British Journal of Rheumatology 1997;**36:**551–559

MEASURING HEALTH-RELATED QUALITY OF LIFE IN RHEUMATOID ARTHRITIS: VALIDITY, RESPONSIVENESS AND RELIABILITY OF EUROQOL (EQ-5D)

N. P. HURST, P. KIND,* D. RUTA,† M. HUNTER and A. STUBBINGS

*Economic & Health Outcomes Unit, Department of Rheumatology, Western General Hospitals Trust, Crewe Road, Edinburgh EH4 2XU, *Centre for Health Economics, University of York, York YO1 5DD and †Department of Public Health, Tayside Health Board, Dundee*

SUMMARY

The EuroQol (EQ-5D) generic health index comprises a five-part questionnaire and a visual analogue self-rating scale. The questionnaire may be used as a health index to calculate a 'utility' value or as a health profile. The validity, reliability and responsiveness of EQ-5D were tested in 233 patients with rheumatoid arthritis stratified by functional class. EQ-5D demonstrated moderate to high correlations with measures of impairment and high correlations with disability measures. Stepwise regression models showed that EQ-5D utility values and visual analogue scores were explained best as a function of pain, disability, disease activity and mood ($R^2 \sim 70\%$), although other variables (side-effects, years of education) were required to explain the visual analogue scores. The EQ-5D health index and visual analogue scale are more responsive than any of the other measures, except pain and doctor-assessed disease activity. The reliability of the EQ-5D index and EQ-5D visual analogue scale is as good or better than that of all other instruments except the Health Assessment Questionnaire. Some patients with severe long-standing disease had health states which attracted utility values below zero, i.e. from a societal perspective they were regarded as being in states 'worse than death'. The practical and ethical implications of these utility valuations are discussed, and at present the utility values should be used and interpreted with caution. With this caveat, EQ-5D is simple to use, valid, responsive to change and sufficiently reliable for group comparisons. It is of potential use as an outcome measure in clinical trials, audit and health economic studies, but further work is required on its performance in other clinical contexts and on the interpretation of the utility values.

KEY WORDS: Quality of life, Utility, Health status, Outcome, Rheumatoid arthritis, Disease activity, EuroQol, Validity, Responsiveness, Reliability.

THERE is growing interest in the development of generic instruments which can be used to measure health-related quality of life (HR-QOL) across a wide spectrum of diseases and conditions. So-called condition-specific instruments clearly have an essential role in the measurement of those aspects most closely related to disease process; simple examples include the erythrocyte sedimentation rate (ESR) in inflammatory conditions, serum creatinine in renal failure or peak flow rate in asthma. However, there is also a need for generic instruments which capture the overall impact of disease as well as the beneficial and detrimental effects of treatment on the individual [1]. A further consideration is that priorities for resource allocation within the NHS will be based increasingly on evidence of the cost-effectiveness of medical interventions on HR-QOL [2]; the reliability of such evidence is substantially dependent on the validity of the methods used to measure health status. Against this background, we have been evaluating the performance of two different generic instruments—the MOS-Short Form 36 (SF36) health profile [3] and the EuroQol (EQ-5D) [4, 5]—in rheumatoid arthritis (RA) and in this report we present our findings on the performance of the second of these measures.

The EQ-5D is a two-part instrument. Part 1 records self-reported problems on each of five 'domains': mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each domain is divided into three levels of severity corresponding to no problem, some problem and extreme problem. By combining one level from each of the five domains a total of 3^5, i.e. 243 'health states' are defined [4, 6], and in a previous small-scale study of EQ-5D in RA [5], the weights used for these 243 states had been obtained from visual analogue scale (VAS) ratings. However, since then a set of values has been obtained from a large representative sample of the adult population of England, Scotland and Wales [6]. A time trade-off procedure (TTO) was used to elicit utility weights for EQ-5D health states from some 3395 respondents. These weights lie on a scale on which full health and death score 1 and 0, respectively. Some severe health states attract negative scores, indicating that from a societal perspective being in these states is regarded as worse than death. Part 2 of the questionnaire records the subject's self-assessed VAS rating of health on a vertical 20 cm line on which the best and worst imaginable health states score 100 and 0, respectively.

Data from EQ-5D can be represented in three distinct forms. Part 1 may be presented either as a profile ($EQ\text{-}5D_{profile}$), based on the unweighted responses indicating a patient's level of problem in each of the five domains, or as a health index ($EQ\text{-}5D_{utility}$) by applying a suitable weighting system such as the utilities obtained from the UK national survey. The VAS rating in Part 2 can be interpreted directly as a

Submitted 12 August 1996; revised version accepted 3 December 1996.

quantitative measure of the patient's valuation of their own global health status (EQ-5D$_{vas}$).

In the current study, we have tested the validity, responsiveness and reliability of all three forms of EQ-5D in a sample of RA patients stratified according to functional class.

PATIENTS AND METHODS

Sample size

A sample size of 240 was selected on the basis that a relationship between any two measurements would be detected at the 5% significance level if their true correlation was >0.2, with an 80% power, and that a 20% drop-out rate would occur. The sample was stratified by functional class [7] to obtain a broad cross-section of disease severity. To achieve this, recruitment of consecutive patients into each functional class continued until 60 patients had been entered in each class.

Statistical methods

The change scores over time for all instruments were normally distributed, but the distribution of several of the instrument scores at baseline and follow-up, including for example the HAQ and EQ-5D$_{utility}$, were non-Gaussian. Analysis and comparison of data by either parametric or non-parametric methods gave virtually identical results, and in general parametric statistics gave more conservative estimates of significance. For this reason, both parametric and non-parametric tests of association and difference are presented to allow comparison of results. Only non-parametric methods were used to analyse the EQ-5D$_{profile}$ scores.

The construct validity of EQ-5D was tested first by examining the correlation between EQ-5D scores, scores from condition-specific instruments and measures of socioeconomic status. Stepwise linear multiple regression analysis was then used to model the relationship between EQ-5D and condition-specific instruments; plots of residuals from the regression equations were normally distributed. The stability of regression models was checked by repeating regression analyses using 3 months follow-up data.

Responsiveness to clinical change was tested in patients reporting change in their arthritis over 3 months. A change score with 95% confidence intervals (CI) was calculated for each instrument. The standardized response mean (SRM), which is a measure of 'signal to noise' ratio, and is defined as the ratio of mean change (δ) to the standard deviation (σ) of the change scores (i.e. δ/σ_{change}) in the population of patients reporting change, allows a direct comparison of the responsiveness of each instrument [8, 9]. SRMs were calculated for each instrument in the group of patients reporting improvement. σ_{change} may be affected both by measurement error and by variance in the biological response. To try to reduce the effect of biological variance on the SRM, we therefore also calculated an SRM (designated SRM*) using the σ_{change} in stable subjects, i.e. those reporting no change over 3 months; this enables direct comparison of δ in those reporting improvement to intra-subject variation over time in stable subjects.

Reliability was tested under two sets of conditions: first over a 3 month period in patients reporting no change in their arthritis, and in a second test, a group of 31 patients was asked to complete a second set of questionnaires after a 2 week interval. Parametric and non-parametric methods were used to test reliability. A change score with 95% CI and a reliability coefficient (intra-class correlation coefficient) (ICC) [10] was computed for each instrument. The ICC, which is derived from analysis of variance, is defined as $ICC = \sigma^2_{pat}/(\sigma^2_{pat} + \sigma^2_{error})$, where σ^2_{pat} is the estimated variance due to patients and σ^2_{error} is the estimated error variance. Values of ICC thus vary from 1 (perfectly reliable) to 0 (totally unreliable). The ICC was chosen in preference to the Pearson correlation which may overestimate reliability [10]. Also, because some of the scales have ordinal characteristics, 'Goodman and Kruskal's gamma', which provides a non-parametric measure of concordance, was computed for each scale.

Patient selection

The case notes of consecutive patients identified from clinic booking lists were reviewed to identify those with RA [11] 2 weeks before each out-patient clinic. Relatively few patients in functional class 4 attended as out-patients so these were also identified on admission to the rheumatology ward (Western General Hospital NHS Trust, Edinburgh) and by contacting GPs and nursing homes in the Lothian and Fife Regions. However, only 50 could be recruited into class 4, 10 fewer than the target number. In all, 245 RA patients were identified, 12 declined to participate and 233 RA patients were recruited. At 3 months, 224 were available for review, four had died and six had withdrawn because they were too ill or unwilling to continue.

The study was approved by the relevant medical ethics committee and all patients gave written consent.

Data collection

Demographic, socioeconomic data, American College of Rheumatology (ACR) disease activity measures [12]—swollen and graded tender joint count [13], modified Stanford Health Assessment Questionnaire (HAQ) [14], patient- and doctor-assessed disease activity (Likert scale), 10 cm visual analogue pain scale (pain-VA), erythrocyte sedimentation rate (ESR)—the Hospital Anxiety and Depression (HAD) Scale [15], presence or absence of co-morbidity or drug side-effects, radiological erosions (ever or never present) were collected. Patient questionnaires were presented in a single booklet, in half of which questionnaires were compiled in reverse order to avoid bias due to 'questionnaire fatigue'. Questionnaires were mailed to patients with a covering letter and consent form. Patients were asked to complete the forms just prior to their clinic appointment. On clinic attendance, the metrologist checked the responses for completeness to ensure that questions had not been omitted in error.

They did not, however, encourage or prompt responses to questions patients did not wish to answer. In the case of some severely disabled patients, the metrologist had to read the questions out and fill in the questionnaire on behalf of the patient. Assessments were performed at baseline, 3 and ~12 months. Here, we report the results of the 3 month follow-up.

Three metrologists were available, but over 95% of assessments were carried out by only two metrologists; to reduce inter-observer variation in the assessment of joint scores, the metrologists underwent a period of standardization training on six patients.

RESULTS

Patient characteristics

A total of 245 patients were identified for recruitment. Of these, 12 declined to take part and 233 (95%) were recruited. The mean age and duration of arthritis according to functional class are shown (Table I). The mean duration of RA increases by 5 or 6 yr between each functional class. At 3 month follow-up, 224 (96%) were available for review—four had died and six had withdrawn because they were too ill or unwilling to continue.

EQ-5D as a health profile (EQ-5D$_{profile}$)

The unweighted response (i.e. 1 = no problems, 2 = some problems, 3 = extreme problems) to the EQ-5D may be used as a descriptive profile. As a preliminary test of validity, the unweighted responses for the self-care, pain/discomfort and anxiety/depression domains were compared with the HAQ, pain-VA and HAD scales, respectively (Table II). For each of these three condition-specific scales, there is significant deterioration in score as the unweighted EQ-5D response deteriorates from level 1 to 3 (Table II). The usual activities and mobility domains do not have a direct counterpart to enable such a comparison.

The median unweighted score for patients in each functional class is shown (Table III). The percentage of patients reporting problems for each of the five EQ-5D domains is also presented according to functional class (Fig. 1). With increasing functional class, the proportion of patients reporting some or severe problems increases progressively in each of the five domains (Kruskal–Wallis test, $P < 0.001$). In functional class 4, patients with residual capacity to take a few steps or to transfer, frequently reported that they had difficulty walking rather than being unable to walk.

TABLE II
Mean (σ) and median (interquartile range) of condition-specific scales for patients classified according to unweighted score (1, 2 or 3) for three EQ-5D domains: self-care, pain/discomfort and anxiety/depression

	N	Mean (σ)	Median (IR)
	HAQ		
Self-care			
1	81	0.67 (0.53)	0.63 (0.63)
2	112	1.72 (0.51)*	1.81 (0.75)†
3	36	2.55 (0.28)*	2.63 (0.34)†
	Pain VA-scale		
Pain/discomfort			
1	11	4.6 (4.6)	4 (8)
2	158	42.0 (22.6)*	45 (33)†
3	61	78.1 (12.2)*	80 (19)†
	HAD-mood		
Anxiety/depression			
1	116	8.8 (5.0)	8.0 (7.0)
2	102	18.1 (6.4)*	18.0 (10)†
3	12	24.8 (7.3)*	26.5 (11.8)†

*Unpaired *t*-test: all $P < 0.001$.
†Mann–Whitney *U*-test: 1 *vs* 2: all $P < 0.0000$; 2 *vs* 3: all $P < 0.000$ except for mood $P = 0.0024$.

EQ-5D as a health index (EQ-5D$_{utility}$) and self-rating scale (EQ-5D$_{vas}$)

Several hypotheses regarding the construct validity of EQ-5D$_{utility}$ and EQ-5D$_{vas}$ were tested. These included the hypotheses that lower values would be associated with worse functional class, lower socioeconomic class, dependency (i.e. patient reported living with a 'carer' as opposed to a spouse or partner), increased disease activity measured using the ACR core set, lowered mood, medical co-morbidity and drug side-effects.

Functional class (Table IV). The EQ-5D$_{utility}$ value discriminates well between each functional class; patients in class 4 have mean EQ-5D$_{utility}$ close to zero with many patients having health states rated 'worse than death' in terms of the population-based weights.

The EQ-5D$_{vas}$ discriminates well between classes 1, 2 and 3, but not between classes 3 and 4. For comparison, the HAQ scores for each functional class

TABLE I
Patient characteristics

Functional class	*N*	Age		Duration of RA	
		yr (S.D.)	(range)	yr (S.D.)	(range)
I	60	49 (14)	(24–77)	5 (7)	(0.15–30)
II	63	53 (15)	(21–80)	11 (12)	(0.2–65)
III	60	59 (12)	(26–87)	16 (11)	(1–40)
IV	50	65 (11)	(39–86)	23 (14)	(4–58)
Males	45	58 (13)	(26–79)	9 (8)	(0.2–29)
Females	188	55 (15)	(21–87)	14 (13)	(0.2–65)
Total	233	56 (14)	(21–87)	13 (13)	(0.2–65)

TABLE III
Median unweighted response for each EQ-5D domain by functional class

Functional class	Mobility	Self-care	Usual activities	Pain	Mood
I	1	1	2	2	1
II	2	2	2	2	1
III	2	2	2	2	2
IV(a)*	2	3	3	2	2
IV(b)*	3	3	3	2	2

*IV(a) = patients with some residual capacity to walk within the home; IV(b) = patients totally unable to walk.

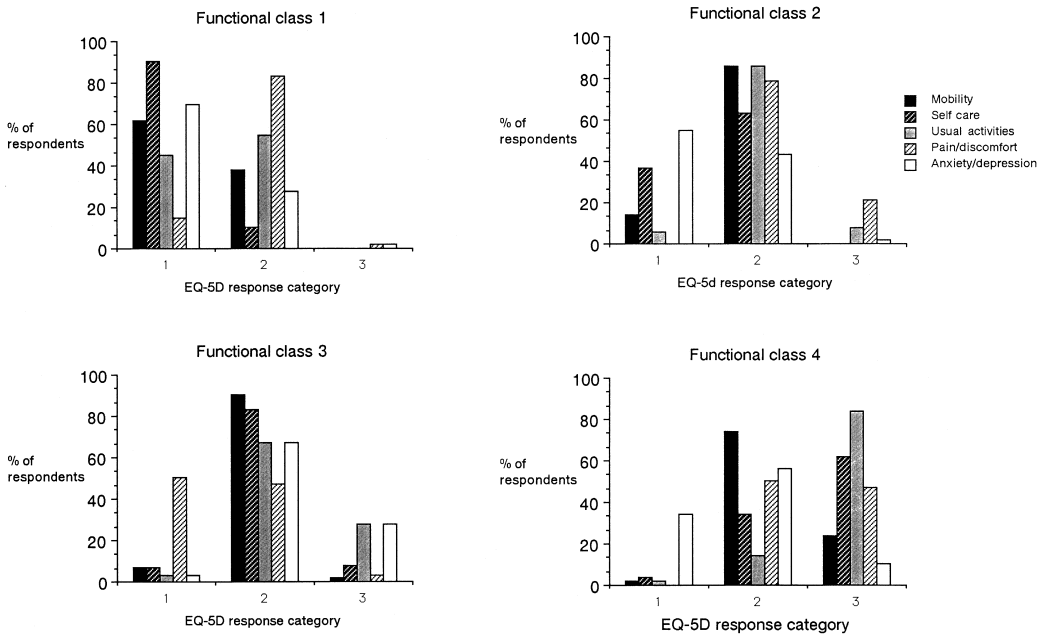

FIG. 1.—EQ-5D health profiles classified by functional class. The percentage of respondents reporting no problems (response = 1) diminishes, while the percentage with some problems (response = 2) or extreme problems (response = 3) increases in each EQ-5D domain with increasing functional class.

are reported which show the expected decline with increasing functional class.

Socioeconomic status, social support and employment. The majority of patients lived with a spouse or partner (67%), lived in owner-occupied property (65%) or were unemployed (71%).

Mean EQ-5D$_{utility}$ was significantly lower in those living with a spouse/partner compared with those living independently ($P < 0.05$), and patients who reported living with a 'carer' had significantly lower EQ-5D$_{utility}$ than either of these two groups ($P < 0.01$). The type of accommodation was used as a proxy for socioeconomic class; patients living in owner-occupied property had significantly higher EQ-5D$_{utility}$ than those living in private or council-rented accommodation.

On the EQ-5D$_{vas}$ scale, patients living independently rated their health significantly higher than those living with a carer ($P < 0.01$) or a spouse ($P < 0.01$), but there was no difference between the latter two groups. No significant differences were detected when patients were classified according to accommodation.

Patients still in employment had significantly higher EQ-5D$_{utility}$ and EQ-5D$_{vas}$ scores than those who were retired due to disability or who were otherwise not employed ($P < 0.001$).

Disease activity. Both EQ-5D$_{utility}$ and EQ-5D$_{vas}$ show similar and statistically significant correlations with ACR disease activity measures and each of the other variables (Table V). Correlations were strongest with measures of disability. Both were also significantly correlated with HAD score, duration of RA, radiological erosions, years of education, co-morbidity and age. In general, EQ-5D$_{utility}$ values are more strongly correlated with measures of disease activity than EQ-5D$_{vas}$.

Stepwise forward linear multiple regression showed that HAQ, HAD-mood, pain-VA and patient-assessed disease activity were significant and consistent predictors of EQ-5D$_{utility}$ values both at baseline and at 3 months follow-up; at the 3 month assessment, the ESR also entered the regression equation (Table VI). HAQ, HAD-mood and pain-VA were also consistent predictors of the EQ-5D$_{vas}$ score at both baseline and 3 month assessment (Table VI). At baseline, three other independent variables (side-effects, patient-assessed

TABLE IV
Mean (σ) and median (interquartile range) EQ-5D$_{utility}$ and EQ-5D$_{vas}$, and HAQ scores classified by functional class

Functional class	N	Mean (σ)	Median (IR)
EQ-5D$_{utility}$			
1	60	0.73 (0.14)	0.73 (0.11)
2	63	0.47 (0.26)*	0.59 (0.41)†
3	60	0.24 (0.31)*	0.12 (0.53)†
4	50	0.02 (0.31)*	0.08 (0.37)†
EQ-5D$_{vas}$			
1	60	76.8 (14.7)	80 (21)
2	63	58.3 (19.2)*	60 (30)†
3	60	43.6 (17.5)*	44 (23)†
4	50	45.0 (23.2) ns	50 (30) ns
HAQ score			
1	59	0.49 (0.45)	0.38 (0.63)
2	62	1.22 (0.42)*	1.25 (0.63)†
3	58	1.93 (0.33)*	2.00 (0.38)†
4	50	2.45 (0.33)*	2.50 (0.38)†

*Unpaired t-test: $P < 0.001$; ns = not significant.
†Mann–Whitney U-test: $P < 0.000$; ns = not significant.

TABLE V
Correlation between EQ-5D and disease-specific measures and demographics at baseline assessment. R is the Spearman rank correlation coefficient

	$EQ\text{-}5D_{utility}$ R	$EQ\text{-}5D_{vas}$ R
†HAQ	−0.78	−0.61
Functional class	−0.74	−0.55
†Pain-VA scale	−0.73	−0.63
†RA activity (patient assessed)	−0.57	−0.52
HAD-mood	−0.56	−0.59
†Joint score-tender	−0.55	−0.52
Duration of RA	−0.45	−0.33
†Disease activity (doctor)	−0.43	−0.47
†Joint score-swollen	−0.43	−0.45
XR erosions (present/absent)	0.42	0.32
†ESR	−0.39	−0.29
Years of education	0.33	0.28
Co-morbidity (present/absent)	−0.28	−0.28
Age	−0.29	−0.17*
Rh factor (present/absent)	0.17*	0.12*
Drug side-effects (present/absent)	−0.16*	−0.23**

*$P > 0.01$; **$P = 0.001$; all others $P < 0.000$.
†ACR disease activity set.

disease activity and years of education) also entered the regression equation for the $EQ\text{-}5D_{vas}$ score, showing that the model is sensitive to other factors.

The β coefficients for HAQ, HAD-mood and pain-VA were generally consistent between baseline and 3 month assessments in the regression equations for the $EQ\text{-}5D_{utility}$ and $EQ\text{-}5D_{vas}$ values, respectively, again confirming the predictive value of these three variables. The R^2 for none of the models was improved by more than 1% when all variables were included in the regression equations.

TABLE VI
Stepwise regression models for $EQ\text{-}5D_{utility}$ and $EQ\text{-}5D_{vas}$ *vs* ACR disease activity measures and other medical and demographic factors

(a) $EQ\text{-}5D_{utility}$

Variable	Baseline ($R^2 = 67\%$) β coefficient	3 month review ($R^2 = 74\%$) β coefficient
HAQ score	−0.188***	−0.157***
HAD-mood	−0.008**	−0.008***
Pain-VA scale	−0.003**	−0.003**
Disease activity (patient)	−0.068*	−0.100***
ESR	ns	−0.001*
Constant	1.12***	1.20***

(b) $EQ\text{-}5D_{vas}$

Variable	Baseline ($R^2 = 65\%$) β coefficient	3 month review ($R^2 = 67\%$) β coefficient
HAQ score	−8.98***	−9.23***
HAD-mood	−0.722***	−0.41*
Pain-VA scale	−0.17**	−0.38***
Side-effects	−5.92*	ns
Disease activity (patient)	−4.26*	ns
Years of education	0.814*	ns
Constant	95***	95***

***$P < 0.001$; **$P < 0.01$; *$P < 0.05$; ns = not significant.

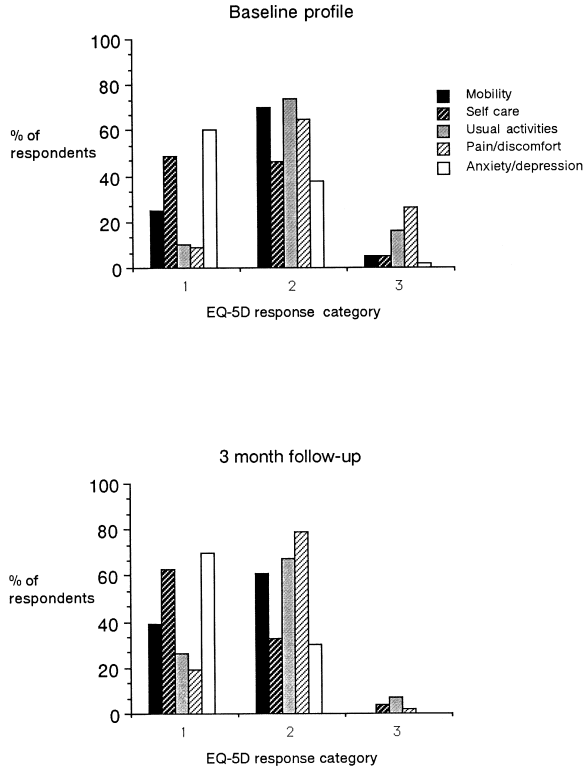

FIG. 2.—Change in EQ-5D profile for patients reporting improvement in activity of RA. The percentage of patients reporting some or extreme problems is lower with a corresponding increase in the percentage of patients reporting no problems in each domain at 3 months follow-up in patients self-reporting improvement in RA ($n = 56$).

Sensitivity to change

Fifty-seven patients reported improvement, 73 deterioration and 93 no change in their arthritis over 3 months.

Change in $EQ\text{-}5D_{profile}$. The change in unweighted score for each EQ-5D domain, except the anxiety/depression domain, was significantly related to category of self-reported change in RA (same, better, worse) over 3 months (Kruskal–Wallis test; mobility $P < 0.001$; self-care $P < 0.05$; usual activities $P < 0.01$; pain/discomfort $P < 0.001$; anxiety/depression $P = 0.4$). For illustration, the change in profile for patients reporting improvement over 3 months is shown (Fig. 2). The percentage of patients reporting extreme problems declines with a corresponding change in the percentage of patients reporting no or some problems in each domain.

Change scores and standardized response means for $EQ\text{-}5D_{utility}$ and $EQ\text{-}5D_{vas}$ (Table VII). All instruments recorded improvement in patients self-reporting improvement. Although all instrument scores declined in patients reporting worsening, the magnitude of change in this group was smaller and in some instances—$EQ\text{-}5D_{utility}$, joint swelling, ESR and HAD-mood—statistically insignificant (not shown). Inspection of

TABLE VII
Mean and 95% CI for change scores (0–3 months) and standardized response means (SRM) in patients reporting improvement over 3 months

	N	Mean change	95% CI for change	SRM	95% CI for SRM	SRM*
Disease activity-doctor	46	+0.7	0.5, 0.9	1.0	0.71, 1.29	1.0
Pain-VA scale	56	+22	15, 29	0.85	0.58, 1.12	1.10
EQ-5D$_{vas}$	56	+12.4	7.9, 16.8	0.71	0.45, 0.96	1.0
EQ-5D$_{utility}$	56	+0.22	0.13, 0.30	0.70	0.41, 0.96	1.0
HAD-mood	55	+2.6	1.5, 3.7	0.65	0.38, 0.93	0.62
Joint swelling	56	+2.1	1.2, 3.0	0.64	0.37, 0.92	0.70
Joint tender	54	+4.1	2.2, 6.0	0.59	0.32, 0.86	0.68
Disease activity (patient)	56	+0.5	0.3, 0.8	0.5	0.3, 0.8	0.71
ESR	49	+5.5	0.4, 10.6	0.31	0.04, 0.60	0.32
HAQ	53	+0.12	0.04, 0.20	0.40	0.13, 0.67	0.41

standardized response means, SRM and SRM* reveals that the SRM* (calculated using variance estimates in patients reporting no change) generally gives the highest values. However, regardless of which method is used, HAQ score and ESR appear relatively unresponsive compared to EQ-5D, pain-VA scale, joint scores and disease activity scores.

Regression analysis of change scores for EQ-5D$_{utility}$ and EQ-5D$_{vas}$ over 3 months. Change scores for EQ-5D$_{utility}$ or EQ-5D$_{vas}$ are significantly correlated with change in each of the condition-specific measures ($P = 0.01$ or greater) except ESR (not shown). Linear forward stepwise regression showed that change in HAQ, HAD-mood, pain-VA, patient-assessed disease activity and self-reported side-effects accounted for 42% of the variance in change in EQ-5D$_{utility}$ (Table VIII). If all variables were included in the equation, the R^2 increased to 48%. These results are consistent with the earlier finding (Table VIa) that pain, function and mood were strong predictors of EQ-5D$_{utility}$ at baseline and 3 months.

Change in HAD-mood, pain-VA, patient- and doctor-assessed disease activity, and self-reported side-effects predicted 48% of the variance in change in EQ-5D$_{vas}$ score (Table VIII). If all variables were included in the model, the R^2 increased to 54%. It should be noted that change in HAQ score did not enter the regression equation, but otherwise the result is broadly in agreement with the regression analysis performed at baseline and 3 months (Table VIb).

TABLE VIII
Linear stepwise regression model for change (δ) in EQ-5D$_{utility}$ and EQ-5D$_{vas}$ between baseline and 3 months *vs* change in ACR disease activity measures and other clinical and demographic factors

Variable	δ EQ-5D$_{utility}$ ($R^2 = 42\%$) β coefficient	δ EQ-5D$_{vas}$ ($R^2 = 48\%$) β coefficient
δ HAQ score	0.165*	ns
δ HAD-mood	0.0127*	0.62*
δ Pain-VA scale	0.0020*	0.21***
δ Disease activity (patient)	0.096**	4.32*
δ Side-effects	−0.090*	−5.72*
δ Disease activity (doctor)	ns	3.33*
Constant	ns	ns

***$P < 0.001$; **$P < 0.01$; *$P < 0.05$; ns = not significant.

Independent variables tested in stepwise regression (0.05 limits) were: change (δ) in ACR disease activity measures plus δ HAD-mood, age, duration of RA, years of full-time education. Co-morbidity and side-effects were coded as 1 = absent and 2 = present at baseline and 3 months; δ side-effects or δ co-morbidity are thus 0 = no change; −1 = new problem reported; +1 = problem no longer reported.

Reliability

The EQ-5D$_{profile}$ for patients reporting 'no change in RA' showed no significant change in any of the five domains (Wilcoxon test, $P > 0.2$).

In patients reporting no change, the 95% CI for mean change in all instruments span zero except for the joint swelling score, HAD-mood and doctor-assessed disease activity, each of which improved significantly from baseline (Table IX). The 95% CI for individual change scores are wide.

The reliability coefficients (ICC) and Goodman and Kruskal's gamma for each instrument are shown (Table IX). Over a 3 month test period, the HAQ is clearly the most reliable instrument, but the EQ-5D$_{utility}$ and EQ-5D$_{vas}$ demonstrate greater reliability than several of the condition-specific instruments. In the 31 patients asked to complete another set of questionnaires over a shorter period of 2 weeks, the ICCs for EQ-5D$_{vas}$ and EQ-5D$_{utility}$ increased slightly (Table IX), but their relative reliability remained unchanged. Reliability assessed using non-parametric tests of concordance (Goodman and Kruskal's gamma) gave very similar results except that the relative reliability of the Likert scale for both patient- and doctor-assessed disease activity was improved compared to other instruments.

DISCUSSION

There is no universally accepted definition or method of measuring HR-QOL [16]. Measurement of 'health' is problematic, not least because the boundaries between health and disease are poorly defined. Perceptions of health and responses to disease are often profoundly affected by individual beliefs and attitudes, as well as by social and economic incentives and pressures. There are also widely differing cultural, ethnic and religious attitudes to the concept of health. Calman [17] has defined quality of life as 'the extent to which an individual's hopes and ambitions are matched

TABLE IX
Mean change scores, reliability coefficients (ICC) and Goodman and Kruskal's gamma† in patients reporting no change in RA over 3 months or over 2 weeks*

	N	Mean change	95% CI for individual	ICC	95% CI for ICC	Gamma†
HAQ	88	+0.04	−0.51, 0.59	0.94	(0.84–1.04)	0.88
	31*	−0.05	−0.60, 0.55	0.92	(0.74–1.1)	0.83
EQ-5D$_{vas}$	91	+2.6	−26.8, 32.0	0.70	(0.60–0.80)	0.57
	31*	−3.1	−29, 23	0.85	(0.67–1.03)	0.71
EQ-5D$_{utility}$	93	+0.02	−0.43, 0.47	0.73	(0.63–0.83)	0.69
	31*	−0.02	−0.44, 0.41	0.78	(0.60–0.96)	0.80
VA-pain scale	91	+0.68	−38, 39	0.75	(0.65–0.85)	0.51
	31*	+1.26	−35, 37	0.75	(0.57–0.93)	0.64
Tender joint score	85	+0.75	−10.8, 12.3	0.78	(0.68–0.88)	0.67
Disease activity (patient)	93	+0.05	−1.24, 1.34	0.61	(0.51–0.71)	0.78
Swollen joint score	88	+1.23	−4.63, 7.09	0.56	(0.46–0.66)	0.52
Disease activity (physician)	80	+0.18	−1.26, 1.61	0.65	(0.55–0.75)	0.78

and fulfilled by experience'. The value of this definition is that it highlights the idea that self-perceptions of HR-QOL may represent the gap between an individual's reality and their expectations in those aspects of their life affected by their health. Most clinicians are familiar with the paradox of the patient who is disproportionally disabled and handicapped by a relatively minor medical problem while another patient with objective evidence of severe disability perceives their HR-QOL to be good. Such patients may have adjusted their expectations over time, narrowing the gap between expectations and reality. HR-QOL may, therefore, be regarded as the resultant of a complex interaction between mental attitude, social adjustment and disease.

The development and origins of the item content of EQ-5D have been described [18] and our study provides good empirical evidence that the unweighted EQ-5D domains cover dimensions of health which are regarded as relevant to patients with arthritis. This was demonstrated by a highly significant relationship between unweighted patient responses on three of the EQ-5D domains and their scores on relevant condition-specific measures. In common with many other generic instruments, the EQ-5D domains cover different levels of impact of disease on the individual, i.e. impairment, disability and handicap. It has been argued that inclusion of different levels of disease impact in a single instrument creates difficulty in determining what such instruments are measuring [16, 19]; for example, disability has a much closer relationship to disease impairment than handicap, while handicap may be considered closer to, if not synonymous with, HR-QOL. Thus, although from the patient's perspective it is the extent to which they are disadvantaged in fulfilling their normal roles, i.e. their degree of handicap, that is of greatest importance, levels of impairment and disability may act as proxy indicators of handicap. Any attempt to capture overall HR-QOL in a single index may therefore incorporate descriptors of impairment and disability, so long as their impact on HR-QOL (or the consequent handicap for the individual) is assessed through some kind of subjective valuation procedure, as is the case for the EQ-5D.

In this paper, we have analysed the performance of EQ-5D in terms of its validity, responsiveness and reliability. If EQ-5D is a valid measure of HR-QOL, one would expect the values elicited to be modestly correlated with measures of impairment, e.g. the ESR or joint score, but more highly correlated with patients' subjective perceptions of their disabilities, for example with the HAD-mood and HAQ scores. This was found to be the case for both EQ-5D$_{utility}$ and EQ-5D$_{vas}$, with slightly higher correlations observed for EQ-5D$_{utility}$. The stepwise regression models provided further confirmatory evidence of construct validity. For example, the variables retained in the models for both EQ-5D$_{utility}$ and EQ-5D$_{vas}$, i.e. physical function, pain, anxiety/depression and patient-assessed disease activity, reflect those aspects of health one would expect to have significant impact on quality of life in patients with RA. The regression model for EQ-5D$_{vas}$ was less stable than the model for EQ-5D$_{utility}$, and patient-assessed disease activity, side-effects and educational level, which entered the model at baseline, were not significant predictors of EQ-5D$_{vas}$ at 3 months.

EQ-5D$_{utility}$ shows the predicted relationships with functional class and socioeconomic status, higher values being associated with higher functional class, employment, higher socioeconomic status and greater independence. It should be noted that some patients with more severe disease attracted utility values below zero, i.e. from a societal perspective they had a health state regarded as worse than death. This, of course, cannot be interpreted either to mean that such patients wish to die or that the societal perspective is that such patients be allowed to die, it merely represents the fact that normal individuals asked to consider existing in such health states would regard themselves as better off dead. Nonetheless, a small number of severely disabled patients did volunteer profoundly pessimistic views of their own health state. The problems associated with derivation of health utilities is discussed in detail by Drummond *et al*. [20]. As discussed below, the self-rated health of patients on the EQ-5D$_{vas}$ scale

diverges from the societal view in severely disabled patients, and raises important ethical and practical questions regarding the use and interpretation of utility values.

Higher self-rating scores on the EQ-5D$_{vas}$ were associated with living independently or being employed, but in contrast to EQ-5D$_{utility}$, the scores did not distinguish between those in functional classes 3 and 4, those living with a spouse rather than a 'carer' or those living in different types of accommodation. These data suggest that the EQ-5D$_{vas}$ detects a more optimistic self-valuation of health in those with more advanced disease than that given by external observers, a phenomenon well recognized by clinicians. The mechanism of this effect is not clear; denial or adjustment to chronic disease is one possible explanation and it has been previously noted that health state valuations differ according to experience of illness. This again raises important questions, such as whether the patient's or society's valuation of health should be used [20, 21]. An alternative explanation is that the two instruments are measuring different aspects of health status or HR-QOL. The tariff for EQ-5D$_{utility}$ is derived, using TTO methodology [22], from third person valuations of 'theoretical' health states using individuals who may have no experience of ill-health. When patients evaluate their own health on the EQ-5D$_{vas}$ scale, it cannot be assumed that they are evaluating health in the same way as normal individuals or over the same time frame. They may, therefore, have quite different perceptions of severity. Whatever the explanation for the difference between EQ-5D$_{utility}$ and EQ-5D$_{vas}$ scores, the discrepancy requires further study since it has implications for the application of EQ-5D$_{utility}$ valuations in cost–utility studies and resource allocation.

The ability to detect clinically important change is an essential requirement of any instrument purporting to measure health outcomes. Firstly, we have shown that change in the unweighted score for each domain except anxiety/depression on the EQ-5D$_{profile}$ is significantly related to the category of self-reported change in RA, i.e. better, same or worse. Condition-specific measures, because of their narrower focus, are often considered to be more responsive to clinical change; however, in our study EQ-5D$_{utility}$ and EQ-5D$_{vas}$ were found to be more responsive, as measured by the SRM, than most of the condition-specific measures. Regression models confirmed that change in EQ-5D$_{utility}$ was predicted by change in disability, mood, pain, patient-assessed disease activity and self-reported drug side-effects, providing further evidence that EQ-5D is measuring clinically relevant change. Because the gold standard for improvement was self-reported change in RA, the comparisons we have made may not represent a full test of the relative sensitivity of measures of disease process such as the ESR, which in this context changed very little. However, both EQ-5D$_{utility}$ and EQ-5D$_{vas}$ perform well, and it can be concluded that these instruments are highly responsive to self-reported improvement in RA and that this reflects clinically important changes. It will be important to confirm this finding in drug intervention studies, e.g. using second-line therapy, where an attributable improvement in health would be anticipated.

Reliability was tested by examining the stability of instrument scores in patients reporting 'no change' in their condition over 3 months and, in a smaller group of subjects, over 2 weeks. A 3 month period was chosen to provide a very conservative test and to give a useful indication of performance under conditions comparable to those in routine clinical practice where measurement intervals may be as long as 3 or 4 months. There are no absolute standards of reliability, but as a guide it is only appropriate to use change scores to assess main effects with an instrument if the variance between subjects exceeds the error variance of measurement, i.e. the reliability of the instrument exceeds 0.5 [10]. The trend to improvement observed in patients reporting no change over 3 months also highlights the importance of considering instrument reliability and stability of controls when using change scores rather than *t*-tests to estimate main effects [10]. Two suggested standards of reliability for tests used to make decisions about individuals are coefficients of 0.94 or 0.85 [10]. However, it must be remembered that any recommendations for standards of reliability are arbitrary and context specific since large sample sizes will be more tolerant of unreliability than smaller samples. Over either the 3 month or 2 week interval, the HAQ was very reliable (ICC = 0.94 or 0.92). EQ-5D also performed moderately well in comparison to the other instruments over 3 months, with ICCs of 0.70 for the EQ-5D$_{vas}$ and 0.73 for EQ-5D$_{utility}$. When tested over 2 weeks, the reliability of the EQ-5D$_{utility}$ and EQ-5D$_{vas}$ improved slightly, showing that test–retest over a 3 month period may slightly underestimate instrument reliability. Examination of the 95% CI for individual change scores shows that the interpretation of scores from any of these instruments may be difficult in individual patients. Very similar results were obtained when non-parametric measures of concordance were used. The only obvious difference is that the relative reliability of the Likert scales for patient- or doctor-assessed disease activity were improved using non-parametric methods, but in general EQ-5D was at least as reliable as standard ACR measures of disease activity.

The data we report here confirm that EQ-5D has construct validity in RA, is at least as responsive to self-reported clinical change and as reliable as many of the condition-specific instruments used in RA. The EQ-5D$_{vas}$ is reliable and clearly useful for measuring changes in perceived health. In addition, the EQ-5D$_{profile}$ may be used as a simple health profile illustrating in which areas a patient or group of patients is reporting problems and where changes have occurred over time. While further work is required to explore the scaling of EQ-5D$_{utility}$, and in particular the valuation of severe health states in relation to death, the EQ-5D would appear suitable for use as a simple generic instrument for measuring net changes in overall health alongside

condition-specific instruments, and may be of particular value in studies of cost-utility and cost-effectiveness [23]. Additional studies to examine the responsiveness of EQ-5D under conditions where attributable change occurs, e.g. after drug intervention, would be very useful. EQ-5D is simple to use and it would be feasible to use EQ-5D in a routine clinical setting alongside a measure of disability such as HAQ, for example for purposes of audit. An acid test of the clinical value of EQ-5D is whether the instrument, presented either as a simple profile or as a summary utility score or as the visual analogue scale, would in fact influence clinical decisions in a routine setting.

Acknowledgements

This work was supported by a grant from the Chief Scientists Office of the Scottish Home & Health Department. The authors are also very grateful to all the patients who willingly gave of their time to complete the various assessments.

References

1. Lambert CM, Hurst NP. Health economics as an aspect of health outcome: basic principles and application in rheumatoid arthritis. Br J Rheumatol 1995;34:774–80.
2. Robinson R. The policy context. Br Med J 1993;307:994–6.
3. Garratt AM, Ruta D, Abdalla MI *et al*. The SF36 health survey instrument: an outcome suitable for routine use within the NHS. Br Med J 1993;306:1440–4.
4. The EuroQol group. EuroQol—a new facility for the measurement of health related quality of life. Health Policy 1990;16:199–208.
5. Hurst NP, Jobanputra P, Hunter M, Lambert CM, Lochead A, Brown H. Validity of EuroQol—a generic health status instrument in patients with rheumatoid arthritis. Br J Rheumatol 1994;33:655–62.
6. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. Discussion paper 138. York: University of York, 1995.
7. Steinbrocker O, Traeger CH, Batterman RC. Therapeutic criteria in rheumatoid arthritis. J Am Med Assoc 1949;140:659–62.
8. Guyatt G, Walters S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171–8.
9. Katz JN, Larson MG, Phillips CB *et al*. Comparative measurement sensitivity of short and longer form health status instruments. Med Care 1992;30:917–25.
10. Streiner DL, Norman GR. Health measurement scales—A practical guide to their development and use. Oxford: Oxford University Press, 1994.
11. Arnett FC, Edworthy SM, Bloch DA *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
12. Felson DT, Anderson JJ, Boers M *et al*. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;6:729–40.
13. Egger MJ, Huth DA, Ward JR *et al*. Reduced joint count indices in the evaluation of rheumatoid arthritis. Arthritis Rheum 1985;28:613–9.
14. Pincus T, Summey JA, Sorraci SA *et al*. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
15. Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scand 1983;67:361–70.
16. Carr AJ, Thomson PW, Kirwan JR. Outcome series. Quality of life measures. Br J Rheumatol 1996;35:275–81.
17. Calman KC. Quality of life in cancer patients—an hypothesis. J Med Ethics 1984;10:124–7.
18. Williams A. The measurement and valuation of health: A chronicle. Discussion paper 136. York: University of York, 1995.
19. Fitzpatrick R, Badley EM. An overview of disability. Br J Rheumatol 1996;35:184–7.
20. Drummond MF, Stoddart GL, Torrance GW, eds. Cost-utility analysis. In: Methods for the evaluation of health care programmes. Oxford: Oxford University Press, 1994, 112–48.
21. Kind P, Dolan P. The effect of past and present illness experience on the valuations of health states. Med Care 1995;33:AS255–63.
22. Torrance G, Thomas WH, Sackett DL. A utility maximisation model for evaluation of health care programs. Health Serv Res 1972;7:118–33.
23. Robinson R. Cost utility analysis. Br Med J 1993;307:859–62.