Scale Characteristics and Mapping Accuracy of the US EQ-5D, UK EQ-5D, and SF-6D in Patients with Rheumatoid Arthritis

FREDERICK WOLFE, KALEB MICHAUD, and GENE WALLENSTEIN

ABSTRACT. ***Objective.*** To compare the US EQ-5D with the UK EQ-5D and the SF-6D in patients with rheumatoid arthritis (RA). To provide mappings for each of the scales based on clinical variables.

Methods. We studied 12,424 patients with RA with 66,958 longitudinal observations using linear regression. In our mapping models we used the Health Assessment Questionnaire (HAQ) as a continuous predictor variable and as individual items. More complex models included the addition of a visual analog pain scale, the mood scale from the SF-36, and demographic and comorbidity covariates. We compared various models using root mean squared error (RMSE), in-sample and out-of-sample mean absolute error (MAE), and other measures of prediction accuracy and model fit.

Results. At any level of clinical severity, the US EQ-5D always had a higher utility score than the UK EQ-5D; and overall, the US scores were 0.094 units higher. The best models explained 64% to 72% of variance in utility scores, with RMSE values of 0.07 (SF-6D), 0.11 (EQ-5D US), and 0.17 (UK EQ-5D). There was a substantial increase in predictive accuracy by using pain and mood as predictor variables in the mapping.

Conclusion. The US EQ-5D differs from the UK version and from the SF-6D in mean scores and ranges. When determined by mapping, the US EQ-5D has a much lower prediction error than the UK EQ-5D. Simple mapping models that use HAQ and pain have acceptable error rates, although more complex models that include mood scores and individual HAQ items substantially improve predictive accuracy. (J Rheumatol First Release June 15 2010; doi:10.3899/jrheum.100043)

Key Indexing Terms:
EQ-5D UTILITIES
HEALTH ASSESSMENT QUESTIONNAIRE
SF-6D UTILITIES
CONVERSION

Recently, Bansback, *et al* proposed a method to use the Health Assessment Questionnaire disability index (HAQ) to estimate preference-based single measures of health status or utilities[1]. Almost all current assessments of utilities in rheumatology studies rely on measures that include either the EQ-5D, the Short Form-6D (SF-6D), or the Health Utilities Index II or III (HUI-II, HUI-III)[2,3]. Given the existence of one of these measures, the results of clinical trials can be transformed into quality-adjusted life-years (QALY) gained or lost as a result of the intervention, and this, in turn, can be expressed in cost-utility analyses as cost per QALY. One QALY is the equivalent of one extra year of life lived in perfect health over a specified number of years. The cost per QALY for rheumatoid arthritis (RA) biologic therapy ranges from US $40,000 to US $68,000[4,5,6,7]. Utilities and QALY allow comparison between treatments in the same disease, for example a comparison of 2 biologics, as well as different treatments across illnesses, thus allowing healthcare economists and regulatory authorities to understand the comparative costs and benefits using a single standard.

Investigators have been most interested in using the EQ-5D because of the relative restricted range of the SF-6D and its apparently reduced responsiveness, although that finding has recently been called into some question[8,9]. The length and difficulty of administering and scoring the HUI is also somewhat limiting for that questionnaire. The EQ-5D is a 5-item questionnaire that has 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. However, 3 of the dimensions fall within the domain of function. Each of the 5 questions has 3 levels, with 1 denoting no problems and 3 indicating extreme problems[10]. The number of theoretically possible health states is 243 (3^5). The EQ-5D is commonly reported as a pre-

From the National Data Bank for Rheumatic Diseases, and University of Kansas School of Medicine, Wichita, Kansas; University of Nebraska Medical Center, Omaha, Nebraska; and Pfizer Global Research and Development, New London, Connecticut, USA.

Supported by a grant from Pfizer, Inc.

F. Wolfe, MD, National Data Bank for Rheumatic Diseases, University of Kansas School of Medicine; K. Michaud, PhD, University of Nebraska Medical Center, National Data Bank for Rheumatic Diseases; G. Wallenstein, PhD, Pfizer Global Research and Development.

Address correspondence to Dr. F. Wolfe, National Data Bank for Rheumatic Diseases, 1035 N. Emporia, Suite 288, Wichita, KS 67214. E-mail: fwolfe@arthritis-research.org

Accepted for publication March 5, 2010.

ference-based single index number that is derived from the answers to the 5 EQ-5D questions. This index number is obtained by applying algorithms that link the responses with average valuations obtained from the general public. The predominance of functional items in the EQ-5D suggested the possibility that EQ-5D scores could be estimated with sufficient accuracy for use in cost-utility analyses. This HAQ to EQ-5D mapping has been used to estimate cost-utility analyses in multiple studies, and its basis has been analyzed and explored in detail by Bansback, *et al*[1].

There are a number of practical problems with the use of utilities. Most importantly, Marra, *et al* have shown in 313 patients with RA that the agreement among 4 different utility scales was poor, and if applied to cost-utility analyses would yield quite different cost/QALY results[11]. A second issue that concerns the EQ-5D is that all RA studies that utilize that questionnaire have been based on the scoring algorithm derived from UK weightings, including Canadian and UK studies. But Johnson, *et al* reported that the average difference in valuations between US and UK EQ-5D was 0.10, with higher scores being found in the US EQ-5D[12]. They also reported that "the magnitude of the difference in the US and UK valuations was not constant across EQ-5D health states; greater differences in valuations were present in health states characterized by extreme problems." Their recommendation that "EQ-5D index scores generated using valuations from the US general population should be used for studies aiming to reflect health state preferences of the US general public" would create problems in interpreting multinational studies and in the comparison of results of observational studies that used the different valuations.

While the Bansback study[1] developed predictive models for the UK EQ-5D, they did not address the US valuations. In addition, models that predict utility scores only from the HAQ cannot adequately address the contribution of pain and mood. Very low utility scores and states "worse than death" derive from the contribution of pain and mood[13].

In this report, we provide data that compare valuations of the US EQ-5D, UK EQ-5D, and SF-6D scales at all levels of the HAQ, as well as at important levels of RA outcomes. In addition we describe the differences between the scales at different levels of RA and HAQ severity. Based on a sample size of 12,098 patients with 63,406 observations, we provide a series of maps via regression algorithms that convert HAQ, and HAQ, pain and mood scores, to US and UK EQ-5D and SF-6D results.

MATERIALS AND METHODS

Patients. We used the National Data Bank for Rheumatic Diseases (NDB) longitudinal study of RA outcomes[14,15] to evaluate utility scores, mapping predictors, and the association of clinical outcomes with utility scores. Patients in this study were diagnosed and referred to the NDB by US rheumatologists. They received no compensation for participation. Patients who were referred to the NDB to be participants of drug safety registries were excluded from analysis because they might have been selected because of the severity of their RA. At 6-month intervals, patients completed complex survey questionnaires by mail or by the Internet. Administration of the SF-6D and the EQ-5D began simultaneously in the NDB assessment of July 2002. The ending date for the current report was the questionnaire of January 2009.

Utilities measures. The EQ-5D, described in some detail above, is a 5-item questionnaire that has 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. From the 243 possible health states derived from 5 questions and 3 levels, a single index number for each state is obtained based on valuations obtained from persons in the general population. In the UK these valuations were based on a population survey of 3,995 persons in the UK using 10-year time tradeoffs, and published in 1995[10]. The valuations were used widely in the US, Canada, and the UK, and are the valuations used in clinical trials[4,5,6,7]. We refer to this version of the EQ-5D as the UK EQ-5D. The worst UK EQ-5D score observed in the current study was –0.59.

In 2005, US valuations of the EQ-5D health states first became available for 42 common health states based on time tradeoff, and then expanded to all 243 states by regression analysis[12,16]. EQ-5D US scores are known to be lower than UK scores[12]. However, the US EQ-5D has not been studied previously in RA. The lowest US EQ-5D score observed in the current study was –0.11.

The third utility measure studied in this report was the SF-6D[17]. First reported in 2002, it utilizes 11 questions from the SF-36[18] to create 6 domains and 249 health states. Valuations for the health states were obtained from 836 persons in the UK general population using the standard gamble method. The worst score observed in the current study was 0.34.

Covariates and predictor variables. We used a series of predictor variables to estimate EQ-5D values (Table 1). The HAQ disability index is a widely used measure of functional status across rheumatic diseases[19,20]. The HAQ consists of 20 items (scored 0–3) in 8 different domains. Each domain contains 2 to 3 questions based on a common theme: dressing, standing, eating, walking, toileting, reach, grip, and instrumental activities. A score is derived for each domain based on the most abnormal item in that domain. In addition, the use of 14 aids and devices to help with function is taken into consideration in the item scoring by increasing all item scores at the level of "with no difficulty [0]" or "with some difficulty [1]" to "with much difficulty or with help [2]" if an aid or device is used with the item. The final HAQ score ranges from 0 to 3 and is the average score of the 8 categories.

The HAQ-II is a reliable and valid 10-item questionnaire that provides scores in the range 0–3, performs at least as well as the HAQ, and is simpler to administer and score[21]. Psychometrically improved, it has a reduced floor effect, and scores that are very similar to those of the HAQ, thus allowing comparison of group data using the HAQ and HAQ-II[22]. The HAQ-II can be substituted for the HAQ in clinical care[22]. Because the HAQ-II has not been used widely in clinical trials, and details of its performance would double the length of this report, we do not give HAQ-II results here, except as a brief summary at the end of the Results section.

We assessed pain using a 21-point 0–10 visual analog scale (VAS) in which higher values indicated more pain. Mood was assessed using the mood (mental health) scales of the SF-36[23]. Comorbidity was measured by a patient-reported composite comorbidity index (range 0–9) consisting of 11 present or past comorbid conditions including pulmonary disorders, myocardial infarction, other cardiovascular disorders, stroke, hypertension, diabetes, spine/hip/leg fracture, depression, gastrointestinal (GI) ulcer, other GI disorders, and cancer[24,25].

The levels of formal education were categorized as 0–8, 9–11, 12, 12–15, and ≥ 16 years. Based on preliminary evaluations, we dichotomized RA duration as < 8 years and ≥ 8 years.

Validation: outcome variables. To characterize the "clinical significance" of differences in utility scores across the 3 utility measures, we compared utility scores at levels of the following clinical measures. Self-rated current health was obtained with a question from the SF-36 questionnaire: "In general, would you say your health is: Excellent, Very Good, Good, Fair,

Table 1. Characteristics of 12,424 patients with rheumatoid arthritis.

Variable	Mean (SD) or %
Age, yrs	61.2 (13.0)
Sex, % male	21.4
White, not of Hispanic origin, %	93.6
Education, yrs, %	
0–8	1.4
8–11	4.6
12	34.3
13–15	26.9
16	32.7
Comorbidity index, %	
No comorbid conditions	24.0
Score 1	28.7
Score 2	21.6
Score 3	13.9
Score 4–9	7.0
Disease duration, median yrs (IQR)	12.8 (12.8, 21.7)
EQ-5D Walking, %	
Have no problems walking about	52.4
Have some problems walking about	47.4
I am confined to bed	0.3
EQ-5D Self-care, %	
Have no problems with self-care	76.9
Have some problems with self-care	22.7
I am unable to wash/dress myself	0.4
EQ-5D Performing usual activities, %	
Have no problems performing usual activities	494
Have some problems performing usual activities	48.2
I am unable to perform usual activities	2.3
EQ-5D Pain, %	
I have no pain or discomfort	17.1
I have moderate pain or discomfort	73.8
I have extreme pain or discomfort	9.1
EQ-5D Anxiety-depression, %	
I am not anxious or depressed	67.1
I am moderately anxious or depressed	30.5
I am extremely anxious or depressed	2.3
HAQ disability index (0–3)	1.03 (0.73)
HAQ-II disability index (0–3)	1.00 (0.67)
SF-6D (0–1)	0.69 (0.13)
US EQ-5D (0–1)	0.73 (0.19)
UK EQ-5D (0–1)	0.64 (0.28)
Mood [SF-36 mental health] (0–100)	72.1 (19.4)
VAS pain (0–10)	3.8 (2.7)

HAQ: Health Assessment Questionnaire disability scale; HAQ-II: HAQ disability scale II; SF-6D: Short-form 6D; US EQ-5D: United States EQ-5D; UK EQ-5D: United Kingdom EQ-5D; IQR: interquartile range.

Poor." Disability status (able to work) was determined by self-report. This measure is a valid, broader measure than an assessment of receipt of work disability pension because it also assesses disability in nonworkers, particularly homemakers and those past the retirement age[26]. Total joint replacement (Yes or No) measures the influence of chronic RA severity and activity, as joint replacement is the end product of RA activity. Total direct medical costs, adjusted to 2007, were calculated from hospitalization, treatment, and utilization data as described[27]. Comorbidity: We used self-reported comorbidities to compute a composite comorbidity index (range 0–9) comprising 11 present or past comorbid conditions including pulmonary disorders, myocardial infarction, other cardiovascular disorders, stroke, hypertension, diabetes, spine/hip/leg fracture, depression, GI ulcer, other GI disorders, and cancer[24,25,28]. Widespread pain index (WPI): In this index patients indicate in which of 19 body areas they had pain during the last week. These areas were those previously described as part of the Regional Pain Scale, now renamed the Widespread Pain Index (WPI)[29]. The WPI is a measure of the degree of widespread pain, and is strongly correlated with poor health status. Fibromyalgia, as measured by survey fibromyalgia criteria[30], is associated with very poor health status.

Predictive models of US EQ-5D, UK EQ-5D, and SF-6D. To predict utilities from surrogate measures, we used linear regression techniques for analysis of each of the utility scales. We also performed analyses using the HAQ-II instead of the HAQ. A central issue for the study analyses was which functional form of the HAQ or HAQ-related variables was best. Although we modeled the HAQ using fractional polynomial regression in preliminary analyses, a fractional polynomial functional form was not an improvement over other forms, and we did not include fractional polynomial regression in the output tables. We used the HAQ as a single continuous variable (HAQ score) and as a categorical variable of 25 categories (0, .125, .25, .375, etc.). However, the categorical form did not perform better than the continuous form, and we elected to use only the continuous form in followup analyses of Table 4. Bansback, *et al* found that treating 8 HAQ domains as categorical variables provided a useful model[1], and we used categorical HAQ domains as one of our functional forms in initial analyses. Finally, we also used the 20 categorical HAQ items, as did Bansback[1]. In the 20 categorical HAQ items analyses we incorporated the contribution of HAQ aids and devices sections into the item scores and the domain scores, and did not analyze them separately. Domains were not used in HAQ-II analyses, as the HAQ-II does not create domains.

In additional analyses we added covariates. We used the 21-step (0–10) VAS pain scale as a continuous scale, and similarly employed the mood scale as a continuous variable. We considered these variables as primary covariates, as the EQ-5D questionnaire has one item for pain and an additional item for mood. We added the comorbidity index as a categorical variable because we believed that the effect of comorbidity on utility scores might offer information that might not be picked up by the HAQ and other covariates. In addition, we adjusted for age, age-squared, sex, RA duration, and education level. Finally, in many preliminary models we incorporated interaction terms between sex and HAQ and sex and duration. These were not included in final analyses because their effect was mostly nonsignificant, complicated model use, and did not add to overall prediction accuracy.

Model selection. We evaluated each model statistically and graphically. In particular we used quantile-quantile plots to evaluate how well the predicted utility followed the observed utility. In our primary analyses we utilized one randomly selected observation from each of the 12,424 patients. However, while there were 12,424 HAQ and utility scores, there were only 10,895 patients who completed all the 20 HAQ items. Therefore, so that all models would use the same sample, we restricted analyses to the 10,895 patients with complete data.

To test out-of-sample error and to evaluate changes over time, we used 8,669 observations for each patient, obtained 6 months after the primary observation. Only 8,669 observations were used because not all patients had 2 consecutive observations within the 6-month window.

As we suspected that many models might be useful clinically, it was not our goal to select the best model. Instead, we described each model in terms of its predictive accuracy and fit. For predictive accuracy at the group level, we used the root mean squared error (RMSE) and the mean absolute error (MAE), and at the patient's level we used the Bland-Altman limits of agreement (LOA) statistic[31] and the correlation between observed and predicted values. The RMSE, also known as the standard error of the estimate (SEE), is the square root of the average squared prediction error. The RMSE favors prediction models that do not produce particularly large errors[1]. The MAE represents the average difference between the actual and predicted utility scores. "The RMSE attaches greater weight to larger errors and favors prediction models that do not produce particularly large errors at the expense of models that are off by a modest amount."[32] We used the MAE to deter-

mine "in-sample" and "out-of-sample" errors. Lower error scores indicate better prediction models. RMSE and MAE should be used in analyses of individual measures (e.g., US EQ-5D) and not used to compare different measures (e.g., US EQ-5D vs UK EQ-5D vs SF-6D). To evaluate model fit, we determined the adjusted R-square, Akaike information criterion (AIC), and Bayesian information criterion (BIC). Higher values indicate better fit for the R-square, and lower values a better fit for the AIC and BIC. Data were analyzed using Stata version 11.0 (Stata Corp., College Station, TX, USA).

The study was approved by the Via Christi Institutional Review Board, Wichita, Kansas.

RESULTS

The study data were derived from patients with RA with a median duration of RA of 12.8 years. The mean age was 61.2 (SD 13.0) years, and 21.4% of participants were men (Table 1). Four of the 5 EQ-5D item variables were almost binary (Table 1). For example, "Confined to bed" was endorsed by only 0.3% and "Unable to wash/dress myself" by 0.4%. By contrast, HAQ (1.03, SD 0.73), mood (2.7, SD 1.8), and VAS pain scores (3.8, SD 2.7) had wide variability.

Values for the key scales were SF-6D 0.69 (SD 0.13), UK EQ-5D 0.64 (SD 0.28), US EQ-5D 0.73 (SD 0.19), and HAQ 1.03 (SD 0.73). The mean difference between UK and US EQ-5D scores was 0.094 units. The observed range of the SF-6D was 0.34 to 1.00, with only 5.0% of scores < 0.5. The range of UK EQ-5D was –0.59 to 1, with 15.0% of scores < 0.5. The US EQ-5D ranged from –0.11 to 1, and 14.0% of the scores were < 0.50. Thus the UK EQ-5D scores are shifted to the left and the scale has a lower limit compared with the US EQ-5D.

The consequences of these distributional differences can be seen in Figure 1, where mean utility scores are plotted at each level of HAQ score. Although the SF-6D aligns closely with the UK EQ-5D at HAQ values up to 1.0, the curves diverge after that. The observed minimum mean score of the SF-6D is 0.50, compared with 0.23 observed for the US EQ-5D and –0.06 for the UK EQ-5D. Scores were always higher ("better") for the US EQ-5D compared with the UK EQ-5D, and the difference in scores increased with increasingly more extreme HAQ scores.

To study the relation between clinical scores and utilities, and the relation between change in scores over 6 months, we utilized a correlation matrix of the key study variables (Table 2). HAQ and pain were correlated with the 3 utility scores at values between 0.625 and 0.681; slightly lower correlations were noted with mood. The correlation between the SF-6D and US EQ-5D was 0.689, and between SF-6D and UK EQ-5D was 0.673. We also examined the correlation between changes in the various scores in questionnaires administered 6 months apart. HAQ change correlated with US EQ-5D change –0.300, UK EQ-5D change –0.289, and SF-6D change –0.258. Pain change correlated with US EQ-5D change –0.363, UK EQ-5D change –0.364, and SF-6D change –0.258. The change in SF-6D was correlated with US EQ-5D change at 0.260 and UK EQ-5D change at 0.250.

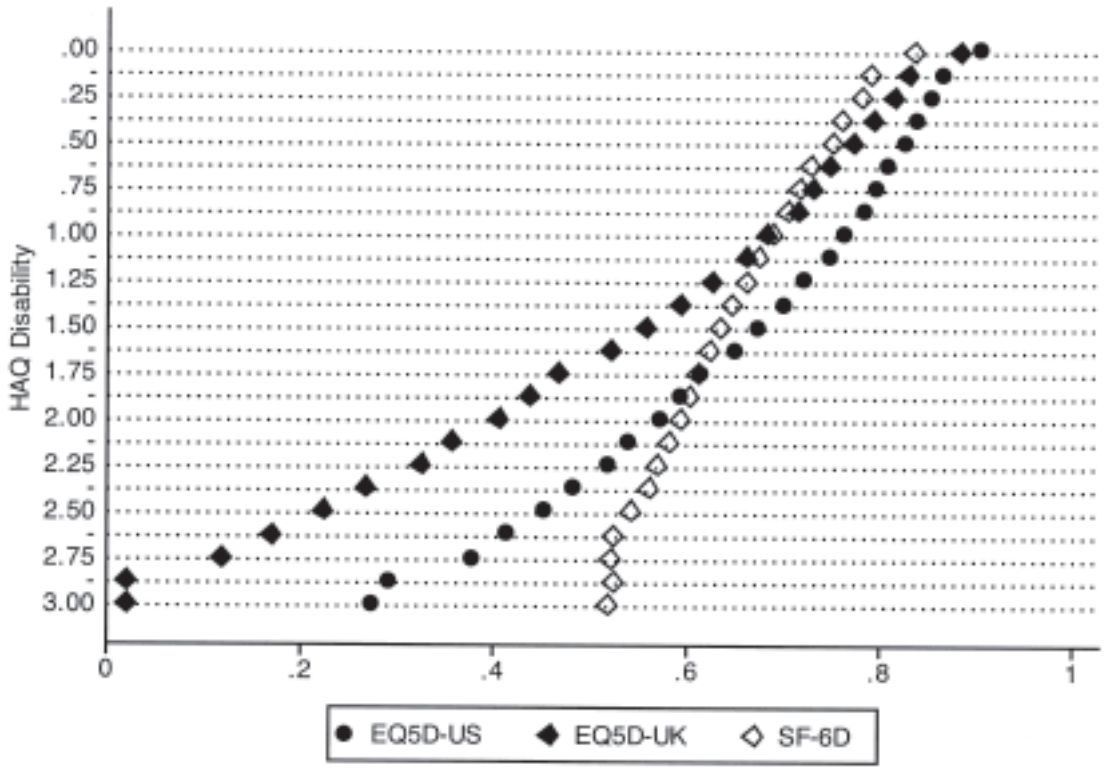

Figure 1. Mean values of US EQ-5D, UK EQ-5D, and SF-6D at all levels of the Health Assessment Questionnaire disability scale.

Table 2. Correlation of study variables and 6-month change in study variables, sorted by strength of association with HAQ and change in HAQ (n = 8,667).

Variable	HAQ	Pain	Mood	US EQ-5D	UK EQ-5D	SF-6D
HAQ	1.000	0.597	0.353	–0.679	–0.663	–0.674
US EQ-5D	–0.679	–0.681	–0.517	1.000	0.993	0.689
SF-6D	–0.674	–0.625	–0.664	0.689	0.673	1.000
UK EQ-5D	–0.663	–0.680	–0.504	0.993	1.000	0.673
Pain	0.597	1.000	0.403	–0.681	–0.680	–0.625
Mood	0.353	0.403	1.000	–0.517	–0.504	–0.664
Variable	Δ HAQ	Δ Pain	Δ Mood	Δ US EQ-5D	Δ UK EQ-5D	Δ SF-6D
Δ HAQ	1.000	0.308	0.110	–0.300	–0.289	–0.258
Δ Pain	0.308	1.000	0.107	–0.363	–0.364	–0.258
Δ US EQ-5D	–0.300	–0.363	–0.197	1.000	0.986	0.260
Δ UK EQ-5D	–0.289	–0.364	–0.187	0.986	1.000	0.250
Δ SF-6D	–0.258	–0.258	–0.394	0.260	0.250	1.000
Δ Mood	0.110	0.107	1.000	–0.197	–0.187	–0.394

HAQ: Health Assessment Questionnaire disability scale; HAQ-II: HAQ disability scale II; SF-6D: Short-form 6D; US EQ-5D: United States EQ-5D; UK EQ-5D: United Kingdom EQ-5D; Mood: SF-36 mental health score.

We also examined mean utility scores for important clinical subgroups, as shown in Table 3. In conditions when patients were severely affected — as in "poor health" and high regional pain scores — UK EQ-5D scores were very low compared with other scale scores, and SF-6D scores were unable to become low enough to adequately represent the adverse health condition. For example, patients reporting "poor health" had SF-6D scores of 0.53, US EQ-5D scores of 0.45, and UK EQ-5D scores of 0.21. For RA patients with the highest (most abnormal) WPI values, the associated utility scores were 0.63, 0.46, and 0.23; and for fibromyalgia occurring in RA, the scores were 0.57, 0.55, and 0.46. In terms of agreement with each other, the SF-6D and US EQ-5D were similar for "disabled," "total joint replacement," high levels of comorbidity, and survey fibromyalgia. Overall, the SF-6D and US EQ-5D appear often to identify similar groups, as suggested by Figure 1, particularly in comparison with the UK EQ-5D.

Table 3. Utility scores by important RA status variables.

Variable	%	US EQ-5D	UK EQ-5D	SF-6D
Current Health				
Excellent	6.7	0.92	0.90	0.85
Good	49.0	0.81	0.75	0.74
Fair	36.3	0.66	0.54	0.62
Poor	8.0	0.45	0.21	0.53
Disabled				
No	86.0	0.76	0.68	0.71
Yes	14.0	0.56	0.38	0.58
Total joint replacement				
No	78.1	0.74	0.65	0.69
Yes	21.9	0.70	0.59	0.67
Comorbidity score				
0	23.2	0.81	0.75	0.75
1	27.4	0.76	0.68	0.71
2	22.3	0.72	0.62	0.67
3	13.8	0.68	0.57	0.64
4–9	13.4	0.62	0.48	0.61
Total medical costs (mean US$)				
Quartile 1 (740)	25.0	0.77	0.69	0.72
Quartile 2 (2,922)	25.0	0.72	0.61	0.68
Quartile 3 (10, 0.13)	25.0	0.74	0.66	0.69
Quartile 4 (22, 010)	25.0	0.71	0.60	0.67
Widespread Pain Index score				
0	12.4	0.87	0.83	0.81
8	4.6	0.68	0.57	0.64
19	2.2	0.46	0.23	0.63
Survey fibromyalgia				
No	81.6	0.77	0.70	0.71
Yes	18.4	0.55	0.37	0.57

Mapping the US and UK EQ-5D and the SF-6D. To develop a predictive model, and usable predictive results, we took several approaches. First, we attempted to predict each of the 5 EQ-5D item results by ordered and binary logistic regression using the HAQ items to predict the 3 functional questions, the VAS pain score to predict the pain question, and the SF-36 mood score to predict the anxiety and depression question. This method proved unsuccessful because of very high rates of misclassification of each item. Such results might have been expected by the distribution of the EQ-5D components that, except for pain, are almost binary, and incapable of identifying the nuances of function, pain, and mood (Table 1).

We then turned to linear regression to model the relationship between utility scores and predictor variables. The method of approach is illustrated in detail for the US EQ-5D (Table 4), and in slightly less detail for the UK EQ-5D and the SF-6D. Results of the analyses of Table 4, in terms of beta coefficients that can be used for clinical prediction, are presented in Table 5 and Appendix 1. In Table 4, each addi-

Table 4. Predictive performance of predictive models (n = 10,895).

Description	US EQ-5D R-square Adj. (RMSE) ISMAE/OSMAE	UK EQ-5D R-square Adj. (RMSE) ISMAE/OSMAE	SF-6D R-square Adj. (RMSE) ISMAE/OSMAE	US EQ-5D AIC (BIC)	US EQ-5D Pearson Correlation (95% LOA)
HAQ only					
HAQ score (continuous)	0.45 (0.14)	0.43 (0.21)	0.44 (0.10)	–12275	0.673
	0.106/0.106	0.158/0.157	0.080/0.080	(–12260)	(± 0.270)
HAQ score (25 categories)	0.46 (0.14)	0.44 (0.21)	0.45 (0.10)	–12437	0.681
	0.103/0.102	0.152/0.151	0.078/0.078	(–12255)	(± 0.267)
8 HAQ domains (24 categories)	0.50 (0.13)	0.48 (0.20)	0.49 (0.09)	–13328	0.711
	0.100/0.100	0.149/0.147	0.075/0.075	(–13145)	(0.257)
20 HAQ items (80 categories)	0.56 (0.12)	0.54 (0.19)	0.51 (0.09)	–14645	0.751
	0.092/0.092	0.135/0.135	0.074/0.074	(–14200)	(± 0.241)
HAQ + pain					
HAQ score (continuous) + VAS pain	0.57 (0.12)	0.55 (0.19)	0.52 (0.09)	–14828	0.753
	0.096/0.096	0.142/0.141	0.074/0.074	(–14806)	(± 0.240)
20 HAQ items (80 categories) + VAS pain	0.63 (0.11)	0.61 (0.17)	0.55 (0.09)	–16414	0.794
	0.087/0.088	0.129/0.128	0.071/0.071	(–15962)	(± 0.222)
HAQ + pain + mood					
HAQ score (continuous) + VAS pain + mood	0.62 (0.11)	0.60 (0.18)	0.68 (0.07)	–16207	0.787
	0.090/0.089	0.135/0.133	0.059/0.060	(–16178)	(± 0.225)
20 HAQ items (80 categories) + VAS pain + mood	0.67 (0.11)	0.64 (0.17)	0.72 (0.07)	–17562	0.817
	0.083/0.083	0.123/0.123	0.057/0.058	(–17102)	(± 0.211)
HAQ + pain + mood + covariates*					
HAQ (continuous) + VAS pain + mood + covariates	0.65 (0.11)	0.61 (0.17)	0.69 (0.07)	–17094	0.808
	0.089/0.088	0.134/0.132	0.059/0.059	(–16664)	(± 0.215)
20 HAQ items (80 categories) + VAS pain + mood + covariates	0.68 (0.10)	0.65 (0.16)	0.70 (0.07)	–18147	0.829
	0.082/0.083	0.123/0.123	0.057/0.057	(–17286)	(± 0.204)

* Model with covariates also includes sex, age, age squared, RA duration, and education level. Observations studied were a single observation per patient from 12,098 patients. However, 9.1% of the 20 HAQ items were left incomplete by patients (missing). Therefore, so that models would be comparable, we only used the 10,895 observations with complete data. RMSE: Root mean square error; ISMAE: In-sample mean absolute error: 10,895 cases; AIC: Akaike information criterion; BIC: Bayesian information criterion; LOA: Bland-Altman limits of agreement.

Table 5. Predictive equations for preference-based utility measures (n = 10,092).

	US EQ-5D			UK EQ-5D			SF-6D		
	Coefficient	Standard Error	t	Coefficient	Standard Error	t	Coefficient	Standard Error	t
HAQ	–0.172	0.002	–95.0	–0.248	0.003	–90.2	–0.120	0.001	–92.7
Intercept	0.911	0.002	405.6	0.895	0.003	262.0	0.811	0.002	506.9
HAQ	–0.107	0.002	–52.8	–0.148	0.003	–48.3	–0.082	0.002	–54.6
VAS Pain	–0.029	0.001	–53.7	–0.044	0.001	–54.2	–0.017	0.000	–41.3
Intercept	0.953	0.002	444.4	0.959	0.003	294.8	0.835	0.002	522.9
HAQ	–0.097	0.002	–50.7	–0.134	0.003	–45.9	–0.070	0.001	–56.9
VAS pain	–0.024	0.001	–46.2	–0.037	0.001	–47.0	–0.010	0.000	–31.2
Mental health	0.002	0.000	38.4	0.004	0.000	35.8	0.003	0.000	76.3
Intercept	0.741	0.006	126.1	0.657	0.009	73.1	0.566	0.004	150.2

HAQ: Health Assessment Questionnaire disability scale; HAQ-II: HAQ disability scale II; SF-6D: Short-form 6D; US EQ-5D: United States EQ-5D; UK EQ-5D: United Kingdom EQ-5D. Mental health: SF-36 mental health t-score. Utility score: Intercept + (Pain* coefficient) + (Mood* coefficient) + (HAQ* coefficient) + ... Pain and mood are only used in calculations when they were collected. Predictive equations for HAQ items are shown in Appendix 1.

tional model is generally shown to provide better fit (adjusted R-square) and predictive accuracy (RMSE, MAE, and for the US EQ-5D: AIC, BIC, Pearson correlation, and LOA). As a measure of the reliability of the study models, we examined the MAE in the development (or in-sample) model and in a second data set (out-of-sample model). As the results of the ISMAE and OSMAE were virtually indistinguishable in each of the 10 models, we present all other data from the primary, in-sample models.

We assumed that the modeling results of this study might

be used under conditions when only HAQ data are available for prediction, or where HAQ and pain data are available for prediction, or where HAQ, pain, and mood data are available for prediction. So we provided analyses to cover each of these uses. If only HAQ data are available, using all 20 HAQ items is superior to using just the HAQ score, assuming the individual HAQ items are available. This observation is true across the 3 utility measures. The data indicate that it is always much better to use a model that includes pain (HAQ + Pain section). Better fit and accuracy is obtained when mood scores are added, although the incremental benefit of the addition of this variable is relatively small. In all cases, using the 20 HAQ items improves fit and accuracy. Differences between the models and model improvement can be seen clearly by observing the AIC, BIC, correlation, and LOA changes.

Prediction of the UK EQ-5D was less satisfactory than prediction of the US EQ-5D or the SF-6D. The RMSE was more than double for the UK EQ-5D compared with the SF-6D in all models, and somewhat less than double for the US EQ-5D compared with the UK EQ-5D. The SF-6D R-square improves substantially with the addition of the mood question. This might be expected to happen as the mood questions are (partially) included in the SF-6D.

Much of the improvement in EQ-5D models that is noted by using HAQ items and pain and mood occurs at lower EQ-5D levels, as shown in Figure 2. Model fit and accuracy deteriorate at levels below 0.5. However, only 14% of US and 15% of UK EQ-5D scores are lower than 0.5. Little is to be gained by adding covariates such as age, sex, RA duration, education, and comorbidity. All these variables are usually not reported in studies or are not available in the forms used in this study.

HAQ-II. Although not specifically reported here, the RMSE of the HAQ-II was 0.11 and the adjusted R-square was 0.65 in the HAQ-II item plus pain and mood mapping to the US EQ-5D, and was 0.17 and 0.62 mapped to the UK EQ-5D. Thus, the HAQ-II is almost identical in its predictive ability compared with the HAQ. Specific model details are available from the authors.

DISCUSSION

The EQ-5D and the SF-6D measures used in rheumatic diseases have been based on preference valuations or weights developed in the UK. North American studies, performed mostly in Canada, have also used the UK preference weightings. The UK EQ-5D weights were described in 1996[10], and the SF-6D was first reported in 2002[17]. The US EQ-5D pref-

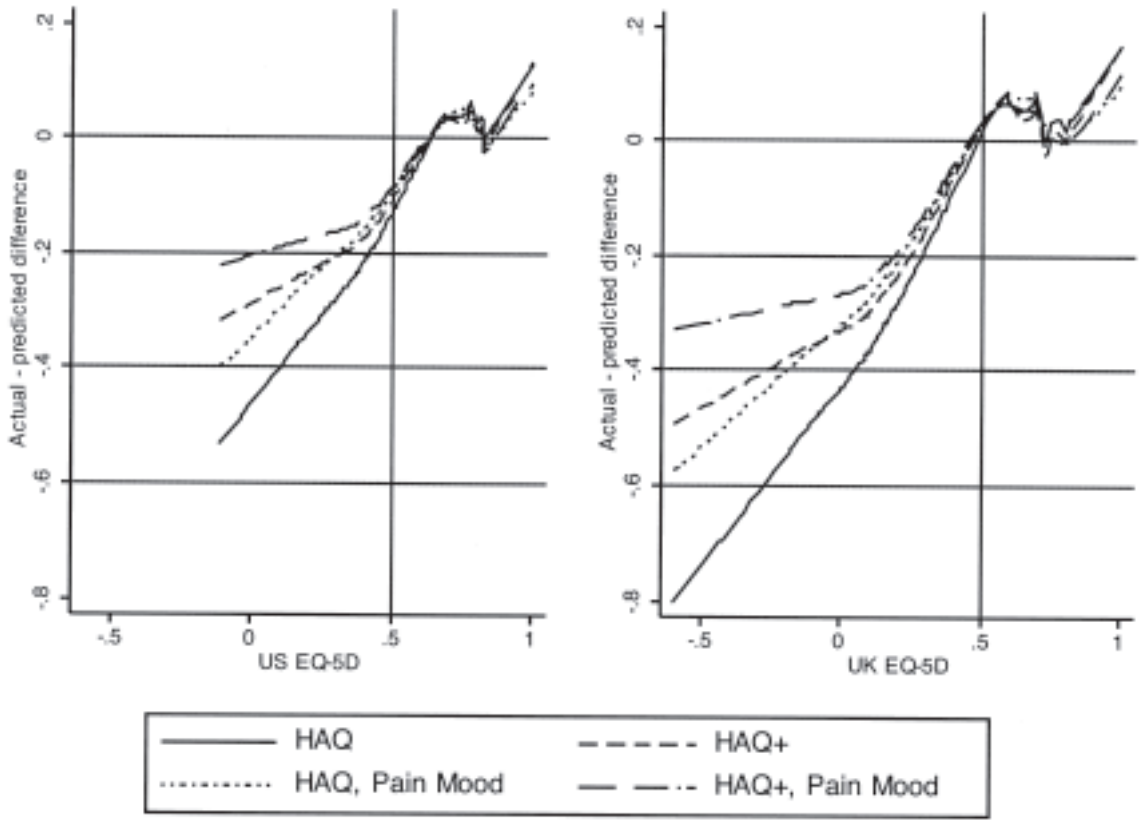

Figure 2. Actual versus predicted differences for the US and UK EQ-5D scales; 14% of US and 15% of UK EQ-5D scores are < 0.5. In the legend, HAQ represent the model using HAQ as a continuous predictor. HAQ+ uses the 80 HAQ item categories. Pain and Mood are additional predictors. HAQ: Health Assessment Questionnaire disability scale.

erence weights were published in 2005[16], but have not been reported previously in patients with rheumatic disease. With the publication of the US EQ-5D weights, it was noted that the mean US EQ-5D score was 0.11 units higher than the mean score of the UK EQ-5D[12], and that differences between the 2 measures were most profound in individuals with poor health status.

Marra, *et al* showed that use of the HUI-2, HUI-3, UK EQ-5D, and SF-6D in patients with RA led to different utility scores and, when applied to cost utility analyses, resulted in different QALY depending on which scale was used[11,33]. We found that for any RA health status state (Table 2), the US EQ-5D always had a higher (better) utility score than the UK version. Overall, we found that the US scores were 0.094 units higher than the UK EQ-5D scores. When the 3 scales were compared using the HAQ as an anchor (Figure 1), the US EQ-5D scores were higher than the SF-6D scores from HAQ values of 0 to 1.75. Thereafter, SF-6D scores were higher owing to the limited scaling of the measure.

In addition to the absolute differences between the UK EQ-5D and the SF-6D scores, the utility score changes after an intervention were larger when the UK EQ-5D was used, compared with the SF-6D. The absolute change differences and responsiveness for the UK EQ-5D and the SF-6D may also depend on baseline RA severity and whether there is improvement or worsening of the clinical state[9] (and Michaud and Wolfe, unpublished data). The US EQ-5D has not yet been studied with respect to changes observed in clinical trials, but it is likely that they will have an intermediate position between the UK EQ-5D and the SF-6D. The above observations present problems in 3 respects: (1) the validity of utility measures, given that they yield different results; (2) the sensitivity of cost utility analyses to the utility measure selected; and (3) the problem of how data should be analyzed when patients in multinational studies are assessed.

The use of mapping of clinical variables to utility variables came about when it was recognized, retrospectively, that economic analyses were valuable, but formal utility scales had not been administered. A wide variety of predictor variables have been mapped in different illnesses[34]. In rheumatic diseases, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) has been mapped to the UK EQ-5D[35,36] and the HUI-3[32]. In RA, mapping efforts have used the HAQ to predict UK EQ-5D and SF-6D scores. Bansback, *et al* provided the first careful and detailed analyses of the HAQ as a predictor of the UK EQ-5D and SF-6D[1]. They found that use of selected "domains" from the HAQ best predicted UK EQ-5D and SF-6D scores, with RMSE values of 0.183 and 0.089 and R-squares of 0.57 and 0.51, respectively. A recent study by Amadi, *et al* found that a HAQ mapped SF-6D score was valid and responsive in early RA[37].

There are essentially 4 approaches to using the HAQ as a predictor: (1) using the calculated final HAQ score in a linear regression; (2) using the final scores in a nonlinear regression or a fractional polynomial regression; (3) using HAQ domains (as used by Bansback, *et al*); or (4) using individual HAQ items as categorical predictors. In agreement with Bansback, *et al*, we found domains to be superior to the HAQ score or to a nonlinear application of the HAQ score (Table 3). However, we found that individual HAQ items were the best predictors when the RSME, MAE, R-square, and other model and prediction statistics were considered (Table 4). However, using individual items is burdensome (Appendix 1), although calculable by computer. Another limitation is that often only mean HAQ scores are available from published studies, therefore eliminating the use of individual items as a potential means of predicting utility values. But if study variable data are available, there is considerable advantage to the item method (Tables 4 and 6).

In almost all settings where HAQ is available, a VAS pain scale is also available. There are substantial gains in model accuracy and fit by using the HAQ and pain scores together to predict utilities (Tables 4 and 6). The model improves further by the incorporation of the SF-36 mental health domain score, but this score, although often a part of clinical trial data, is not ordinarily reported in primary trial results. Similarly, some additional improvement in prediction and fit can be obtained by incorporating other covariates such as age, sex, education, comorbidity, and duration of RA (Table 4). However, the added improvement is small and these covariates are not always available (education) or are collected using a common scale (comorbidity). These results lead us to recommend the continuous HAQ and VAS pain scale (and mood scale, if available) when item data are not available, and the more complex HAQ items as a substitute for the HAQ summary score, if available. The mapping of the utility measures described here expand on methods currently available, and should represent an improvement in validity.

The initial report of clinical relevance of the UK EQ-5D came from Hurst, *et al* in 1997[38]. They found that the EQ-5D demonstrated "moderate to high correlations with measures of impairment and high correlations with disability measures," and was reliable and valid. Recent studies that included change data have confirmed these findings[9]. However, transformation of clinical data to EQ-5D is not without problems. As shown in Table 1, 73.8% of patients endorsed the EQ-5D category of "moderate pain or discomfort." The VAS pain score for those in that category was 3.7 (SD 2.3), indicating large variance and the inability of the EQ-5D to determine important clinical differences. Similarly, the EQ-5D category of "I have no problems walking about" represents a crude clinical measure.

Scott, *et al* raised the issue of "real clinical concern" over the use of utility indices to measure the outcome of clinical

APPENDIX. Coefficients for mapping HAQ items, pain, and mood to US EQ-5D, UK EQ-5D, and SF-6D.

Variables	*Level*	*ES EQ-5D coefficients*			*UK EQ-5D coefficients*			*SF-6D coefficients*		
dresself	1	-0.0209164	-0.0107955	-0.0104924	-0.0286289	-0.0130260	-0.0125969	-0.0157935	-0.0100002	-0.0095682
	2	-0.0510455	-0.0355375	-0.0406870	-0.0790171	-0.0551091	-0.0623983	-0.0235009	-0.0146239	-0.0219615
	3	-0.0200821	-0.0146465	-0.0206087	-0.0202447	-0.0118649	-0.0203045	-0.0178609	-0.0147495	-0.0232451
shampoo	1	-0.0172894	-0.0111708	-0.0085561	-0.0218776	-0.0124447	-0.0087435	-0.0094028	-0.0059004	-0.0021746
	2	-0.0400077	-0.0303082	-0.0253929	-0.0606191	-0.0456659	-0.0387081	-0.0214110	-0.0158589	-0.0088548
	3	0.0010129	-0.0050219	-0.0078813	0.0050864	-0.0042172	-0.0082648	-0.0004873	-0.0039417	-0.0080161
standup	1	-0.0010491	0.0048323	0.0038787	-0.0008022	0.0082649	0.0069150	-0.0043265	-0.0009599	-0.0023187
	2	-0.0096592	0.0030511	0.0001530	-0.0160061	0.0035888	-0.0005136	0.0004044	0.0076800	0.0035503
	3	-0.0256991	-0.0223289	-0.0263074	-0.0347935	-0.0295978	-0.0352294	0.0057495	0.0076786	0.0020096
inbed	1	-0.0271633	-0.0117749	-0.0052891	-0.0415555	-0.0178318	-0.0086511	-0.0188478	-0.0100393	-0.0007975
	2	-0.0923054	-0.0612107	-0.0433716	-0.1460464	-0.0981090	-0.0728575	-0.0416808	-0.0238817	0.0015375
	3	-0.1421166	-0.1081804	-0.0726423	-0.2268627	-0.1745448	-0.1242400	-0.0840812	-0.0646557	-0.0140166
cutmeat	1	-0.0006750	-0.0038233	-0.0004247	0.0000926	-0.0047611	0.0000496	-0.0003820	-0.0021841	0.0026586
	2	-0.0039791	-0.0058636	0.0006446	-0.0042515	-0.0071569	0.0020557	-0.0096142	-0.0106930	-0.0014192
	3	-0.0276155	-0.0197987	-0.0092085	-0.0404031	-0.0283523	-0.0133617	-0.0208702	-0.0163958	-0.0013056
liftcup	1	0.0029545	0.0012531	-0.0018449	0.0066667	0.0040436	-0.0003416	0.0101413	0.0091673	0.0047530
	2	-0.0082273	-0.0088073	-0.0123182	-0.0153799	-0.0162741	-0.0212439	0.0119777	0.0116457	0.0066429
	3	-0.0097699	-0.0062478	-0.0075026	-0.0169245	-0.0114947	-0.0132709	0.0125049	0.0145210	0.0127330
openmilk	1	0.0039678	0.0053398	0.0056801	0.0068099	0.0089251	0.0094067	-0.0062773	-0.0054919	-0.0050071
	2	0.0084268	0.0098826	0.0067215	0.0132897	0.0155340	0.0110594	-0.0005227	0.0003106	-0.0041937
	3	0.0334204	0.0336987	0.0263025	0.0539924	0.0544215	0.0439521	0.0062960	0.0064554	-0.0040836
walkflat	1	-0.0146739	-0.0135528	-0.0134169	-0.0207514	-0.0190230	-0.0188307	-0.0040601	-0.0034184	-0.0032248
	2	-0.0458891	-0.0388965	-0.0392040	-0.0721524	-0.0613722	-0.0618074	-0.0126674	-0.0086648	-0.0091029
	3	-0.0449620	-0.0439822	-0.0413594	-0.0666370	-0.0651264	-0.0614139	-0.0066470	-0.0060861	-0.0023489
climstep	1	-0.0240416	-0.0131477	-0.0119735	-0.0373827	-0.0205881	-0.0189260	-0.0191221	-0.0128863	-0.0112132
	2	-0.0388777	-0.0187443	-0.0171594	-0.0615475	-0.0305087	-0.0282652	-0.0246652	-0.0131406	-0.0108822
	3	-0.0380033	-0.0173043	-0.0165277	-0.0594569	-0.0275460	-0.0264468	-0.0245869	-0.0127385	-0.0116320
washbody	1	-0.0424382	-0.0395975	-0.0379089	-0.0542944	-0.0499151	-0.0475249	-0.0097517	-0.0081257	-0.0057196
	2	-0.0437475	-0.0338301	-0.0311998	-0.0544789	-0.0391897	-0.0354664	-0.0275865	-0.0219096	-0.0181617
	3	-0.1112697	-0.1219570	-0.1120519	-0.1331262	-0.1496023	-0.1355814	-0.0497549	-0.0558724	-0.0417584
tubbath	1	-0.0008469	-0.0011134	-0.0037939	-0.0004310	-0.0008418	-0.0046362	-0.0046408	-0.0047933	-0.0086129
	2	0.0000548	0.0014620	-0.0037095	0.0006201	0.0027894	-0.0045309	-0.0022387	-0.0014332	-0.0088022
	3	0.0148090	0.0100990	-0.0005246	0.0233871	0.0161258	0.0010880	0.0104564	0.0077603	-0.0073774
ontoilet	1	-0.0036051	-0.0072233	-0.0082396	-0.0020176	-0.0075957	-0.0090342	0.0045400	0.0024689	0.0010208
	2	-0.0330435	-0.0316041	-0.0242748	-0.0502560	-0.0480368	-0.0376621	-0.0134460	-0.0126221	-0.0021785
	3	-0.0155227	-0.0373578	-0.0263340	-0.0170024	-0.0506647	-0.0350604	0.0180821	0.0055834	0.0212915
overhead	1	-0.0053139	-0.0021734	-0.0028227	-0.0089031	-0.0040616	-0.0049806	-0.0082846	-0.0064870	-0.0074121
	2	-0.0164994	-0.0096235	-0.0120896	-0.0264855	-0.0158853	-0.0193761	-0.0087096	-0.0047737	-0.0082877
	3	-0.0039453	-0.0027025	-0.0051775	-0.0025962	-0.0006802	-0.0041836	-0.0058711	-0.0051597	-0.0086863
benddown	1	-0.0149354	-0.0087555	-0.0047985	-0.0192464	-0.0097192	-0.0041180	-0.0098695	-0.0063321	-0.0006936
	2	-0.0554777	-0.0429912	-0.0309974	-0.0841508	-0.0649010	-0.0479235	-0.0238867	-0.0167393	0.0003510
	3	-0.0665197	-0.0599750	-0.0431892	-0.1065060	-0.0964164	-0.0726557	-0.0206156	-0.0168693	0.0070492
opencar	1	0.0064116	0.0037864	0.0042868	0.0116082	0.0075610	0.0082693	0.0020613	0.0005586	0.0012717
	2	0.0203300	0.0195750	0.0174581	0.0337677	0.0326038	0.0296072	0.0113343	0.0109021	0.0078856
	3	-0.0035015	0.0071164	-0.0149887	0.0058330	0.0222022	-0.0090878	0.0312814	0.0373593	0.0058613
openjars	1	-0.0088382	-0.0027482	-0.0005143	-0.0114226	-0.0020340	0.0011281	-0.0103717	-0.0068858	-0.0037026
	2	-0.0361213	-0.0214850	-0.0120449	-0.0508676	-0.0283034	-0.0149408	-0.0262160	-0.0178380	-0.0043865
	3	-0.0719955	-0.0468271	-0.0455762	-0.1132833	-0.0744822	-0.0727116	-0.0260366	-0.0116299	-0.0098474
fauceton	1	-0.0043811	-0.0056523	-0.0053269	-0.0055143	-0.0074742	-0.0070134	-0.0041629	-0.0048906	-0.0044268
	2	-0.0215957	-0.0198653	-0.0189281	-0.0297556	-0.0270879	-0.0257613	-0.0108691	-0.0098786	-0.0085432
	3	-0.0163850	-0.0137316	-0.0164460	-0.0197131	-0.0156225	-0.0194647	0.0040029	0.0055218	0.0016540
runerand	1	-0.0308902	-0.0207622	-0.0165810	-0.0483409	-0.0327270	-0.0268085	-0.0303045	-0.0245071	-0.0185492
	2	-0.0683132	-0.0510515	-0.0426032	-0.1067338	-0.0801221	-0.0681634	-0.0457308	-0.0358499	-0.0238118
	3	-0.0811619	-0.0793218	-0.0674274	-0.1227204	-0.1198836	-0.1030470	-0.0568388	-0.0557855	-0.0388370
inoutcar	1	-0.0069681	-0.0056268	-0.0050783	-0.0107932	-0.0087254	-0.0079490	-0.0060022	-0.0052345	-0.0044529
	2	-0.0213781	-0.0136055	-0.0114421	-0.0341098	-0.0221270	-0.0190647	-0.0134266	-0.0089774	-0.0058948
	3	-0.0637539	-0.0602002	-0.0629017	-0.0805562	-0.0750775	-0.0789016	-0.0062906	-0.0042564	-0.0081059
vacuum	1	-0.0387638	-0.0231620	-0.0196066	-0.0549551	-0.0309023	-0.0258696	-0.0624639	-0.0535332	-0.0484671
	2	-0.0568949	-0.0342803	-0.0302655	-0.0829309	-0.0480670	-0.0423840	-0.0810832	-0.0681384	-0.0624176
	3	-0.0680161	-0.0455794	-0.0413184	-0.0997766	-0.0651869	-0.0591554	-0.0871734	-0.0743304	-0.0682588
VAS pain	-		-0.0234228	-0.0203486		-0.0361099	-0.0317584		-0.0134075	-0.0090270
Mental	-			0.0021600			0.0030575			0.0030778
Intercept	-	0.8833296	0.9194756	0.7378074	0.8536995	0.9094242	0.6522703	0.8144202	0.8351106	0.5762477

Variable names are short names for HAQ items. Mental = SF-36 mental health T-score.
Utility score = Intercept + (Pain*coefficient) + (Mood*coefficient) + (HAQ item level*coefficient) + ... Pain and mood are only used in calculations when they were collected. For each HAQ item, only score use the item level that was selected. HAQ item scores are repeated for each item.

care[39]. In addition and in particular, they stated that "HAQ and EuroQol are demonstrably not equivalent, [and] economic evaluations of treatment cost effectiveness should not be based on EuroQol data transformed from HAQ." They noted in their study of 56 patients that 6-month "...changes in HAQ and EuroQol were unrelated (r = 0.08)," while the correlation between changes in the EQ-5D and changes in pain was 0.54. While we found 6-month changes in the HAQ to correlate with changes in the UK EQ-5D at r = 0.300 and changes in the EQ-5D to correlate with changes in pain at r = 0.363 (Table 2), the concerns of Scott, *et al* are important and reflect the ongoing tensions between clinical measurement and patient and societal valuation[40,41].

They also reflect the omnipresent but often unspoken problem of the use of clinical data for administrative decisions at the level of the patient — particularly in the face of measurement error. As shown in Table 3, which measures the difference between actual (observed) and predicted (mapped) values, the best case Bland-Altman LOA was ± 0.204 units and the worst case was ± 0.270 units. Thus, if mapped values are applied at the individual patient level, an unreliable estimate of the actual health state may be obtained, and these differences do not even consider HAQ measurement error. Most cost-effectiveness studies, however, do not use patient-level data, and the RMSE levels found in our study are acceptable for group use. With respect to mapping of EQ-5D data, incorporation of pain, and possibly of mood, provides additional assurance of utility scores that correlate with clinical experience.

A case can be made that the use of any of the mapped models is acceptable. However, all things being equal, the model with the smallest predictive error should be preferred. As shown by Grootendorst, *et al*, the confidence intervals around the predictive values depend on the sample size of the study that the predictions are applied to[32], a finding we also noted (data not shown).

Although mapped utilities can be helpful when actual utility scores are not available, mapped utilities can have additional substantial limitations. Barton, *et al* showed that in patients with osteoarthritis, "mapping models developed from the WOMAC tended to underestimate the QALY gain associated with each of four interventions, compared to that which was derived from actual [UK] EQ-5D scores"[35]. One explanation for this observation is that "prediction errors...tend to be increasingly positive for lower EQ-5D scores and increasingly negative for higher EQ-5D scores," a finding that we observed in the current study in Figure 2 (error direction is reversed by subtraction method) and others have also noted in RA studies of mapped EQ-5D scales. These findings, and the inherent error in mapping[42,43], lead us to advise the use of the 5-item EQ-5D questionnaire or SF-6D rather than relying on secondary mapping.

In summary, the US EQ-5D differs from the UK version and from the SF-6D in mean scores and ranges. When determined by mapping, the US EQ-5D has a much lower prediction error than the UK EQ-5D. Simple mapping models that use HAQ and pain have acceptable error rates, although more complicated models that include individual HAQ items and mood scores improve predictive accuracy and model fit.

REFERENCES

1. Bansback N, Marra C, Tsuchiya A, Anis A, Guh D, Hammond T, et al. Using the Health Assessment Questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis Rheum 2007;57:963-71.
2. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. Pharmacoeconomics 1995;7:490-502.
3. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113-28.
4. Brennan A, Bansback N, Reynolds A, Conway P. Modelling the cost-effectiveness of etanercept in adults with rheumatoid arthritis in the UK. Rheumatology 2004;43:62-72.
5. Kobelt G, Jonsson L, Young A, Eberhardt K. The cost-effectiveness of infliximab (Remicade) in the treatment of rheumatoid arthritis in Sweden and the United Kingdom based on the ATTRACT study. Rheumatology 2003;42:326-35.
6. Wong JB, Singh G, Kavanaugh A. Estimating the cost-effectiveness of 54 weeks of infliximab for rheumatoid arthritis. Am J Med 2002;113:400-8.
7. Vera-Llonch M, Massarotti E, Wolfe F, Shadick N, Westhovens R, Sofrygin O, et al. Cost-effectiveness of abatacept in patients with moderately to severely active rheumatoid arthritis and inadequate response to methotrexate. Rheumatology 2008;47:535-41.
8. 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.
9. 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.
10. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol. York: Publications Unit, Centre for Health Economics, University of York; 1996.
11. 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.
12. Johnson JA, Luo N, Shaw JW, Kind P, Coons SJ. Valuations of EQ-5D health states: are the United States and United Kingdom different? Med Care 2005;43:221-8.
13. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Farragher TM, Verstappen SM, et al. 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.
14. Wolfe F, Michaud K. A brief introduction to the National Data Bank for Rheumatic Diseases. Clin Exp Rheumatol 2005;23:S168-S71.
15. Wolfe F, Michaud K, Li T, Katz RS. EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: Comparisons with RA, non-inflammatory rheumatic disorders, and fibromyalgia. J Rheumatol 2010;37:296-304.
16. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care 2005;43:203-20.

17. 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.
18. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36). 1. Conceptual framework and item selection. Med Care 1992;30:473-83.
19. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
20. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. Arthritis Rheum 2000;43:2751-61.
21. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire. II: A revised version of the Health Assessment Questionnaire. Arthritis Rheum 2004;50:3296-305.
22. Wolfe F. Why the HAQ-II can be an effective substitute for the HAQ. Clin Exp Rheumatol 2005;23 Suppl:S29-S30.
23. Ware JE Jr, Kosinski M. The SF-36 physical and mental health summary scales: A manual for users of Version 1, second edition. Lincoln, NE: Quality Metric, Inc.; 2001.
24. Michaud K, Wolfe F. The development of a rheumatic disease research comorbidity index for use in outpatients patients with RA, OA, SLE and fibromyalgia (FMS) [abstract]. Arthritis Rheum 2007;56 Suppl:S596.
25. Michaud K, Wolfe F. Comorbidities in rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007;21:885-906.
26. Wolfe F, Michaud K, Choi HK, Williams R. Annual and lifetime productivity costs and income losses in persons with rheumatoid arthritis [abstract]. Arthritis Rheum 2003;48 Suppl:S243.
27. Michaud K, Messer J, Choi HK, Wolfe F. Direct medical costs and their predictors in persons with rheumatoid arthritis: a 3 year study of 7,527 patients. Arthritis Rheum 2003;48:2750-62.
28. Wolfe F, Michaud K, Li T, Katz BS. Chronic conditions and health problems in rheumatic diseases: comparisons with rheumatoid arthritis, noninflammatory rheumatic disorders, systemic lupus erythematosus and fibromyalgia. J Rheumatol 2010;37:305-15.
29. Wolfe F. Pain extent and diagnosis: development and validation of the regional pain scale in 12,799 patients with rheumatic disease. J Rheumatol 2003;30:369-78.
30. Katz RS, Wolfe F, Michaud K. Fibromyalgia diagnosis: A comparison of clinical, survey, and American College of Rheumatology criteria. Arthritis Rheum 2006;54:169-76.
31. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60.
32. Grootendorst P, Marshall D, Pericak D, Bellamy N, Feeny D, Torrance GW. A model to estimate Health Utilities Index mark 3 utility scores from WOMAC index scores in patients with osteoarthritis of the knee. J Rheumatol 2007;34:534-42.
33. 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.
34. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010;11:215-25.
35. Barton GR, Sach TH, Jenkinson C, Avery AJ, Doherty M, Muir KR. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores? Health Qual Life Outcomes 2008;6:51.
36. Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain. Cost Eff Resour Alloc 2009;7:12.
37. Amjadi SS, Maranian PM, Paulus HE, Kaplan RM, Ranganath VK, Furst DE, et al. Validating and assessing the sensitivity of the Health Assessment Questionnaire Disability Index-derived Short Form-6D in patients with early aggressive rheumatoid arthritis. J Rheumatol 2009;36:1150-7.
38. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol 1997;3:551-9.
39. Scott DL, Khoshaba B, Choy EH, Kingsley GH. Limited correlation between the Health Assessment Questionnaire (HAQ) and EuroQol in rheumatoid arthritis: questionable validity of deriving quality adjusted life years from HAQ. Ann Rheum Dis 2007;66:1534-7.
40. Nord E, Richardson J, Macarounas-Kirchmann K. Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. Int J Technol Assess Health Care 1993;9:463-78.
41. Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should patients have a greater role in valuing health states? Appl Health Econ Health Policy 2005;4:201-8.
42. Kaplan RM, Groessl EJ, Sengupta N, Sieber WJ, Ganiats TG. Comparison of measured utility scores and imputed scores from the SF-36 in patients with rheumatoid arthritis. Med Care 2005;43:79-87.
43. Harrison MJ, Lunt M, Verstappen SM, Watson KD, Bansback NJ, Symmons DP. Exploring the validity of estimating EQ-5D and SF-6D utility values from the Health Assessment Questionnaire in patients with inflammatory arthritis. Health Qual Life Outcomes 2010;8:21.