.44]	Yes
Skin score	-0.15 [-0.29, -0.01]*	Yes	- 0.13 [-0.27, 0.01]**	Yes

HAQ-DI: Health Assessment Questionnaire-Disability Index; PGA: Patient Global Assessment; VAS: Visual Analogue Scale; GI: Gastrointestinal. All $p<0.001$;*$p<0.05$;**$p>0.06$.

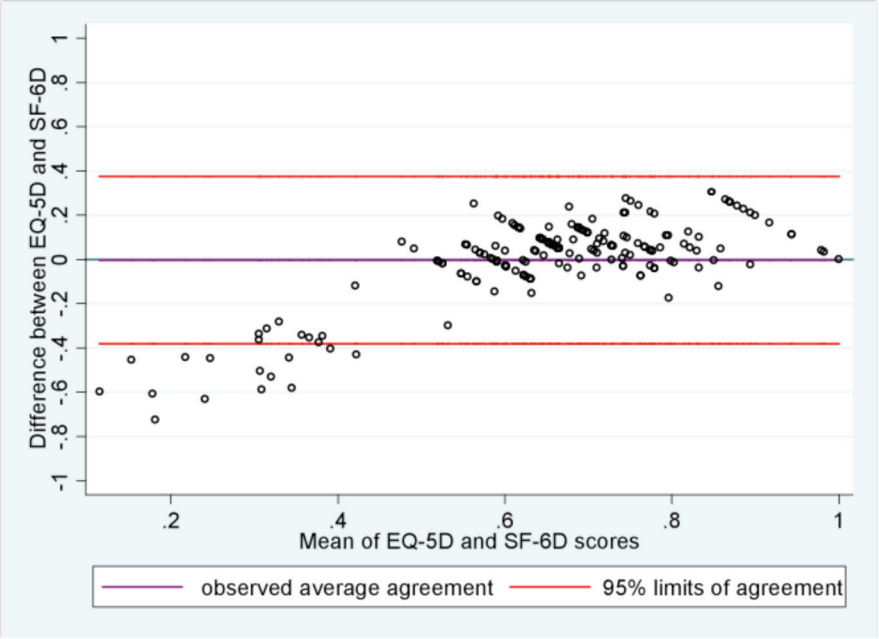

Fig. 1. Bland and Altman plot of differences between EQ-5D and SF-6D for patients with systemic sclerosis.
Lower line: Mean difference +1.96 SD (Value = -0.38); Upper line: Mean difference -1.96 SD (Value = 0.39).

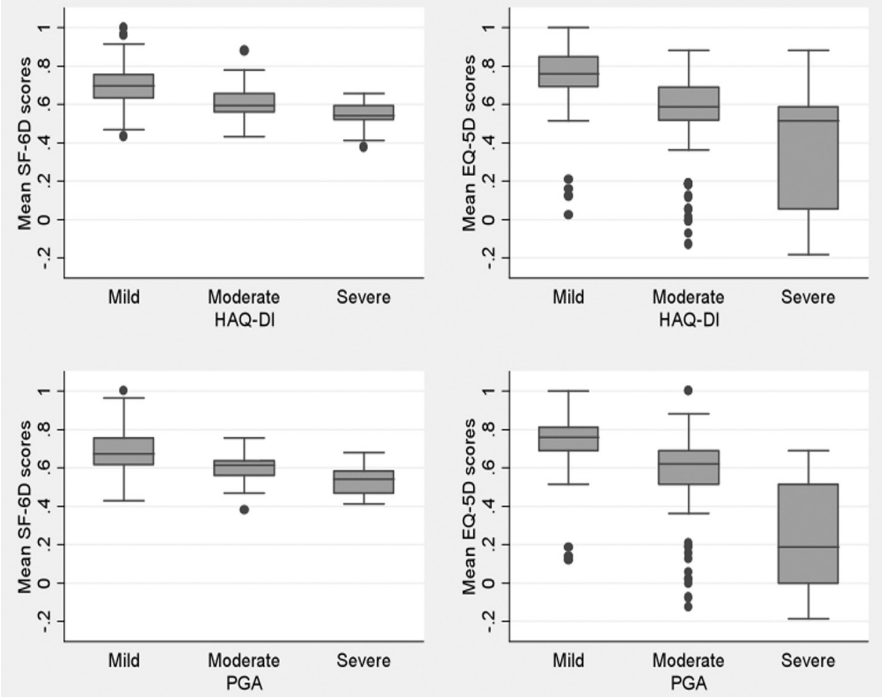

Fig. 2. Discriminative ability of the EQ-5D and SF-6D between mild, moderate and severe HAQ-DI and PGA levels at baseline.
The box-plots present the median, quartiles and extreme values for the EQ-5D and SF-6D utility scores for each HAQ-DI and PGA category.

fect (7.7% and 0.5%, respectively) at baseline. Spearman's correlation between both measures at baseline was r=0.74 (p<0.01), indicating a marked association between EQ-5D and SF-6D utility scores. Lin's concordance coefficient for baseline was 0.49 (95% CI = 0.42–0.56), indicating moderate concordance between the EQ-5D and SF-6D. The Bland-Altman plot (Fig. 1) showed a mean difference of 0.003 (95% CI = 0.024–0.030) but wide limits of agreement (-0.38–0.39). In an additional analysis, only EQ-5D scores <0.45 were taken into account. A structural bias was found for the lower scores on the utility measures. In this group (n=24), the mean difference between the EQ-5D and SF-6D measures was -0.44 (SD=0.14; 95% CI=0.38–0.50).

Validity: construct

Correlations of the EQ-5D and SF-6D with physical functioning measures are displayed in Table II. Consistent with our hypotheses, a marked correlation was found for both utility measures and HAQ-DI and PGA scores, and a small (<0.20) correlation was found with skin score. Our hypotheses regarding a low correlation with the six VAS scores were confirmed for ulcers, Raynaud's phenomenon and gastrointestinal tract. For lung symptoms, a low correlation was found for the EQ-5D, but a moderate correlation with the SF-6D. Furthermore, pain and fatigue both showed a moderate to marked correlation with the EQ-5D and SF-6D. There were no significant differences between the EQ-5D and SF-6D in correlations with HAQ-DI, PGA, the 6 VAS scores and skin score. For the EQ-5D, 77.8% of the *a priori* hypotheses were confirmed, and 66.7% for the SF-6D.

Validity: discrimination

Discrimination between the HAQ-DI and PGA categories for both utility measures at baseline are displayed in Figure 2. According to the MANOVA, both the EQ-5D and SF-6D were able to discriminate between HAQ-DI categories ($F(2,203) = 41.2$, p<0.001 and $F(2,195) = 35.9$, p<0.001, respectively), and between PGA levels ($F(2,200) = 53.3$, p<0.001 and $F(2,192) = 32.9$, p<0.001, respectively). *Post-hoc* analyses revealed that both measures were able to discriminate between all HAQ-DI and PGA categories.

Minimally important differences and responsiveness

MIDs, SRMs and effect sizes for the three utilised anchors are displayed in Table III. For the minimally improved categories, MIDs were 0.05 for the SF-6D and ranged from 0.05 to 0.10 for the EQ-5D. For the minimally deteriorated categories, MIDs for the SF-6D were -0.03 to -0.04, and ranged from -0.12 to -0.14 for the EQ-5D. The MIDs in the minimally changed groups were of larger magnitude than the unchanged groups for both measured and all three anchors. For the SF-6D, the mean MID estimate was 0.05 for the improved and -0.04 for the deteriorated subgroups. The mean MID estimate for the EQ-5D was 0.08 for the improved subgroups, and -0.13 for the deteriorated subgroups. The SRMs and ES for the both utility measures were small to moderate [38]. For the SF-6D, ES and SRM were ranging from -0.29 to -0.41 for the deteriorated categories and from 0.38 to 0.54 for the improved categories. The SRMs and ES for the EQ-5D ranged from -0.50 to -0.55 for the deteriorated categories and from 0.19 to 0.38 for the improved categories.

Discussion

This was the first study to compare measurement properties of the EQ-5D and SF-6D utility measures, to assess the association and agreement between these measures and to estimate MIDs in a sample of patients with SSc. A marked correlation was found between both utility measures, indicating that higher scores on the EQ-5D are associated with higher scores on the SF-6D and vice versa. However, according to the results of this study, the agreement between the EQ-5D and SF-6D was moderate. Especially for lower scores (including about 10% of the sample), results showed that EQ-5D and SF-6D scores, in particular for patients with worse health status, are not reasonably comparable. Further investigation is warranted to assess whether and to what extent the choice for the EQ-5D versus SF-6D to measure utility might

Table III. Minimal important differences and responsiveness (SRM and ES) of the EQ-5D and SF-6D measures.

		SF6D			EQ5D		
Anchor	n[a]	MID [95%CI]	SRM	ES	MID [95%CI]	SRM	ES
GRoC (somewhat worse)	49	-0.03 [-0.06, 0.00]	-0.34	-0.29	-0.12 [-0.18, -0.05]	-0.53	-0.53
GRoC (unchanged)	87	-0.01 [-0.03, 0.01]	-0.14	-0.12	-0.01 [-0.05, 0.04]	-0.04	-0.04
GRoC (somewhat better)	24	0.05 [0.00, 0.09]	0.44	0.38	0.05 [-0.04, 0.14]	0.26	0.19
HAQ-DI (≥0.22 worse)	63	-0.04 [-0.07, -0.02]	-0.41	-0.35	-0.14 [-0.20, -0.07]	-0.55	-0.52
HAQ-DI (unchanged)	95	-0.01 [-0.03, 0.00]	-0.16	-0.11	-0.04 [-0.08, 0.00]	-0.20	-0.15
HAQ-DI (≥0.22 better)	38	0.05 [0.01, 0.09]	0.45	0.54	0.10 [0.01, 0.18]	0.38	0.29
HAQ-DI (≥0.14 worse)	68	-0.04 [-0.06, -0.01]	-0.39	-0.33	-0.13 [-0.20, -0.07]	-0.53	-0.50
HAQ-DI (unchanged)	84	-0.02 [-0.04, 0.00]	-0.20	-0.14	-0.04 [-0.08, 0.00]	-0.20	-0.15
HAQ-DI (≥0.14 better)	42	0.05 [0.02, 0.09]	0.45	0.53	0.08 [0.01, 0.16]	0.35	0.27

GRoC: Global rating or change; HAQ-DI: Health Assessment Questionnaire-Disability Index; MID: Minimal important difference; ES: $D/SD_{baseline}$; SRM: $D/SD_{difference}$. [a]Calculations based on mimimum number of imputations.

have an influence on cost-utility evaluations and treatment decisions.

Neither of the measures showed floor- and ceiling effects, the mean scores were comparable for both measures, and the EQ-5D and SF-6D were both able to discriminate between levels of disability and PGA of health. For the EQ-5D, 77.8% of our *a priori* hypotheses were confirmed, whereas for the SF-6D this percentage was 66.7%. Based on our *a priori* cut-off (≥75% confirmed), our results support the construct validity of the EQ-5D but not for the SF-6D. In contrast to our *a priori* hypotheses, a moderate to marked correlation was found with pain and fatigue for both measures. However, both symptoms are experienced frequently by patients with SSc (27), and were found to have at least moderate impact on daily activities in the majority of patients, indicating that pain and fatigue might significantly reduce health related quality of life in patients with SSc. Therefore, it might not be surprising that these symptoms are associated with lower utility scores in SSc, and our hypotheses regarding pain and fatigue might have been suboptimal.

The mean MIDs for the SF-6D reflect a change of 6.1%–7.7% for SF-6D scores, well within the suggested change of 5–10% that is generally considered a reasonable MID (40). The mean MIDs for the EQ-5D were equivalent of a change of 12.5–20.3%, somewhat larger than the suggested cut-offs (40). The ES and SRM for both measures were consistent with analyses showing MID estimates for health related quality of life instruments ranging from 0.30 to 0.50 SD units (41, 21). In addition, the MIDs for the changed groups were at least twice as large as the estimates in the unchanged groups and could therefore be used to calculate sample sizes for clinical trials involving patients with SSc, and in interpreting the relevance and importance of treatment effects (42).

Compared with the study of Khanna *et al*. (21), the MIDs and ES for the SF-6D found in our sample were larger for the all three anchors. This could possibly be due to differences between samples; the study by Khanna *et al*. included only patients with the diffuse SSc subtype, who generally have a less favourable outcome compared to patients with limited SSc (45), while our sample consisted mainly (approximately 75%) of patients with limited SSc. Furthermore, mean disease duration of the sample by Khanna *et al*. was shorter than that of the sample in the present study. Since the mean SF-6D score in both samples was similar (0.64 versus 0.65), this indicates that the sample of Khanna *et al*. included patients with a worse disease progression, which might be reflected in smaller improvements that are perceived as important by patients.

There are some limitations of the present study that should be taken into account when interpreting our results. First, comparisons between the measures were relative; there was no external standard to compare with, such as physician-rated disease severity. Furthermore, we did not differentiate between baseline levels of health status while assessing MIDs, assuming that minimally important change values are equal for patients with relatively mild and worse disease status. The small numbers of patients per subgroup did not allow us to differentiate among levels of health status. In addition, the MIDs for both utility measures found in this study were relatively small, and compared with recently published standard errors of measurements (SEM) (43), the MIDs found in the present study were of similar magnitude or smaller, indicating that the EQ-5D and SF-6D might not be appropriate for individual patient monitoring. Since SEM also varies across the range of health, notably EQ-5D and SF-6D scores, more research is needed to assess how the MIDs and the SEM are linked in patients with SSc (44). Finally, test-retest reliability for the EQ-5D and SF-6D was not assessed in the present study. In the study of Khanna *et al*. (21), the test-retest reliability of the SF-6D was found to be excellent. In future studies, assessing both utility measures twice in a narrower timeframe contributes to the comparability of the stability over time.

Conclusion

The present study is the first to compare measurement properties and MIDs of two widely used utility measures, notably the EQ-5D and SF-6D, in patients with SSc. Measurement properties were generally acceptable and similar for both measures. Although there was a marked correlation between the measures, the moderate agreement implies that caution is needed when interpreting and comparing results that are

obtained with the EQ-5D and SF-6D. Results from the present study suggest that scores of the EQ-5D and SF-6D are not reasonably comparable, in particular for patients with low utility scores. The MID estimates we provided can be used to calculate sample sizes for clinical trials and guide the interpretation of results, in particular when examining the usefulness of a clinical intervention for SSc patients.

References

1. BRAUER CA, ROSEN AB, GREENBERG D, NEUMANN PJ: Trends in the measurement of health utilities in published cost-utility analyses. *Value Health* 2006; 9: 213-8.
2. BLOOM BS: Use of formal benefit/cost evaluations in health system decision making. *Am J Manag Care* 2004; 10: 329-35.
3. BROOKS R: EuroQol: the current state of play. *Health Policy* 1996; 37: 53-72.
4. RABIN R, DE CHARRO F: EQ-5D: a measure of health status from the EuroQol Group. *Ann Med* 2001; 33: 337-43.
5. BRAZIER J, ROBERTS J, DEVERILL M: The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002; 21: 271-92.
6. WARE JE, JR., SHERBOURNE CD: The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473-83.
7. HARRISON MJ, DAVIES LM, BANSBACK NJ, INGRAM M, ANIS AH, SYMMONS DP: The validity and responsiveness of generic utility measures in rheumatoid arthritis: a review. *J Rheumatol* 2008; 35: 592-602.
8. MARRA CA, RASHIDI AA, GUH D *et al*.: Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Qual Life Res* 2005; 14: 1333-44.
9. SALAFFI F, CAROTTI M, CIAPETTI A, GASPARINI S, GRASSI W: A comparison of utility measurement using EQ-5D and SF-6D preference-based generic instruments in patients with rheumatoid arthritis. *Clin Exp Rheumatol* 2011; 29: 661-71.
10. WHITEHURST DG, BRYAN S, LEWIS M: Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. *Med Decis Making* 2011; 31:E34-44.
11. BRAZIER J, ROBERTS J, TSUCHIYA A, BUSSCHBACH J: A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ* 2004; 13: 873-84.
12. MCDONOUGH CM, TOSTESON AN: Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. *Pharmacoeconomics* 2007; 25: 93-106.
13. KATSUMOTO TR, WHITFIELD ML, CONNOLLY MK: The pathogenesis of systemic sclerosis. *Annu Rev Pathol* 2011; 28: 509-37.
14. LEROY EC, BLACK C, FLEISCHMAJER R *et al*.: Scleroderma (systemic sclerosis): classification, subsets, and pathogenesis. *J Rheumatol* 1988; 15: 202-5.
15. HUDSON M, THOMBS BD, STEELE R, PANOPALIS P, NEWTON E, BARON M, CANADIAN SCLERODERMA RESEARCH GROUP: Health-related quality of life in systemic sclerosis: a systematic review. *Arthritis Rheum* 2009; 61: 1112-20.
16. THOMBS BD, JEWETT LR, ASSASSI S *et al*.: New directions for patient-centred care in scleroderma: the Scleroderma Patient-centred Intervention Network (SPIN). *Clin Exp Rheumatol* 2012; 30: S23-9.
17. TAN A, DENTON CP, MIKHAILIDIS DP, SEIFALIAN AM: Recent advances in the diagnosis and treatment of interstitial lung disease in systemic sclerosis (scleroderma): a review. *Clin Exp Rheumatol* 2011; 29: S66-74.
18. BEYER C, DISTLER O, DISTLER JH: Innovative antifibrotic therapies in systemic sclerosis. *Curr Opin Rheumatol* 2012; 24: 274-80.
19. JAESCHKE R, SINGER J, GUYATT GH: Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials* 1989; 10: 407-15.
20. REVICKI DA, CELLA D, HAYS RD, SLOAN JA, LENDERKING WR, AARONSON NK: Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes* 2006; 27: 4-70.
21. KHANNA D, FURST DE, WONG WK *et al*.: Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. *Qual Life Res* 2007; 16: 1083-92.
22. KWAKKENBOS L, VAN LANKVELD WG, VONK MC, BECKER ES, VAN DEN HOOGEN FH, VAN DEN ENDE CH: Disease-related and psychosocial factors associated with depressive symptoms in patients with systemic sclerosis, including fear of progression and appearance self-esteem. *J Psychosom Res* 2012; 72: 199-204.
23. SUBCOMMITTEE FOR SCLERODERMA CRITERIA OF THE AMERICAN RHEUMATISM ASSOCIATION DIAGNOSTICS A THERAPEUTIC CRITERIA COMMITTEE: Preliminary criteria for the classification of systemic sclerosis (scleroderma). *Arthritis Rheum* 1980; 23: 581-90.
24. CLEMENTS P, LACHENBRUCH P, SIEBOLD J *et al*.: Inter and intraobserver variability of total skin thickness score (modified Rodnan TSS) in systemic sclerosis. *J Rheumatol* 1995; 22: 1281-5.
25. STEEN VD, MEDSGER TA: 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.
26. FRIES JF, SPITZ P, KRAINES RG, HOLMAN HR: Measurement of patient outcomes in arthritis. *Arthritis Rheum* 1980; 23: 137-45.
27. BASSEL M, HUDSON M, TAILLEFER SS, SCHIEIR O, BARON M, THOMBS BD: Frequency and impact of symptoms experienced by patients with systemic sclerosis: results from a Canadian National Survey. *Rheumatology* (Oxford) 2011; 50: 762-7.
28. MOKKINK LB, TERWEE CB, PATRICK DL *et al*.: The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010; 19: 539-49.
29. ANGST F: The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol* 2011; 11: 152.
30. ROYSTON P: Multiple imputation of missing values: update of ice. *Stata J* 2005; 5: 527-36.
31. MARSHALL A, ALTMAN DG, HOLDER RL, ROYSTON P: Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol* 2009; 9: 57.
32. LIN LI: A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; 45: 255-68.
33. BLAND JM, ALTMAN DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 8: 307-10.
34. FRANZBLAU A: A primer of statistics for non-statisticians. New York: Brace & World, 1985.
35. TERWEE CB, BOT SDM, DE BOER MR *et al*.: Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; 60: 34-42.
36. WELLS GA, TUGWELL P, KRAAG GR, BAKER PR, GROH J, REDELMEIER DA: Minimum important difference between patients with rheumatoid arthritis: The patient's perspective. *J Rheumatol* 1993; 20: 557-60.
37. KHANNA D, FURST DE, HAYS *et al*.: Minimally important difference in diffuse systemic sclerosis – results from the D-penicillamine study. *Ann Rheum Dis* 2006; 65: 1325-9.
38. COHEN J: Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Erlbaum; 1988
39. HUSTED JA, COOK RJ, FAREWELL VT, GLADMAN DD: Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* 2000; 53: 459-68.
40. OSOBA D, BEZJAK A, BRUNDAGE M *et al*.: Analysis and interpretation of health-related quality-of-life data from clinical trials: basic approach of The National Cancer Institute of Canada Clinical Trials Group. *Eur J Cancer* 2005; 41: 280-7.
41. GUYATT GH, OSOBA D, WU AW, WYRWICH KW, NORMAN, GR: Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002; 77: 371-83.
42. 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.
43. PALTA M, CHEN HY, KAPLAN RM, FEENY D, CHEREPANOV D, FRYBACK DG: Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Med Decis Making* 2011; 31: 260-9.
44. TERWEE CB, ROORDA LD, KNOL DL, DE BOER MR, DE VET HCW: Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009; 62: 62-7.
45. AL-DHAHER FF, POPE JE, OUIMET JM: Determinants of morbidity and mortality of systemic sclerosis in Canada. *Semin Arthritis Rheum* 2010; 39: 269-77.