EXTENDED REPORT

How do the EQ-5D, SF-6D and the well-being rating scale compare in patients with ankylosing spondylitis?

Annelies Boonen, Désirée van der Heijde, Robert Landewé, Astrid van Tubergen, Herman Mielants, Maxime Dougados, Sjef van der Linden

Ann Rheum Dis 2007;**66**:771–777. doi: 10.1136/ard.2006.060384

See end of article for authors' affiliations

Correspondence to:
Dr A Boonen, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands;
aboo@sint.azm.nl

Accepted 16 December 2006
Published Online First 9 January 2007

Purpose: To compare aspects of validity of EuroQol—5 Dimensions (EQ-5D) and Short-Form—6 Dimensions (SF-6D), two indirect utility instruments, and the well-being rating scale (RS) in ankylosing spondylitis (AS).
Methods: EQ-5D, SF-6D and RS were available for 254 patients fulfilling modified New York criteria. 134 patients were part of an observational cohort and 120 were part of a randomised controlled trial (RCT). Aspects of validity assessed were truth (agreement and correlation with external health measures) and discrimination (differentiation between health states, repeatability and detection of treatment effect).
Results: Median (range) values were 0.69 (−0.08–1.00) for the EQ-5D, 0.65 (0.35–0.95) for the SF-6D and 0.65 (0.14–1.00) for the RS. Agreement (intraclass correlation coefficient) was moderate (0.46–0.55). Instruments correlated equally with disease activity, functioning and quality of life. The SF-6D showed smaller average differences in utility between patients with better and worse disease compared with the EQ-5D and the RS. The smallest detectable difference (SDD) (in the control group of RCT) was 0.36, 0.17 and 0.33 for EQ-5D, SF-6D and RS, respectively. The ability to detect treatment effect (in the intervention trial) showed standardised effect sizes that were moderate for EQ-5D and SF-6D (0.63 and 0.64) and low for the RS (0.23).
Conclusion: In patients with AS, EQ-5D, SF-6D and the RS correlate equally well with external measures of health, but have different psychometric properties. The SDD is most favourable for the SF-6D, but it discriminates less well between patients with different disease severities. The RS has a poorer ability to detect treatment effects. It is difficult to recommend one of the instruments.

Quality-of-life (QoL) instruments measure the overall impact of diseases on individuals and are a relevant outcome in trials and in observational studies.[1] A utility is a specific type of QoL assessment that ranges by definition from 0 to 1, where 0 indicates a health state similar to death and 1 indicates a state of perfect health. Values <0 indicate a health state worse than death. Except for the predefined scale of the utility instrument, there are other distinctions with the usual QoL instruments. Conceptually, utilities reflect the preference for a health state in a choice situation that includes uncertainty. Therefore, utility assessments are choice experiments that include different levels of risk. The level of risk or sacrifice the subject is prepared to take in order to reach a better health status is then transformed in the utility value. A specific application of utilities is that they are the base to calculate quality-adjusted life years (QALYs) by combining the utility value with the time in that particular health state.

Direct and indirect utility elicitation techniques have been developed. Direct utility instruments such as the standard gamble (SG) and time-trade-off (TTO) assess directly the level of risk the subject is willing to take in order to reach an alternative preferable health state.[2] These assessments are usually performed by interview and are time consuming. The indirect utility instruments are multidomain health-status questionnaires completed by patients. These ratings result in a large number of possible health states. The utility of each health state is obtained through a scoring function derived from direct utility assessment of the healthy population. Indirect utilities have the advantage that they can be assessed through self-report questionnaires and are easy to understand. The EuroQol—5 Dimensions (EQ-5D) indirect utility has been proposed in 1990 and is generally accepted.[3] More recently, the Short-Form—6 Dimensions (SF-6D) indirect utility was developed.[4] It has an additional advantage that it can be derived from the Short-Form-36 (SF-36). Other well known indirect utility instruments are the Quality of Well-being[5 6] and the Health Utility Index.[7]

The use of generic indirect utilities in cost–utility analyses is supported by guidelines for pharmacoeconomic evaluations.[8] Although the theoretical concept of utilities implies that one specific health state has one utility score, independent of the way it is measured, different instruments can give different scores.[9–11] This, however, was not confirmed in AS, and previous studies did not explore the underlying comparative test characteristics of the instruments. In addition, although the use of utilities is supported by a clear welfare-based theoretical economic concept, its advantage over a simple general well-being rating scale (RS) is now discussed.[12] As yet, this has not been studied empirically. In this study, we compared some aspects of validity in patients with AS by applying the definitions of Outcome Measures in Rheumatoid Arthritis Clinical Trials.[13] More specifically, this study assesses aspects of "truth" (does the instrument measure what it is supposed to measure) and "discrimination" (repeatability, differentiation between health states and ability to detect a treatment effect as part of the sensitivity to change) of the EQ-5D, SF-6D and the RS instruments.

Abbreviations: ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; EQ-5D, EuroQol—5 Dimensions; ICC, intraclass correlation coefficient; QALY, quality-adjusted life years; QoL, quality of life; RS, rating scale; RTE, relative treatment effect; SDD, smallest detectable difference; SES, standardised effect size; SF-36, Short-Form-36; SF-6D, Short-Form—6 Dimensions; SG, standard gamble; TTO, time-trade-off

Table 1 Differences in both indirect utility indices with regard to the number of questions and domains covered by the original health state measurement, the approach to derive the utility index and its range

Instrument	Number of questions and answer scale	Domains	Number of possible health states*	Method of valuing preference	Range
EQ-5D	5 questions; answering scale from 1–3	Mobility (1Q) Daily activities (1Q) Self-care (1Q) Pain (1Q) Mood (1Q)	243 (42)	TTO†	−0.59 to 1.00
SF-6D	10 questions of the SF-36; different answering scales	Physical function (2Q) Role limitation (4Q) Social function (1Q) Pain (1Q) Mental health (1Q) Vitality (1Q)	18000 (249)	SG†	0.30 to 1.00

EQ-5D, EuroQol—5 Dimensions; SF-6D, Short-Form—6 Dimensions; Q, question; TTO, time-trade-off; SG, standard gamble.
*The number of states used in the development of the utility index are given in brackets.
†EQ-5D utility functions are available for populations from different countries (http://www.euroqol.org) and SF-6D utility functions are derived from the UK populations only.

METHODS

Patients

Two datasets of patients fulfilling modified New York criteria[14] were merged. The first dataset included 143 patients, who were part of an unselected prevalence-based cohort with longitudinal follow-up (Outcome in Ankylosing Spondylitis International Study cohort).[15] The second dataset included 120 patients who took part in a 9-months randomised controlled trial comparing a 3-week spa treatment (n = 80) with usual care (n = 40).[16 17] Both studies were approved by medical ethics committees and all patients gave consent. The patients participating in the intervention trial had to experience pain and functional limitations during the 3 months preceding entry into the study. As the results of all validation tests of the three instruments provided similar results when performed in the two datasets separately, it was decided to merge datasets to enhance the clarity of presentation. Test–retest repeatability and ability to detect a treatment effect could only be assessed in patients participating in, respectively, the control (n = 40) and intervention (n = 80) groups of the clinical trial.

Questionnaires

All patients had at least at one point in time[4] completed both EuroQol[3] and SF-36. Patients in the observational cohort completed both instruments in 1999, which was the third year of the assessment. Patients from the clinical trial completed both instruments at several assessment points during the trial period (1999). For this analysis, the baseline and fourth-week assessments (1 week after the end of the 3-week spa treatment) were considered.

The first five questions of the EuroQol were used to calculate the EQ-5D utility based on the scoring function that was derived from a UK population utility elicitation.[3] The rating scale of the EuroQol was used as the direct well-being score. The rating scale of the EuroQol is by convention a 20 cm thermometer-like vertical line, with the lowest anchor labelled as "worst imaginable health state possible" and the upper anchor labelled as the "best imaginable health state possible". To be used as a utility scale (range 0–1), the ratings on the original scale (range 0–100) were divided by 100. From the SF-36, the SF-6D utility was calculated by applying the scoring function that was also derived from a UK population utility elicitation.[4] Table 1 and appendix describe the details of the EQ-5D and the SF-6D with regard to the health domains, the number of questions included for each domain in the health status questionnaire, the answering scales of these questions, the number of total possible health states and the choice experiment applied (SG or TTO) to obtain the population-derived utility score.

In addition to the indirect utility and well-being RS, patients completed the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) to assess disease activity (BASDAI range 0–10,

Table 2 Characteristics of patients included in each of the datasets

	Combined datasets (n = 254)	Observational cohort (n = 134)	Spa exercise trial (n = 120)	p Value
Male (%)	70	69	73	0.65
Mean (SD) age (years)	48.3 (11.2)	49.0 (12.2)	48.0 (9.0)	
Mean (SD) disease duration*	13.1 (8.1)	14.9 (9.3)	10.9 (5.7)	<0.001
Employed (%)	50 (55 for <65 years)	47 (55 for <65 years)	55 (56 for <65 years)	0.15 (0.29 for <65 years)
Mean (SD) BASDAI	4.2 (2.2)	3.7 (2.2)	4.7 (2.0)	<0.001
Mean (SD) BASFI	4.2 (2.23)	3.9 (2.4)	4.5 (2.0)	0.13
Mean (SD) ASQoL	6.9 (4.5)	6.2 (4.4)	7.8 (4.5)	0.01
Mean (SD; median) EQ-5D	0.64 (0.23; 0.69) Range −0.08 to 1.00	0.62 (0.25; 0.69) Range −0.08 to 1.00	0.66 (0.21; 0.69) Range −0.02 to 1.00	0.43
Mean (SD; median) SF-6D	0.67 (0.12; 0.65) Range 0.35 to 0.95	0.69 (0.13; 0.68) Range 0.51 to 0.95	0.64 (0.11; 0.62) Range 0.35 to 0.89	0.005
Mean (SD; median) RS	0.62 (0.18; 0.65) Range 0.14 to 1.00	0.62 (0.18; 0.65) Range 0.14 to 1.00	0.62 (0.19; 0.65) Range 0.19 to 1.00	0.86

ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; EQ-5D, EuroQol 5 dimensions; SF-6D, Short-Form—6 Dimensions; RS, rating scale.
*Since diagnosis.

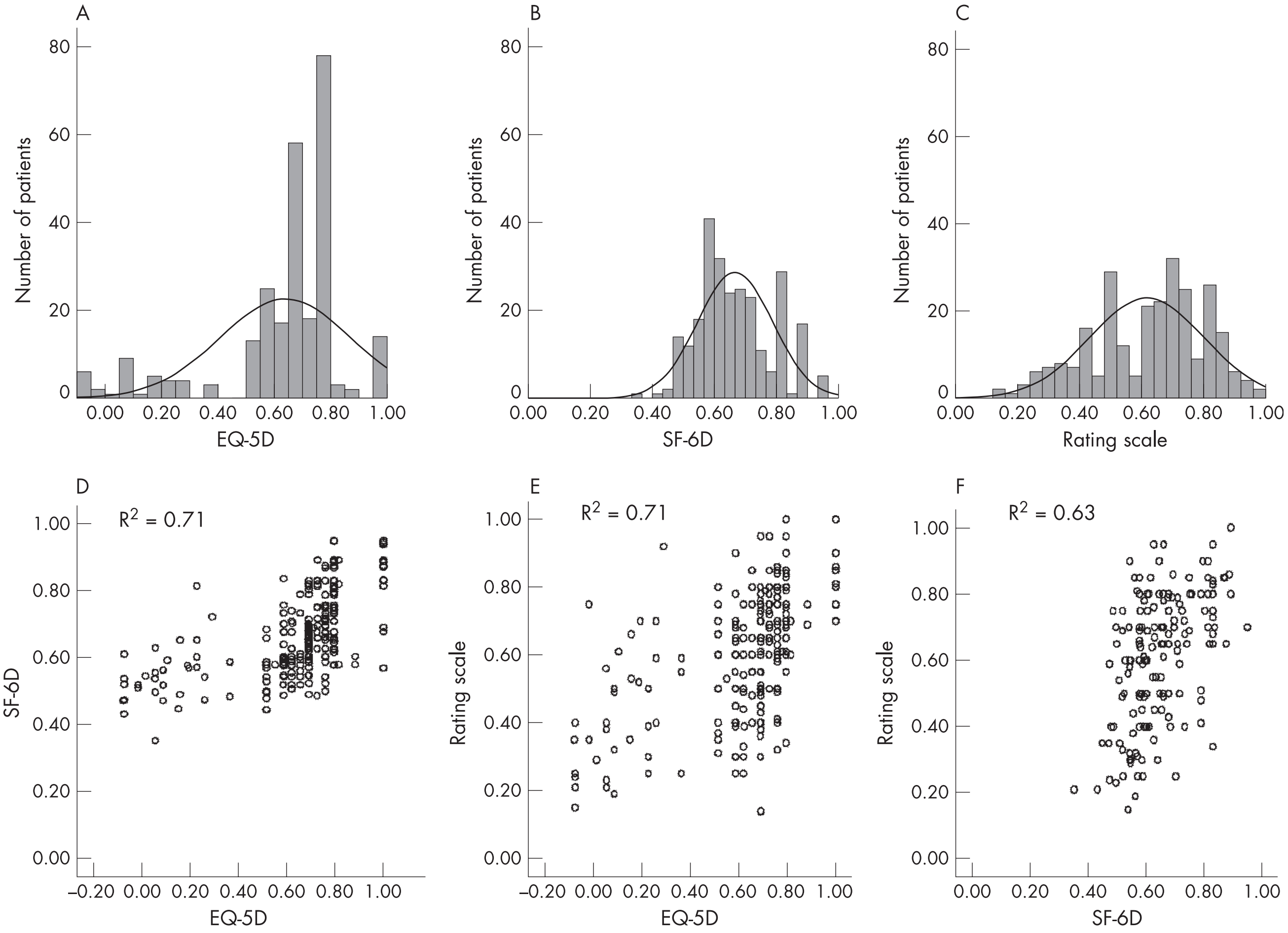

Figure 1 Distribution of EuroQol—5 Dimensions (EQ-5D; A), Short-Form—6 Dimensions (SF-6D; B) and rating scale (C) and scatter plots of EQ-5D with SF-6D (D), EQ-5D with rating scale (E) and SF-6D with rating scale (F).

higher values indicating more active disease),[18] the Bath Ankylosing Spondylitis Functional Index (BASFI) to assess physical functioning (BASFI range 0–10, higher values indicating worse functioning)[19] and the Ankylosing Spondylitis Quality of Life (ASQoL; range 0–18) instrument to assess disease-specific QoL.[20]

Statistical analyses

Descriptive statistics characterised the patient samples, and unpaired t test or χ^2 test tested differences in continuous or categorical variables, respectively.

To assess aspects of truth, the absolute agreement between instruments (EQ-5D, SF-6D and RS) was calculated by single-

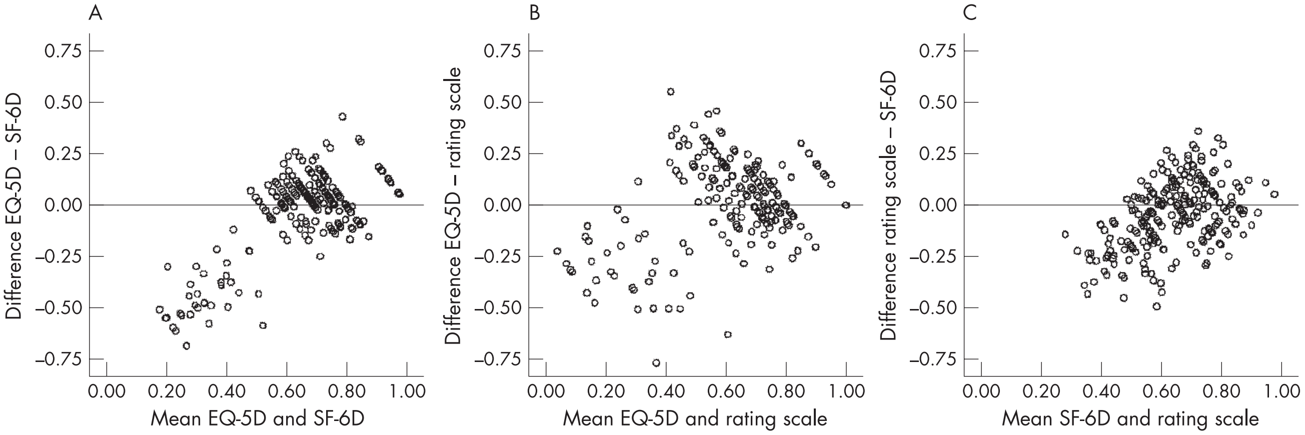

Figure 2 Bland and Altman plots comparing EuroQol—5 Dimensions (EQ-5D) with Short-Form—6 Dimensions (SF-6D) (A), EQ-5D with rating scale (B) and rating scale with SF-6D (C).

Table 3 Correlation of indirect utility instruments and RS with external standards of health state

	BASDAI	BASFI	ASQoL
EQ-5D	−0.55	−0.59	−0.71
SF-6D	−0.59	−0.53	−0.71
RS	−0.59	−0.58	−0.63

ASQoL, Ankylosing Spondylitis Quality of Life; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; EQ-5D, EuroQol—5 Dimensions; SF-6D, Short-Form—6 Dimensions; RS, rating scale.

measure intraclass correlation coefficient (ICC). Scatter plots as well as Bland and Altman plots[21] for each pair of assessments were provided to visualise the level of agreement in relation to the measurement scale. Subsequently, Spearman's correlation of the instruments with external measures of health comprising BASDAI, BASFI and disease-related quality of life (ASQoL) was determined, and scatter plots were provided.

To assess aspects of "discrimination", first, the discriminatory capacity between patients with contrasting health states was assessed by comparing the difference in scores between patients with high opposed to low disease activity, defined by BASDAI (cut-off for higher disease activity at ⩾4) and between patients with better as opposed to worse physical functioning (cut-off for worse functioning at ⩾4). Test–retest repeatability was assessed in the control group of patients of the clinical trial (n = 40) using the single-measure ICC for absolute agreement for the scores at baseline and 4 weeks later, assuming that the health state (and preference for health state) would not change over such a short period of time without intervention. The smallest detectable difference (SDD) was calculated by the limits of agreement method. For each instrument, Bland and Altman[21] plots illustrate the magnitude of the difference between the two assessment points and show how the difference values are distributed over the entire range of the score in function of the average of the scores of the two assessment points. Sensitivity to change aimed to compare the ability of the instruments to detect a treatment effect in the participants of the clinical trial (n = 120) using the standardised effect size (SES). The SES is the difference in the mean change 1 week after the spa treatment, in the intervention and control groups (d_i–d_c) divided by the SD of the pooled change.[22] Because comparison of the SES might be hampered by different scales of the three instruments, the relative magnitude of the treatment effects across the measures was also expressed as relative treatment effect (RTE). This is the difference in change score 1 week after the spa treatment between the intervention and control groups divided by the change score in the control group ($((d_i–d_c)/d_c)$).[22]

RESULTS

Characteristics of the patient samples

Table 2 shows the demographic and clinical characteristics of patients of the individual and the merged databases. Patients in the spa exercise trial had shorter disease duration (10.9 (SD 5.7) vs 14.9 (SD 9.3) years; p<0/001) and somewhat higher disease activity (BASDAI 4.7 (SD 2.2) vs 3.7 (SD 2.4); p = 0.14).

Utilities and RS: truth: agreement and construct validity

Table 1 and fig 1A–C show that the EQ-5D covers a larger range of the utility scale on the left side of the scale and has a more skewed distribution. The lowest EQ-5D value was −0.08 whereas these were 0.35 for the SF-6D and 0.14 for the RS. EQ-5D values cluster between 0.6 and 0.8 (fig 1A). Agreement (ICC) between instruments was only moderate; 0.47 (95% CI 0.40 to 0.56) for the pair EQ-5D and SF-6D, 0.55 (95% CI 0.46 to 0.62) for the pair EQ-5D and RS, and 0.46 (95% CI 0.34 to 0.56) for the pair SF-6D and RS. Patients with EQ-5D utilities between 0.6 and 0.8 show a wide range of SF-6D utility and RS values, showing a ceiling of the EQ-5D (fig 1D,E). As illustrated by the Bland and Altman plots in fig 2, especially for worse health states, the differences between the EQ-5D and either SF-6D (fig 2A) or RS (fig 2B) are important, the EQ-5D giving systematically lower values. Similarly, in the poorer health states, the RS provides systematically lower utility values than the SF-6D (fig 2C).

Correlation with external standards health status was similar and was moderate to good for all external standards, with the best correlations with ASQoL (table 3).

Figure 3 illustrates the EQ-5D differentiates less in the better health states: patients with EQ-5D utilities between 0.6 and 0.8 show a large variability in the spectrum of the ASQoL values.

Discriminant validity: discrimination between disease states, test–retest repeatability and treatment effect

Table 4 shows that the SF-6D discriminates less between groups of patients with lower opposed to higher disease activity (threshold BASDAI ⩾4) or better opposed to worse physical functioning (threshold BASFI ⩾4). The differences in scores between the groups were larger (0.19–0.21 points) for the EQ-5D and RS than for the SF-6D (0.10–0.12 points).

Repeatability in the patients included in the control group of the clinical trial was moderate and lower for the EQ-5D (ICC 0.55 (95% CI 0.29 to 0.73)) than the SF-6D (ICC 0.68 (95% CI

Table 4 Differentiating capacity of the instruments for low and high disease activity (Bath Ankylosing Spondylitis Disease Activity Index) better and worse physical functioning (Bath Ankylosing Spondylitis Functional Index)

	BASDAI <4 (n = 125) Mean BASDAI 2.28 (SD 1.05)	BASDAI ⩾4 (n = 137) Mean BASDAI 5.95 (SD 1.28)	Δ Groups BASDAI <4 and ⩾4 (mean difference 2.67)
EQ-5D	0.73 (0.16)	0.55 (0.26)	0.18 (95% CI 0.13 to 0.24)
SF-6D	0.73 (0.12)	0.61 (0.09)	0.12 (95% CI 0.09 to 0.14)
RS	0.71 (0.14)	0.53 (0.17)	0.18 (95% CI 0.14 to 0.21)
	BASFI <4 (n = 121) Mean BASFI 2.15 (SD 1.16)	**BASFI ⩾4 (n = 143) Mean BASFI 5.88 (SD 1.29)**	**Δ Groups BASFI <4 and ⩾4 (mean difference 3.73)**
EQ-5D	0.74 (0.16)	0.55 (0.25)	0.19 (95% CI 0.14 to 0.24)
SF-6D	0.72 (0.12)	0.62 (0.10)	0.10 (95% CI 0.08 to 0.13)
RS	0.72 (0.14)	0.53 (0.16)	0.21 (95% CI 0.16 to 0.23)

BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; EQ-5D, EuroQol—5 Dimensions; SF-6D, Short-Form—6 Dimensions; RS, rating scale; Δ, difference.
Values are mean (SD).

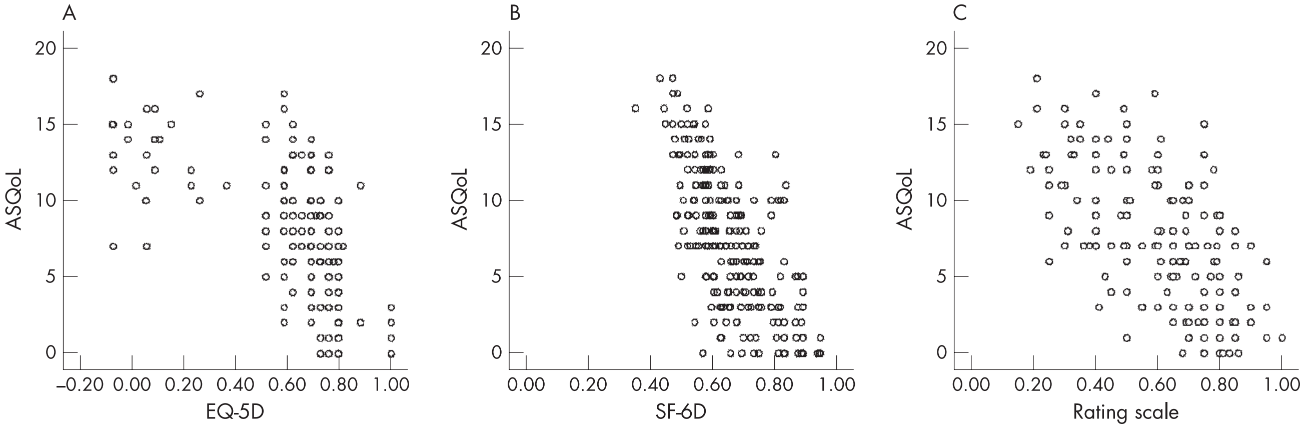

Figure 3 Scatter plots illustrating the relationship between Ankylosing Spondylitis Quality of Life (ASQoL) and EuroQol—5 Dimensions (EQ-5Q; A), Short-Form—6 Dimensions (SF-6D; B) and rating scale (C).

0.46 to 0.81)) and the RS (ICC 0.66 (95% CI 0.45 to 0.81)). The SDD was 0.36 for the EQ-5D, 0.17 for the SF-6D and 0.33 for the RS. Figure 4 illustrates the poor repeatability of the EQ-5D, especially in the lower utility values.

Patients receiving the active spa treatment improved in BASDAI (−1.09 (SD 1.83)), BASFI (−1.00 (SD 1.31)) and ASQoL (−2.03 (SD 3.19)) at week 4 (3 weeks after intervention) compared with baseline. In this group, mean (SD) improvement in EQ-5D, SF-6D and RS were 0.074 (0.23), 0.075 (0.11) and 0.065 (0.18), respectively. The changes in scores were normally distributed. Table 5 shows the mean difference in change in intervention and control groups, the pooled SD of change, the SES and the RTE. Ability to detect change after intervention compared with the control group was better for EQ-5D and SF-6D than for the RS. Also after adjusting for difference in scaling by using the RTE, the ability to detect treatment difference was similar for EQ-5D and SF-6D.

DISCUSSION

Several aspects of validity of the EQ-5D, SF-6D and well-being RS in patients with AS were investigated and are summarised in table 6. The instruments correlate equally well with external measures of health, but agree only moderately among each other. The test–retest repeatability (SDD) was best for the SF-6D and its ability to detect a treatment effect is moderate, but it discriminates less between patients with better and worse health states compared with both other instruments. These findings make it difficult to recommend one of the instruments.

The three instruments are developed to measure generic QoL and can be used to calculate QALYs. They have in common that they rate health on a scale between 0 (representing death for the utilities or the worst possible health for the RS) and 1 (representing perfect health). As the EQ-5D and SF-6D are also utilities, they theoretically measure a different underlying construct than a simple well-being RS. It is mainly argued that the inclusion of a choice experiment in the valuation of a health state better reflects the true welfarist value associated with health. However, few empirical data support the view that the indirect utilities measure a different concept, and a discussion on this issue is underway in the health economic literature.[12] Our data support the view that the relationship with external disease-specific instruments of health state including disease activity, functioning with disease-specific quality of life (ASQoL) are similar for the indirect utilities and the RS. The construct validity of the indirect utilities in relation to the RS should be further explored in AS and other diseases.

It should be emphasised that we included only indirect utility instruments, which are not simply interchangeable with utilities derived for direct assessments.[23 24] A study that compared the RS with the SG in patients with AS and showed better utility with the SG method (0.86) than the RS (0.69).[25]

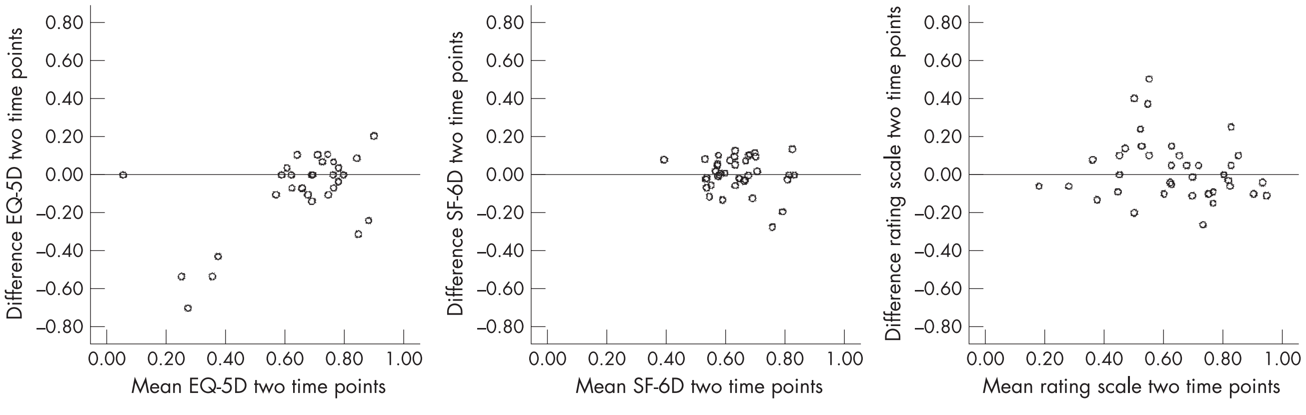

Figure 4 Bland and Altman plots illustrating differences in the repeatability (test–retest) of the EuroQol—5 Dimensions (EQ-5D; A), Short-Form—6 Dimensions (SF-6D; B) and rating scale (C).

Table 5 Comparison of sensitivity to change, specified as the ability of the instruments to detect a treatment effect

	Observed treatment effect	Pooled SD	Standardised treatment effect	Relative treatment effect
EQ-5D	0.14	0.22	0.63	−2.09
SF-6D	0.07	0.11	0.64	11.45
RS	0.04	0.18	0.23	1.67

EQ-5D, EuroQol—5 Dimensions; SF-6D, Short-Form—6 Dimensions; RS, rating scale (0–1).
Analyses were performed in the subgroup that participated in the spa intervention trial (n = 120).
Observed treatment effect: difference in the mean change in the intervention and control groups (d_i–d_c); Standardised effect size: ratio of the treatment effect and the pooled standard deviation (PSD) (d_i–d_c/PSD); Relative treatment effect: relative magnitude of the treatment effect computed as (d_i–d_c/d_c).

Similar findings were reported in RA.[26] Likely, patients adapt to the consequence of the disease and become risk adverse. However, direct utility instruments also differ among themselves.[27] Moreover, they are time consuming and less feasible in (large-scale) intervention studies and longitudinal cohorts. Notwithstanding, utility ratings from both types of studies are necessary for health-economic evaluations.

A simple, but key observation was the difference in the "true" range of the theoretical 0–1 utility scale the instruments actually cover, especially at the lower part of the scale. The lowest observed value was −0.08 for the EQ-5D, 0.35 for the SF-6D and 0.14 for the RS 0.14. It is therefore not surprising that differences between instruments were especially high in the subgroup with worse disease. As a direct consequence, the mean EQ-5D showed larger differences between groups with better and worse disease defined by either BASDAI or BASFI. This will have important consequences when using the instruments in clinical trials and cost-effectiveness analyses that select patients with high disease activity; the mean change (gain) in EQ-5D will be larger and provide more favourable incremental cost–utility values. In a 5-years Markov model on cost–utility of an expensive treatment (drug acquisition about €12 000 per year) in patients with active AS (BASDAI ⩾4) the cost–utility ratio would be €88 000/QALY using the EQ-5D to calculate QALY opposed to €174 000/QALY using the SF-6D. In this model, baseline utility in patients with active AS was 0.49 for EQ-5D but 0.62 for SF-6D. It is clear that the potential gain in EQ-5D utilities is larger when improving disease activity, than for SF-6D utilities. However, the validity of the greater "change" should be critically questioned in view of the observed repeatability. The test–retest repeatability based on the limits of agreement theory was poorest for the EQ-5D and RS, with SDDs of 0.36 and 0.33, respectively, compared with 0.17 for the SF-6D. Comparison should be interpreted in relation to the instrument ranges. When accepting the theoretical range of the scales from 0 to 1, the SDDs would be 36%, 17% and 33% of this (theoretical) scale for the EQ-5D, SF-6D and RS, respectively. If SDD is expressed as a proportion of the true instrument scale, the SDDs are 22%, 24% and 33% for the EQ-5D, SF-6D and RS, respectively, being high for all instruments. A recent paper showed the minimally important difference to be, on average, 0.074 for the EQ-5D and 0.041 for the SF-6D[28] across 11 chronic conditions. In rheumatoid arthritis, the minimal important difference was estimated to be 0.05 for the EQ-5D and 0.03 for the SF-6D.[29] These values need to be questioned when the statistically detectable differences are much higher. This study is the first to compare the ability to detect a treatment effect. The EQ-5D and SF-6D were more sensitive to change compared with the RS. It should be noted that the mean (SD) BASDAI in the spa intervention group was 4.8 (2.0), whereas the mean BASDAI in most biological studies is >6. Taking into account the differences between EQ-5D and SF-6D to discriminate between better and worse disease, the treatment effect should be checked in a population selected with high disease activity.

The comparison of the metric properties of the instruments was methodologically hampered as the EQ-5D showed a high level of skewness compared with the normal distribution of the SF-6D and RS. Classic approaches to study agreement and repeatability assume normality. It should be noted that the change values of the utilities had a near-normal distribution.

In the literature, differences between the EQ-5D and the SF-6D were already observed in other diseases, but not in AS.[9–11] Two studies compared EQ-5D, SF-6D and Health Utility Index in longitudinal cohorts of patients with either RA[10] or a mixture of rheumatological conditions (but not AS).[9] Both reported interchangeable constructs of the indirect utility instruments. Similarly, they confirmed the larger mean change in the EQ-5D in patients who reported change in health state. Test–retest repeatability and sensitivity to change were measured by effect size after dividing patients in groups according to their answer on a transition question (health state better, worse or same)

Table 6 Overview of aspects of validity of the three instruments

	Number of items in health status questionnaire	Calculation	Construct	Distribution	Correlation with external measures of health	Discrimination between groups with better and worse disease	Repeatability	Detecting treatment effect (sensitivity to change)
EQ-5D	5 items	Scoring function freely available	Indirect utility	Skewed left-sided and values cluster between 0.6 and 0.8	Moderate to good	Better than SF-6D and equal to RS	ICC 0.55, SDD 0.36	SES moderate
SF-6D	10 items of the SF-36	Scoring function freely available*	Indirect utility	Near normal	Moderate to good	Lower than EQ-5D and RS	ICC 0.68, SDD 0.17	SES moderate
RS	1 item	Direct value	Direct assessment of well-being	Near normal	Moderate to good	Better than SF-6D and equal to EQ-5D	ICC 0.66, SDD 0.33	SES low

EQ-5D, EuroQol—5 Dimensions; RS, rating scale; SDD, smallest detectable difference; SES, standardised effect size; SF-6D, Short-Form—6 Dimensions; SR-36, Short-Form-36.
*If for academic use.

and analysing the differences between these groups.[9] [10] One of the studies reported no important differences in the repeatability and sensitivity to change, but the second study reported our finding of better sensitivity to change of the SF-6D. It should be realised that the method to estimate repeatability and responsiveness in these studies actually reflects the meaningful difference. Our analyses examined the statistically detectable difference and studied the ability to detect a treatment effect.

SUMMARY AND CONSIDERATIONS

This study first questions whether indirect utility instruments measure a different construct in empirical studies compared with an RS. Second, the results question the psychometric comparability of EQ-5D, SF-6D and the RS. It is difficult to advise which instrument should be preferred. The results of this study call for recommendations on which generic QoL instruments can be used in studies that aim to calculate QALYs.

ACKNOWLEDGEMENTS

We thank Kurt de Vlam, Maryse Meziere and Liesbeth Heuft for data collection in the patients of the OASIS cohort.

....

Authors' affiliations

Annelies Boonen, Désirée van der Heijde, Robert Landew, Astrid van Tubergen, Sjef van der Linden, Department of Internal Medicine, Division of Rheumatology, Caphri Research Institute, University Hospital Maastricht, Maastricht, The Netherlands

Herman Mielants, Department of Rheumatology, University Hospital Gent, Gent Belgium

Maxime Dougados, Department of Rheumatology, Université Réné Descartes, Hôpital Cochin, Paris, France

Funding: There was no sponsorship for this study.

Competing interests: None declared.

REFERENCES

1 **Wolfe F**. Critical issues in longitudinal and observational studies: purpose, short versus long term, selection of study instruments, methods, outcomes, and biases. *J Rheumatol* 1999;**26**:469–72.

2 **Drummond MF**, O'Brien BJ. *Methods for the economic evaluation of health care programmes.* Oxford: Oxford Medical Publications, 1997.

3 **Group TE**. EuroQol— a new facility for the measurement of health-related quality of life. *Health Policy* 1990;**16**:199–208.

4 **Brazier J**, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *J Clin Epidemiol* 1998;**51**:1115–28.

5 **Kaplan RM**, Atkins CJ, Timms R. Validity of a quality of well-being scale as an outcome measure in chronic obstructive pulmonary disease. *J Chronic Dis* 1984;**37**:85–95.

6 **Balaban DJ**, Sagi PC, Goldfarb NI, Nettler S. Weights for scoring the quality of well-being instrument among rheumatoid arthritics. A comparison to general population weights. *Med Care* 1986;**24**:973–80.

7 **Feeney D**, Torrance GW, Gabriel SE. *Health Utilities Index.* Philadelphia, PA: Lippincott-Raven, 1996.

8 **Gold MR**, Russell LB, Siegel JE, Weinstein MC. *Cost-effectiveness in health and medicine.* Oxford, UK: Oxford University Press, 1996.

9 **Conner-Spady B**, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Med Care* 2003;**41**:791–801.

10 **Marra CA**, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, *et al.* Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Qual Life Res* 2005;**14**:1333–44.

11 **Longworth L**, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Econ* 2003;**12**:1061–7.

12 **Parkin D**, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Econ* 2006;**15**:653–64.

13 **Boers M**, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology. *J Rheumatol* 1998;**25**:198–9.

14 **van der Linden S**, Valkenburg H, Catrs A. Evaluation of diagnosis of AS. The modified New York criteria. *Arthritis Rheum* 1984;**27**:361–8.

15 **Boonen A**, van der Heijde D, Landewe R, Guillemin F, Spoorenberg A, Schouten H, *et al.* Costs of ankylosing spondylitis in three European countries: the patient's perspective. *Ann Rheum Dis* 2003;**62**:741–7.

16 **van Tubergen A**, Landewe R, van der Heijde D, Hidding A, Wolter N, Asscher M, *et al.* Combined spa-exercise therapy is effective in patients with ankylosing spondylitis: a randomized controlled trial. *Arthritis Rheum* 2001;**45**:430–8.

17 **Van Tubergen A**, Boonen A, Landewe R, Rutten-Van Molken M, Van Der Heijde D, Hidding A, *et al.* Cost effectiveness of combined spa-exercise therapy in ankylosing spondylitis: a randomized controlled trial. *Arthritis Rheum* 2002;**47**:459–67.

18 **Calin A**, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougados M. Defining disease activity in ankylosing spondylitis: is a combination of variables (Bath Ankylosing Spondylitis Disease Activity Index) an appropriate instrument? *Rheumatology (Oxford)* 1999;**38**:878–82.

19 **Calin A**, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, *et al.* A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. *J Rheumatol* 1994;**21**:2281–5.

20 **Doward LC**, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, *et al.* Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Ann Rheum Dis* 2003;**62**:20–6.

21 **Bland JM**, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;**8**:135–60.

22 **Buchbinder R**, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. *Arthritis Rheum* 1995;**38**:1568–80.

23 **Suarez-Almazor ME**, Conner-Spady B. Rating of arthritis health states by patients, physicians, and the general public. Implications for cost-utility analyses. *J Rheumatol* 2001;**28**:648–56.

24 **Ariza-Ariza R**, Hernandez-Cruz B, Carmona L, Dolores Ruiz-Montesinos M, Ballina J, Navarro-Sarabia F. Assessing utility values in rheumatoid arthritis: a comparison between time trade-off and the EuroQol. *Arthritis Rheum* 2006;**55**:751–6.

25 **Bakker C**, Rutten M, van Doorslaer E, Bennett K, van der Linden S. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;**21**:269–74.

26 **Tijhuis GJ**, Jansen SJ, Stiggelbout AM, Zwinderman AH, Hazes JM, Vlieland TP. Value of the time trade off method for measuring utilities in patients with rheumatoid arthritis. *Ann Rheum Dis* 2000;**59**:892–7.

27 **Suarez-Almazor ME**, Conner-Spady B, Kendall CJ, Russell AS, Skeith K. Lack of congruence in the ratings of patients' health status by patients and their physicians. *Med Decis Making* 2001;**21**:113–21.

28 **Walters SJ**, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res* 2005;**14**:1523–32.

29 **Marra CA**, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, *et al.* A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Soc Sci Med* 2005;**60**:1571–82.

APPENDIX

DESCRIPTION OF THE UTILITY INSTRUMENTS

The EuroQol instrument consists of two parts.[3] The first part comprises five questions, each addressing a different attribute (domain) of health status (EuroQol—5 Dimensions) covering mobility, self-care, daily activities, pain and mood. Each of the five questions can be answered on a three-point categorical scale. This results in a total of 243 possible health states. Utility values were derived by elicitating time-trade-off (TTO) values for a selected set of 42 health states by interviewing a representative sample of a non-institutionalised general population ($n = 2997$). In the TTO technique, the subject has to indicate the number of expected life years he/she is willing to give up to reach that ''preferred'' health state compared with the health state in question. The final utility transformation results in an equation that provides utility values ranging from -0.59 to 1.00. In addition to the UK population utilities, equations have been derived by a similar approach from several other populations including the US, Canada, Japan, New Zealand, Zimbabwe and most European countries (http://www.euroqol.org).

The SF-6D is based on 10 questions covering 6 health attributes (domains) of the original SF-36, which includes 36 questions covering 8 health attributes (domains).[4] Of the 10 selected questions, 2 questions relate to physical function, 4 to role limitation and 1 each to social function, pain, mental health and vitality. The answering scales vary for each of the questions. Of the total of 18 000 possible health states, 249 were selected to elicitate utilities during interviews applying the standard gamble (SG) in a sample of 611 subjects of the general UK population. In the SG, the subject has to decide whether he/she is willing to stay in the health state at question or is willing to take a gamble to reach either ''perfect health for a specific (variable) time period'' or ''immediate death''. The final utility equation results on a utility value ranging from 0.30 to 1.00.