Comparison of the EQ-5D and the SF-6D Utility Measures in 813 Patients with Early Arthritis: Results from the ESPOIR Cohort

CÉCILE GAUJOUX-VIALA, ANNE-CHRISTINE RAT, FRANCIS GUILLEMIN, RENÉ-MARC FLIPO, PATRICE FARDELLONE, PIERRE BOURGEOIS, and BRUNO FAUTREL

ABSTRACT. ***Objective.*** The revolution of early aggressive therapy in early arthritis (EA) has fueled the search for better approaches to establish cost-effectiveness. Our objective was to compare the EuroQol EQ-5D health outcome measure and the SF-6D and to investigate their relationship to clinical variables in a large prospective cohort of patients with EA.

Methods. The EQ-5D and SF-6D utility measures were longitudinally assessed in 813 patients with EA. Agreement and aspects of validity (construct validity, discrimination) were assessed.

Results. At baseline, mean values for EQ-5D were 0.52 ± 0.31 (range –0.59 to 1.0) and for SF-6D were 0.58 ± 0.11 (range 0.30 to 0.92), with a bimodal distribution for the EQ-5D. Agreement was low for patients with severe disability or active disease: the utility was systematically lower with EQ-5D. The intraclass correlation coefficient was 0.42 at baseline and increased to 0.53 at 6 months and 0.57 at 1 and 2 years. Correlations between the 2 utility scores and the Health Assessment Questionnaire were good, and remained similar and stable over 2 years (r = –0.70). Correlations with the Disease Activity Score for 28 joints and the physical component of the MOS 36-item Short-form Health Survey (SF-36) were moderate to good and stable. In contrast, correlation with the mental component of the SF-36 was better with the SF-6D, and the correlation with pain, weak at baseline, improved at 6 months and remained stable thereafter. The SF-6D was better able to discriminate patients with high disease activity.

Conclusion. There was systematic disagreement between EQ-5D and SF-6D in EA, especially in patients with worse clinical outcomes. Using the 2 instruments could be appropriate to conduct sensitivity analyses of cost-utility ratios because the instruments measure utility with closely similar measured properties, but at different levels. (First Release May 1 2011; J Rheumatol 2011;38: 1576–84; doi:10.3899/jrheum.101006)

Key Indexing Terms:
UTILITY SF-6D EQ-5D EARLY ARTHRITIS

From INSERM, CIC-EC CIE6, Nancy; Epidémiologie et Evaluation Cliniques, CHU Nancy, Nancy; Paul Verlaine Metz University, Paris Descartes University, EA 4360 Apemac, Nancy; Paris 6 – Pierre et Marie Curie University; Department of Rheumatology, Pitié-Salpêtrière Hospital, Paris; Department of Rheumatology, Lille University 2, Lille; and Department of Rheumatology, Amiens University, Amiens, France.

Supported by the French Society of Rheumatology. An unrestricted grant from Merck Sharp and Dohme (MSD) was allocated for the first 5 years of the ESPOIR cohort study. Two additional grants from INSERM were obtained to support part of the biological database. The French Society of Rheumatology, Abbott, Amgen, and Wyeth also supported the ESPOIR cohort study.

C. Gaujoux-Viala, MD, INSERM, CIC-EC CIE6; Epidémiologie et Evaluation Cliniques, CHU Nancy; Paul Verlaine Metz University, Paris Descartes University, EA 4360 Apemac; Paris 6 – Pierre et Marie Curie University; Department of Rheumatology, Pitié-Salpêtrière Hospital; A-C. Rat, MD, PhD; F. Guillemin, MD, PhD, INSERM, CIC-EC CIE6; Epidémiologie et Evaluation Cliniques, CHU Nancy; Paul Verlaine Metz University, Paris Descartes University, EA 4360 Apemac; R-M. Flipo, MD, PhD, Department of Rheumatology, Lille University 2; P. Fardellone, MD, PhD, Department of Rheumatology, Amiens University; P. Bourgeois, MD; B. Fautrel, MD, PhD, Paris 6 – Pierre et Marie Curie University; Department of Rheumatology, Pitié-Salpêtrière Hospital.

Address correspondence to Dr. C. Gaujoux-Viala, Université Paris 6–Pierre & Marie Curie, AP-HP, Groupe hospitalier Pitié-Salpêtrière, Service de Rhumatologie, 83 boulevard de l'Hôpital, 75651 Paris cedex 13, France. E-mail: cecilegaujouxviala@yahoo.fr

Accepted for publication February 9, 2011.

Preference-based measures of health have become important for estimating health states to calculate quality-adjusted life years, which are an essential component of cost-utility analysis. The EuroQol EQ-5D health outcome measure[1] and the SF-6D[2] are indirect preference-based health-related quality of life (HRQOL) instruments increasingly being used for economic evaluation of clinical interventions and health programs. Although the theoretical concept of utility implies that one specific health state has one utility score, regardless of how it is measured, different instruments can give different scores[3]. A review of these measures concluded that, among other items, a comparison of the preference-based measures across a range of conditions and severity is needed[4].

Several mainly cross-sectional studies have therefore compared EQ-5D and SF-6D scores for patients with a particular clinical condition; a common finding is small but important differences between the utility estimates by the 2 measures[5,6,7]. However, few comparisons exist in rheumatoid arthritis (RA)[8,9,10,11], especially in Europe, and no comparison has yet been conducted for early arthritis (EA), except a recent article comparing only the responsiveness in

a limited sample of patients with very early inflammatory arthritis (4–11 weeks' duration; n = 182)[12]. The broad expansion of drug development for RA and the revolution of early aggressive therapy have fueled the search for better approaches to establish cost-effectiveness in EA, but consensus is lacking on the choice of utility instrument. The choice of instrument may affect both the results of future studies of new biologic agents and their cost-effectiveness. There is a need for consensus based on the relative merits of the instruments from evidence of their practicality, reliability, construct validity, and discriminant validity, as well as their overall suitability for evaluative purposes. Thus, if the instrument properties are close but the utility levels elicited by the 2 instruments are different, sensitivity analyses using the 2 levels of utility could be appropriate to determine cost-utility ratios.

Our aim was to compare the EQ-5D and SF-6D in terms of their utility values and performance — i.e., acceptability (missing values), construct validity, and discriminant ability — in a large group of patients with EA over a period of 2 years.

MATERIALS AND METHODS

Patients. Between December 2002 and March 2005, we recruited 813 patients with EA from 14 French regional centers in the ESPOIR cohort[13]. Inclusion criteria were age 18 to 70 years, more than 2 swollen joints for > 6 weeks and < 6 months, suspected or confirmed diagnosis of RA, and taking no disease-modifying antirheumatic drugs or steroids (except if < 2 weeks). Patients were followed every 6 months during the first 2 years, then every year for at least 10 years. At baseline and at each visit, data for a set of clinical and biological variables were recorded, including that from the Disease Activity Score for 28 joints (DAS28), a composite index of disease activity[14]. At each visit, patients completed self-administered patient-reported outcome measures, including a functional ability questionnaire, the Health Assessment Questionnaire (HAQ)[15], and HRQOL questionnaires, the EQ-5D, and the MOS 36-item Short-form Health Survey (SF-36)[16]. The protocol of the ESPOIR Cohort study was approved by the ethics committee of Montpellier, France. All patients gave their signed informed consent before inclusion.

Utility measurement. The utility concept was developed by health economists. Assessment of utility assigns a numeric value from 0 to 1 for health states, 0 indicating death and 1 a state of perfect health. The values reflect the preference for a health state in a situation of choice that includes uncertainty or sacrifice (e.g., life-years). While methods such as standard gamble and time tradeoff may be used to measure health states directly, they are less suitable for clinical research and less widely used for feasibility reasons. Instead, indirect utility assessment techniques (EQ-5D and SF-6D) have been developed. The indirect health utility assessments involve population-assigned weights to calculate utility scores for particular health states from multidomain health-status questionnaires completed by patients[17] (Table 1).

Statistical analysis. EQ-5D and SF-6D utility scores were calculated by use of the scoring algorithms developed by Dolan[1] and Brazier, *et al*[2], respectively. Descriptive statistics [mean and standard deviation, median and interquartile range (IQR), minimum, maximum] and distributions of the EQ-5D and SF-6D utility scores were computed. Ceiling and floor effects were assessed and compared and considered present if > 15% of the respondents achieved the highest or lowest possible score[19]. The within-subject difference in mean utility scores of the 2 instruments was tested at baseline by paired t test. To test the difference between the 2 instruments, a limit of 0.03 between the scores was chosen on the basis of the smallest estimate of the minimal important difference (MID) for the SF-6D or EQ-5D published[7,10].

Agreement. The paired utility scores were presented graphically as scatterplots. Agreement between measures was analyzed by the intraclass correlation coefficient (ICC) and Bland-Altman plots for the entire sample and for subgroups categorized by disease activity (DAS28 ≤ 3.2, 3.2–5.1, and > 5.1) and functional ability (HAQ ≤ 1, 1–2, and > 2). The ICC was based on a 2-way random mixed-effects model, with absolute agreement. The Bland-Altman plots illustrate the magnitude of the difference between the 2 utility measures (SF-6D – EQ-5D) and show the distribution of the difference values over the entire range of the utility score.

Because the lower bounds of the 2 instruments differ and to document the agreement without this difference in scale, we standardized the utility scores. EQ-5D and SF-6D scores were transformed linearly to fit the range 0–1 to retain scale proportionality (based on the theoretically possible range).

Construct validity. To investigate whether the EQ-5D and SF-6D are valid measures of EA health status, we used Spearman's product-moment correlation to compare values for the 2 instruments with those for external measures of health, the HAQ, DAS28, and SF-36. Spearman correlation coefficients were compared with an appropriate t test[20].

Discriminant validity. One-way ANOVA was used to test whether the utility scores differed among different disease activity states and functional groups. The hypothesis is that utility scores decreased with higher disease activity and functional ability at the same timepoint. The influence of sociodemographic factors was analyzed by t test or ANOVA. The ability of the EQ-5D and SF-6D instruments to detect differences between health status measures by external indicators was tested by the relative efficiency statistic, widely used in HRQOL studies but only recently used to test utility[21]. The statistic is calculated as the ratio of the square of the t statistic of the comparator instrument (here SF-6D utility score) to the square of the t statistic of the reference instrument (here EQ-5D utility score). A relative efficiency score > 1.0 indicates that the SF-6D is more efficient than the EQ-5D in detecting differences. We used the cutoff points currently used to define the activity states of RA (DAS28 ≥ 3.2 for low disease activity, DAS28 > 5.1 for high disease activity; and HAQ score > 1 with a sharp drop in work capacity)[22].

All analyses involved use of SAS v9.1 (SAS Institute, Cary, NC, USA). A $p < 0.05$ was considered statistically significant.

RESULTS

Characteristics of the population. Table 2 shows the demographic and clinical characteristics of the 813 patients in the ESPOIR cohort at inclusion. In total, 578 (71.3%) patients fulfilled the American College of Rheumatology criteria for RA[23], which confirmed that patients were at high risk of developing RA. At 2 years, 692 patients were still being followed, and all characteristics, except for erosions and DAS28, were similar to those of the initial population.

Global utility scores. The distribution of utility scores was bimodal for the EQ-5D and near-normal for the SF-6D (Figure 1). At baseline, the mean utility score for the EQ-5D was 0.518 ± 0.306 (median 0.656, IQR 0.255–0.725); the mean utility score for the SF-6D was 0.582 ± 0.114 (median 0.580, IQR 0.519–0.646). The mean difference in utility scores for the 2 measures was 0.064 (95% CI –0.42 to 0.55) at baseline and was significantly different from 0.03, the MID for evaluative purposes ($p < 0.0001$).

The EQ-5D generated a minimum value of –0.594 and a

Table 1. Overview of instrument properties of the EQ-5D and SF-6D.

Domains (no. levels for each domain)	No. Questions	No. Possible Health States	Valuation Technique	Range
EQ-5D				
Mobility (3)	5 questions	243	Time tradeoff	–0.59 to 1.00
Usual activities (3)				
Self-care (3)				
Pain/discomfort (3)				
Anxiety/depression (3)				
SF-6D				
Physical function (6)	11 questions of the SF-36*	18,000	Standard gamble	0.296 to 1.00
Role limitation (4)				
Social function (5)				
Pain (6)				
Mental health (5)				
Vitality (5)				

* The SF-6D can also be calculated using the SF-12 as well as SF-36[18].

Table 2. Characteristics of patients included in the ESPOIR cohort at baseline (n = 813).

Characteristic	Mean ± SD
Age, yrs	48.11 ± 12.56
Female sex, n (%)	624 (76.7)
Years of education, n (%)	
< 5	101 (12.4)
6–12	457 (56.2)
> 12	255 (31.4)
Employed, n (%)	481 (59.2)
Married or living together, n (%)	594 (73)
Disease Activity Score for 28 joints†	5.11 ± 1.31
Health Assessment Questionnaire score	0.979 ± 0.684
Erythrocyte sedimentation rate	29.4 ± 24.6
C-reactive protein level*, mg/l	20.3 ± 32.4
Rheumatoid factor*, n (%)	376 (45.8)
Anti-CCP2 antibodies*, n (%)	315 (38.8)
van der Heijde modified Sharp score**	5.97 ± 10.14
EQ-5D score	0.52 ± 0.31
SF-6D score	0.58 ± 0.11

† 9.5% of patients had a DAS28 ≤ 3.2, 44.6% DAS28 3.2–5.0 and 45.9% DAS28 > 5.1. * Baseline C-reactive protein level (normal < 10 mg/l); IgM and IgA rheumatoid factor (ELISA, Menarini, France; positive > 9 IU/ml); and anti-CCP2 antibodies (ELISA, DiaSorin, France; positive > 50 U/ml) were detected in all patients using the same technique in a central laboratory (Paris-Bichat). ** Of 715 (22.3%) patients, 160 had erosions on hands and/or feet at baseline.

maximum value of 1.0, with 11.8% of patients in health states considered worse than dead and 1.5% with a corresponding utility score of 1.0. In contrast, the SF-6D generated a minimum value of 0.301 and a maximum value of 0.923. Thus, no significant floor or ceiling effect was found at baseline. However, at 6 months, 6% of patients had an EQ-5D utility score of 1.0, and this proportion increased at 1 year, then remained stable over time, at ~12%. The proportion of patients with an SF-6D utility score of 1.0 remained low, between 0.5% and 0.7%. Few missing values were observed: 1.2% for the SF-6D and 0.6% for the EQ-5D at baseline, and 1% and 0.3%, respectively, at 2 years.

Agreement. A scatterplot of the EQ-5D and SF-6D utility scores is shown in Figure 2; the Spearman product-moment correlation coefficient was 0.71 ($p < 0.0001$). This high correlation between SF-6D and EQ-5D was stable over 2 years. However, deviations from the 45-degree line of perfect agreement are evident, particularly at the low end of the utility scales.

At baseline, ICC agreement between the instruments was low, 0.42 (95% CI 0.37–0.48), but increased to 0.53 (95% CI 0.47–0.58) at 6 months and 0.57 (95% CI 0.52–0.62) at 1 and 2 years. Agreement decreased with increasing disease activity and functional disability at each timepoint (Table 3).

At baseline, the Bland-Altman plot displayed lack of agreement between the 2 measures, with a systematic variation in the EQ-5D and SF-6D scores: less healthy individuals (mean score < 0.4) showed high scores on the SF-6D, and healthier individuals (mean score > 0.5) showed high scores on the EQ-5D (Figure 3). The Bland-Altman limits of agreement for the 2 utility scores ranged from –0.42 to 0.55 for all patients. The lack of agreement was notable at the low end of the utility scale and increased with increasing disease activity. The agreement improved at 6 months and then remained stable: the Bland-Altman limits of agreement were from –0.34 to 0.36 for all patients. Despite improvement, agreement still tended to be poor with increased disease activity (Figure 3).

Recalculated ICC values, with transformation of the 2 utility scores to fit the range 0–1, were higher than without transformation and were stable over time and ranged from 0.64 to 0.68 (Table 3). No decrease in agreement with increasing disease activity or functional disability was observed with transformed ICC (data not shown).

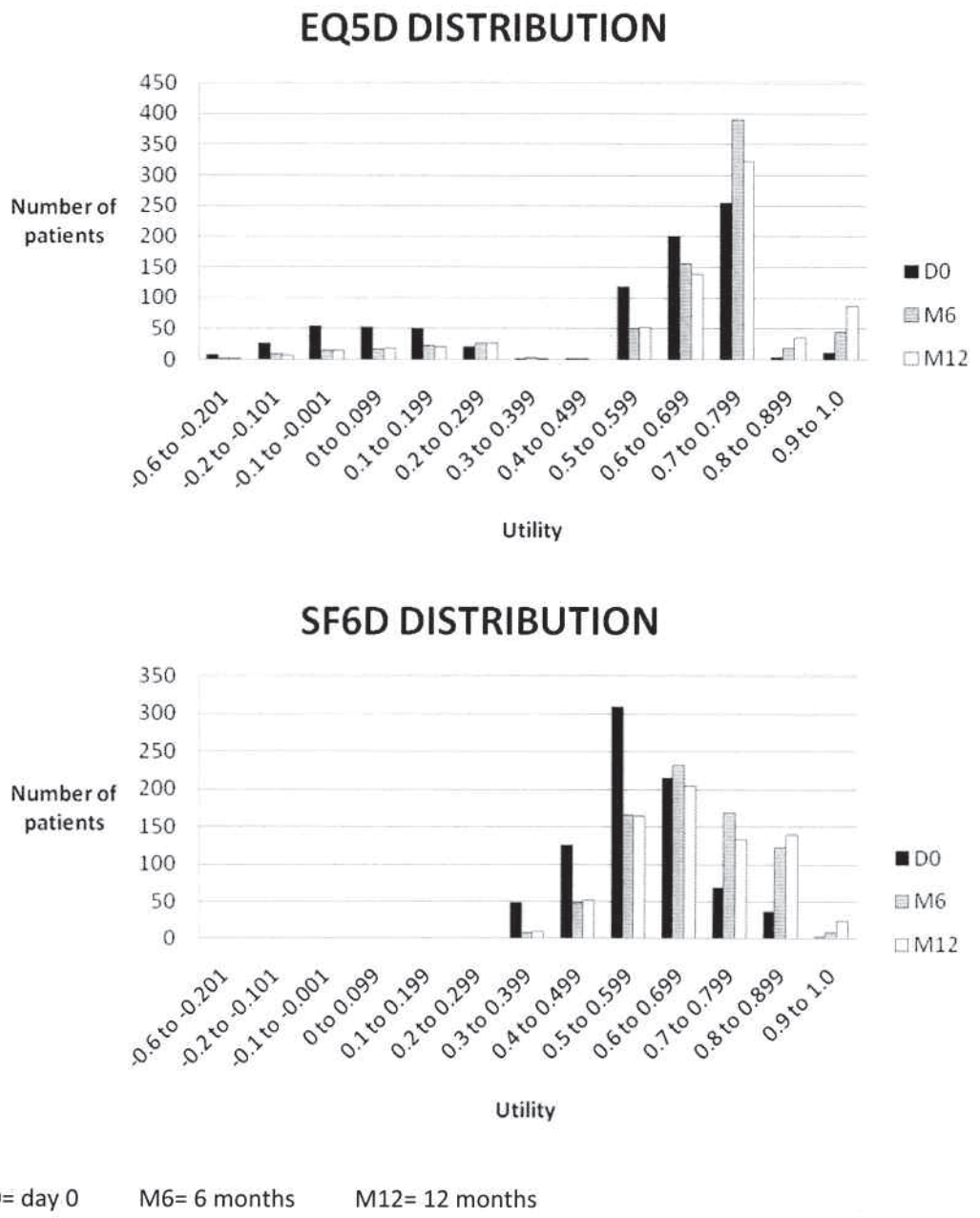

Figure 1. Frequency distribution of SF-6D and EQ-5D utility scores over time.

Construct validity. At baseline, correlation with the DAS28 was similar and moderate (r = –0.47 and –0.42 for the SF-6D and EQ-5D, respectively, $p < 0.04$), and correlations with the HAQ score and the physical component of the SF-36 were similar and good (r = –0.70 with the HAQ for both utility measures, and r = 0.64 and 0.59 for the SF-6D and EQ-5D, respectively, with the physical component of the SF-36, $p < 0.01$). However, correlation with the mental component of the SF-36 was better with the SF-6D than with the EQ-5D (r = 0.69 and 0.53, $p < 0.0001$), and correlation with pain at rest was weak (r = –0.35 and –0.28, respectively, $p < 0.006$). Correlation with the HAQ score, DAS28, and the physical component of the SF-36 remained stable over the 2 years. Correlation with the mental component of the SF-36 and pain at rest was markedly improved at 6 months and then remained stable, but was always better with the SF-6D than the EQ-5D for the mental component of the SF-36 (r = 0.77–0.80 for the SF-6D and 0.61–0.62 for the EQ-5D) and was stable and similar for the 2 utility measures for pain (r = –0.45 for the SF-6D and EQ-5D at 6 months and r = –0.52 to –0.55 thereafter; Table 4).

Discriminant validity. The utility scores did not differ by age

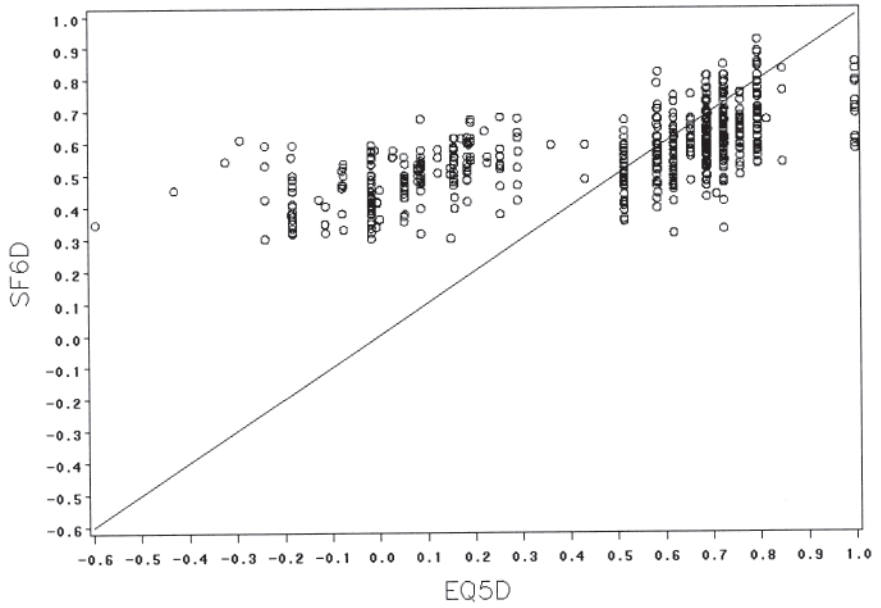

Figure 2. Comparison of EQ-5D with SF-6D.

Table 3. Intraclass correlation coefficients (ICC) between SF-6D and EQ-5D for all patients and for several subgroups categorized by disease activity and by functional disability over time.

	Global ICC		ICC by Disease Activity		ICC by Functional Disability	
Time, mo	Initial	Recalculated*	DAS28	Initial (95% CI)	HAQ	Initial (95% CI)
Baseline	0.42 (0.36; 0.48)	0.64 (0.59; 0.68)	≤ 3.2	0.42 (0.19; 0.60)	≤ 1	0.36 (0.27; 0.43)
			3.2–5.1	0.42 (0.33; 0.51)	1–2	0.27 (0.16; 0.38)
			> 5.1	0.36 (0.26; 0.45)	> 2	0.17 (–0.10; 0.41)
M6	0.53 (0.47; 0.58)	0.64 (0.59; 0.68)		0.52 (0.44; 0.59)		0.48 (0.41; 0.54)
				0.48 (0.39; 0.57)		0.32 (0.16; 0.47)
				0.38 (0.19; 0.55)		0.22 (–0.38; 0.69)
M12	0.57 (0.52; 0.62)	0.68 (0.64; 0.72)		0.58 (0.51; 0.64)		0.56 (0.51; 0.62)
				0.47 (0.36; 0.56)		0.31 (0.14; 0.46)
				0.34 (0.09; 0.54)		0.35 (–0.36; 0.81)
M18	0.58 (0.52; 0.62)	0.68 (0.64; 0.71)		0.69 (0.53; 0.65)		0.56 (0.50; 0.61)
				0.43 (0.31; 0.53)		0.32 (0.16; 0.47)
				0.48 (0.21; 0.68)		0.27 (–0.61; 0.85)
M24	0.57 (0.52; 0.62)	0.66 (0.62; 0.70)		0.58 (0.51; 0.64)		0.55 (0.49; 0.61)
				0.41 (0.29; 0.52)		0.28 (0.10; 0.44)
				0.37 (0.08; 0.60)		–0.03 (–0.62; 0.58)

* EQ-5D and SF-6D utility scores were transformed linearly to fit the range 0–1 to compare scales and retain scale proportionality.

(p = 0.14 and p = 0.12 for the SF-6D and EQ-5D, respectively), sex (p = 0.12 and p = 0.50), or marital status (p = 0.55 and p = 0.29). Both utility scores increased with number of years of education. Both utility measures showed statistically significant differences by disease activity (DAS28 low, moderate, and high disease activity) and functional disability (HAQ ≤ 1, 1–2, > 2) (p < 0.0001). Both measures generated utility scores that decreased with increasing disease activity or functional disability. The difference in scores between the low and high disease activity groups was greater for the EQ-5D (0.25, 95% CI 0.17–0.33) than for the SF-6D (0.12, 95% CI 0.09–0.15).

Considering the cutoff point for low disease activity (DAS28 ≤ 3.2) at baseline (n = 75), the relative efficiency score was 1, so the SF-6D had the same efficiency as the EQ-5D in identifying patients with low disease activity. Considering the cutoff point for high disease activity (DAS28 > 5.1) at baseline (n = 360), the relative efficiency score was 1.40, so the SF-6D was 40% more efficient than the EQ-5D in identifying patients with high disease activity. When patients were dichotomized at baseline in terms of functional disability (HAQ > 1; n = 347), the relative efficiency score was 1.29, so the SF-6D was 29% more efficient than the EQ-5D in identifying patients with HAQ > 1. But

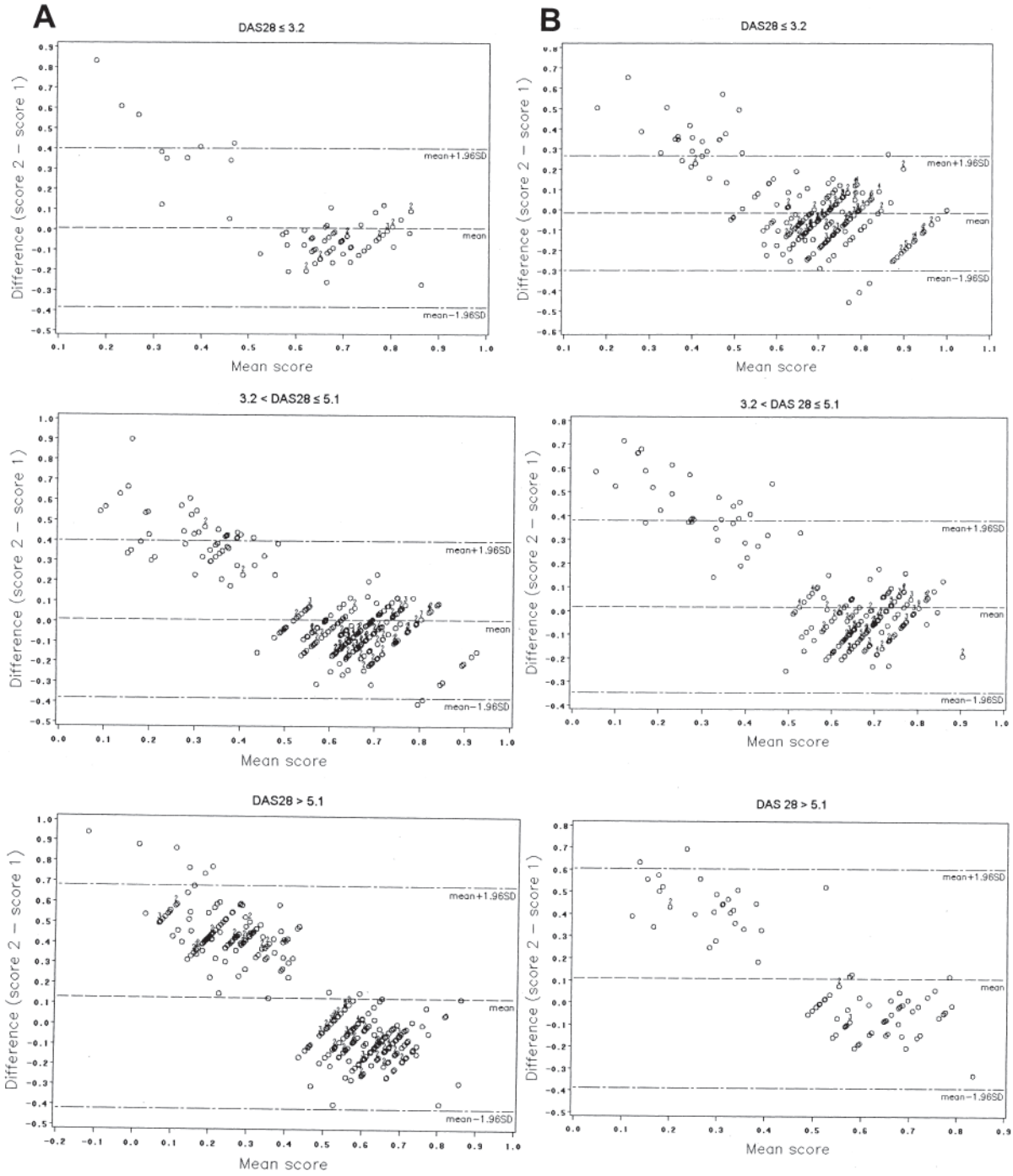

Figure 3. Bland-Altman plots of differences in SF-6D and EQ-5D utility scores for all patients by disease activity at baseline (A) and at 6 months (B). Score 2 = SF-6D; score 1 = EQ-5D.

for HAQ > 2 (n = 56), the relative efficiency score was 0.70, so the EQ-5D was 30% more efficient than the SF-6D in identifying patients with HAQ > 2 (Table 5).

DISCUSSION

Although the correlation between the 2 utility scores, the EQ-5D and SF-6D, was high, descriptive statistics revealed systematic disagreement at both the low and high ends of the utility scales. In particular, EQ-5D values < 0.5 corresponded to markedly high SF-6D scores. In addition, a wide range of SF-6D scores (0.58–0.85) was associated with an EQ-5D score of 1.0. Bland-Altman plots also displayed lack of agreement between the 2 measures, particularly at the low end of the utility scales. Our results were similar to those

Table 4. Correlations of the EQ-5D and SF-6D with external measures of health, the HAQ, DAS28, and SF36. Data are Spearman's product-moment correlation.

Time, mo	DAS28		HAQ		SF-36 Physical Component		SF-36 Mental Component		Pain at Rest	
	EQ-5D	SF-6D	EQ-5D	SF-6D	EQ-5D	SF-6D	EQ-5D	SF-6D	EQ-5D	SF-6D
Baseline	–0.42	–0.47	–0.70	–0.70	0.59	0.64	0.53	0.69	–0.28	–0.35
M6	–0.43	–0.41	–0.60	–0.60	0.59	0.64	0.62	0.79	–0.45	–0.45
M12	–0.50	–0.51	–0.68	–0.67	0.69	0.72	0.61	0.77	–0.59	–0.58
M18	–0.50	–0.53	–0.68	–0.69	0.69	0.73	0.62	0.80	–0.52	–0.55
M24	–0.48	–0.52	–0.69	–0.71	0.66	0.73	0.61	0.77	–0.55	–0.54

Table 5. Discriminant capacity of the EQ-5D anc SF-6D.

Outcome	SF-6D	p	EQ-5D	p	Relative Efficiency
Disease activity					
DAS28 ≤ 3.2	0.653	< 0.0001*	0.646	< 0.0001*	1.00
3.2–5.1	0.619		0.611		
> 5.1	0.532		0.400		1.40
Functional disability					
HAQ ≤ 1	0.637	< 0.001*	0.659	< 0.0001*	1.29
1–2	0.524		0.385		
> 2	0.426		0.036		0.70

*ANOVA.

found for heterogeneous RA[8,11]. The explanation could lie in the difference in the "true" range of the theoretical 0–1 utility scale the instruments actually cover. The lowest observed value was –0.594 for the EQ-5D and 0.301 for the SF-6D. Therefore, the observation that differences between instruments were especially high with worse disease is not surprising. As a consequence, the mean EQ-5D showed larger differences between groups with better and worse disease defined by the DAS28 or HAQ. This result has important consequences when using the instruments in clinical trials and for cost-effectiveness analyses of patients with high disease activity: the gain in EQ-5D will be larger and will provide more favorable incremental cost-utility values[9]. To determine whether the poor agreement was due only to differences in the scaling of these 2 instruments, we recalculated the ICC after transforming utility values into a 0–1 scale. After rescaling, the ICC were increased but remained moderate. This finding suggests that observed differences in the ICC are not due merely to differences in the scaling of these 2 instruments. Mean SF-6D utility scores exceeded mean EQ-5D utility scores by 0.064, which is significantly higher than the MID for the SF-6D (MID = 0.033)[24] and the EQ-5D (MID = 0.03, postulated to be the minimum clinically important difference because it is the smallest of the coefficients in the York weights, that is, the smallest difference in moving from one level to another on any of the 5 dimensions)[10].

Several reasons might explain the differences between the utility scores. First, the health descriptive system of the SF-6D does not allow for negative values and so assigns a 0.296 value to the most severe health state produced by the descriptive system, whereas the EQ-5D score allows for negative scores[25]. Second, EQ-5D utility scores are based on time tradeoff, which tends to result in high values for mild states, whereas SF-6D scores are based on standard gamble, which tends to result in high values for severe states[26,27]. A further explanation for why healthier individuals showed higher scores on the EQ-5D than on the SF-6D is that the SF-6D may be more sensitive (because of its larger descriptive system) for patients experiencing mild to moderate health problems[2]. Lower utility scores were observed for EQ-5D in patients with severe disabilities. This result may be explained by the content of the EQ-5D. Of the 5 dimensions, 4 (mobility, self-care, usual activity, and pain/discomfort) are likely to be particularly affected in patients with EA. A study comparing EQ-5D and SF-6D in 7 diseases[25] showed larger mean differences between the 2 instruments in osteoarthritis than in diseases focusing on pain and discomfort such as irritable bowel syndrome. We found that the patients with a score worse than death on EQ-5D (n = 90) had higher scores on the pain and physical function, and a large proportion of these patients scored maximum on the pain dimension and moderate on all other dimensions (data not shown), confirming results of a study investigating the health states of patients with inflammatory arthritis with a score worse than death on EQ-5D[28]. Differences between the utility scores may also be confounded by the valuation and/or scoring methods. The instruments use different operational definitions of the domains and functional levels within each domain.

The level of agreement for the 2 measures improved at 6 months and then remained stable. The first explanation for this observation is that disease activity decreased with treatment, and agreement was better for healthier patients. However, considering agreement in different disease-activity and functional-ability groups, we still observed improvement of agreement at 6 months, especially for low disease activity.

Correlations of the 2 scales with the DAS28, HAQ score, and the physical component of the SF-36 were moderate to good and were stable over 2 years. In contrast, scores for the

mental component of the SF-36 and pain at rest correlated better with the SF-6D than the EQ-5D at baseline, were improved at 6 months and remained stable thereafter, and were similar for pain. The improvement in the correlations could be explained by the importance of the mental health component and how patients deal with and are able to cope with a recent diagnosis of a chronic disease in terms of utility. The improvement in agreement and correlations at 6 months could also be explained by patients becoming used to completing questionnaires. These results should be interpreted with caution, keeping in mind that the SF-6D is derived from the SF-36 (using 11 of the 36 questions). Stronger correlations found between the SF-36 and SF-6D than between the SF-36 and EQ-5D do not necessarily mean that the SF-6D has better properties, and are in part due to the fact that the SF-36 and SF-6D use the same items. However, it is interesting that correlations with the physical component of the SF-36 were similarly good for both measures of utility, whereas correlation with the mental component of the SF-36 was better with the SF-6D than with the EQ-5D.

Our study has some limitations. We did not compare the test–retest reliability of these 2 instruments. Comparison of the metric properties of the instruments was hampered because the EQ-5D scores showed a high level of skewness compared with the normal distribution of the SF-6D scores. Classical approaches to study agreement assume normality. Of note, the change values of the utilities showed a near-normal distribution. Finally, the scoring algorithms used for the 2 instruments were developed from data for a general population in the United Kingdom because no such algorithm was available in France at the time of the study. Use of an algorithm from the same population for both the EQ-5D and SF-6D might result in a more valid comparison.

One of the strengths of this study is that a broad group of patients with EA was included. The ESPOIR cohort aims to include all patients with EA regardless of disease level, age, and sex, and our study shows the performance of the instruments in a real-life setting. The study also includes a large number of patients with longitudinal assessment.

Further research to examine the psychometric properties of the EQ-5D and SF-6D, in particular sensitivity to change, would strengthen the limited evidence currently available to analysts. Future research should focus on understanding the reasons for the differing performance of the 2 utility measures in EA. The objective is to determine which of the 2 instruments is the more pertinent, or if cost-utility analysis should include both EQ-5D and SF-6D in sensitivity analyses.

ACKNOWLEDGMENT

We thank Nathalie Rincheval, who did expert monitoring and data management; and investigators who recruited and followed the patients: F. Berenbaum, Paris-Saint Antoine; M.C. Boissier, Paris-Bobigny; A. Cantagrel, Toulouse; B. Combe, Montpellier; M. Dougados, Paris-Cochin; P. Goupille, Tours; F. Liote, Paris-Lariboisière; X. Le Loet, Rouen; X. Mariette, Paris-Bicêtre; O. Meyer, Paris-Bichat; A. Saraux, Brest; T. Schaeverbeke, Bordeaux; J. Sibilia, Strasbourg.

REFERENCES

1. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095-108.
2. 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.
3. 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.
4. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Assess 1999;3:1-164.
5. Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. Health Econ 2003;12:1061-7.
6. Xie F, Li SC, Luo N, Lo NN, Yeo SJ, Yang KY, et al. Comparison of the EuroQol and Short Form 6D in Singapore multiethnic Asian knee osteoarthritis patients scheduled for total knee replacement. Arthritis Rheum 2007;57:1043-9.
7. Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged > or = 45 years. Health Econ 2008;17:815-32.
8. Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care 2004;42:1125-31.
9. Marra CA, Marion SA, Guh DP, Najafzadeh M, Wolfe F, Esdaile JM, et al. Not all "quality-adjusted life years" are equal. J Clin Epidemiol 2007;60:616-24.
10. 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.
11. Lillegraven S, Kristiansen IS, Kvien TK. Comparison of utility measures and their relationship with other health status measures in 1041 patients with rheumatoid arthritis. Ann Rheum Dis 2010;69:1762-7.
12. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Verstappen SM, Watson K, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. Qual Life Res 2009;18:1195-205.
13. Combe B, Benessiano J, Berenbaum F, Cantagrel A, Daures JP, Dougados M, et al. The ESPOIR cohort: a ten-year follow-up of early arthritis in France: methodology and baseline characteristics of the 813 included patients. Joint Bone Spine 2007;74:440-5.
14. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44-8.
15. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
16. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). Conceptual framework and item selection. Med Care 1992;30:473-83.
17. Khanna D, Tsevat J. Health-related quality of life — an introduction. Am J Manag Care 2007;13 Suppl 9:S218-23.
18. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851-9.
19. Veenhof C, Bijlsma JW, van den Ende CH, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis

questionnaires: a systematic review of the literature. Arthritis Rheum 2006;55:480-92.
20. Guilford JP, Fruchter B. Fundamental statistics in psychology and education. New York: McGraw-Hill; 1965:190.
21. Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. Health Econ 2005;14:1169-89.
22. Kobelt G, Woronoff AS, Richard B, Peeters P, Sany J. Disease status, costs and quality of life of patients with rheumatoid arthritis in France: the ECO-PR Study. Joint Bone Spine 2008;75:408-15.
23. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
24. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4.
25. 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.
26. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ 1996;15:209-31.
27. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. J Health Econ 2006;25:334-46.
28. Harrison MJ, Lunt M, Verstappen SM, Watson KD, Bansback NJ, Symmons DP. Why do patients with inflammatory arthritis often score states "worse than death" on the EQ-5D? An investigation of the EQ-5D classification system. Value Health 2009;12:1026-34.