The Validity and Responsiveness of Generic Utility Measures in Rheumatoid Arthritis: A Review

MARK J. HARRISON, LINDA M. DAVIES, NICK J. BANSBACK, MARY INGRAM, ASLAM H. ANIS, and DEBORAH P.M. SYMMONS

ABSTRACT. ***Objective.*** Cost-utility analysis is increasingly important as healthcare providers aim to invest scarce resources in interventions offering the greatest health benefit. The ability to attach utility values to health states is essential, and is increasingly performed using generic scales. However, the evidence regarding the validity of generic utility scales in rheumatoid arthritis (RA) is unclear. We summarize and review evidence on the validity and comparative performance of generic utility scales in RA.

Methods. We searched the English-language medical literature for studies using utilities in RA between 1980 and mid-2006. Reports describing primary evidence of the validity or performance of a generic utility scale in RA were selected, summarized, and reviewed using the OMERACT filter.

Results. In total 923 articles were identified, of which 228 reported the use of utility scales in RA; 26 studies related to the validation or evidence of generic utility scales in RA, the EQ-5D, Health Utility Index-2 (HUI2) and HUI3, SF-6D, and Quality of Well-Being Scale. The EQ-5D, HUI2 and HUI3, and SF-6D all have consistent evidence of construct validity and responsiveness in RA, but each has limitations.

Conclusion. The EQ-5D and HUI3 have been the most extensively studied instruments and show validity and responsiveness for use in RA, but both instruments have limitations. The SF-6D is relatively new and appears to have potential for use in milder RA, but needs further evaluation. More longitudinal head-to-head evaluation of measures is needed across the spectrum of RA disease severity to further investigate their comparative properties, and to seek consensus on the best utility measure for use in economic evaluation. (First Release Feb 15 2008; J Rheumatol 2008;35:592–602)

Key Indexing Terms:

RHEUMATOID ARTHRITIS | UTILITIES | QUALITY OF LIFE
QUALITY-ADJUSTED LIFE-YEARS | COST-EFFECTIVENESS | VALIDITY

Utility is the preference (rated in the presence of choice) for a health state relative to perfect health (scored 1) and death (scaled 0)[1]. Utility may be converted into quality-adjusted life-years (QALY) by multiplying the time spent in a health state by its utility. Cost-utility analysis (CUA) — or cost per QALY gained — is increasingly used to evaluate interventions, and has been adopted by organizations providing guidance on the use of new and existing interventions in many countries including the UK[2], Canada[3], and the USA[4].

Utility can be valued directly using techniques such as the standard gamble (SG), time tradeoff (TTO), and rating scales (RS). These differ in alignment with utility theory, and have been described comprehensively[5]. Direct measurement is generally unsuitable for use in large numbers of patients as it can be complicated, time-consuming, and costly. Instead, information about patients' health status collected using self-administered generic health questionnaires can be mapped to societal preferences for described health states to give estimates of health utility. Examples of such preference-based instruments include the SF-6D[6], EuroQol[7], and Health Utility Index (HUI)[8]. These utility measures allow quantification and comparison of the effects of diseases with multiple outcomes, such as rheumatoid arthritis (RA), on a single scale both within and across specialties. However, each must be validated within each disease in which it is used.

The use of generic preference-based measures in RA is increasing; however, there are gaps and conflicts in the evidence base regarding their validity in this setting. We used a filter developed by the OMERACT group[9] to summarize and review the evidence of the validity, and comparative

From the Arthritis Research Campaign (arc) Epidemiology Unit, and Health Economics Research at Manchester (HERMAN), The University of Manchester, Manchester, UK; Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital, Vancouver, British Columbia; and Department of Healthcare and Epidemiology, University of British Columbia, Vancouver, British Columbia, Canada.

M.J. Harrison, MSc, Research Assistant; M. Ingram, MSc, Librarian-Information Officer; D.P.M. Symmons, MD, Professor of Rheumatology and Musculoskeletal Epidemiology, Arthritis Research Campaign (arc) Epidemiology Unit; L.M. Davies, MSc, Reader, Director of Health Economics Research, Health Economics Research at Manchester; N.J. Bansback, MSc, Health Economist, Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital; A.H. Anis, PhD, Professor of Health Economics, Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital, and Department of Healthcare and Epidemiology, University of British Columbia.

Address reprint requests to Prof. D.P.M. Symmons, arc Epidemiology Unit, Stopford Building, The University of Manchester, Oxford Road, Manchester, M13 9PT, UK. E-mail: deborah.symmons@manchester.ac.uk

Accepted for publication November 17, 2007.

performance of the most popular preference-based measures (referred to here as utilities) in RA. We emphasize areas for future research and make recommendations on which measures to use based on the available evidence.

MATERIALS AND METHODS

Literature review. A comprehensive literature review of Medline, Embase, and Web of Science was conducted to identify utilities used in RA and musculoskeletal disease from 1980 to mid-2006. The scope was then limited to identifying and summarizing studies describing the use of utility measures or scores in patients with RA.

Search components were the condition (combined as: rheumatoid arthritis or RA or musculoskeletal disease or polyarthritis) and the subject (utilities). The subject was created using 2 strands, general terms (combined as: quality-adjusted life-years or QALYs or preference-based measures or utilities or utility) and named generic utility measures (combined as: EQ-5D [Euroqol or EQ5D or EQ-5D] or SF-6D [SF6D or SF-6D] or Health Utilities Index [Health Utilities Index or HUI$], Quality of Well-Being scale [Quality of Well-Being Scale or QWB]. General terms for utility measures aimed to ensure identification of all relevant studies using values, while the names of measures most commonly used across specialties[10] were included to ensure identification of studies validating these measures. The final search strategy for utilities in RA combined the utility terms (combined as: general terms or named generic utility measures) and the condition terms.

We screened the abstracts of identified reports and retained those using utility measures, QALY, or economic analysis in RA (MH and NB). At the second stage, reports describing primary evidence relating to the validity or performance of a generic utility in RA were selected. Other relevant studies were identified by checking reference lists of included studies (MH and NB).

Assessment of evidence. The OMERACT filter[9] framework was used to classify evidence into feasibility, truth, and discrimination of utilities in RA. Quantitative data under each heading were assessed using commonly used criteria in reviewing evidence[11-16,17]. Evidence from studies directly measuring and comparing multiple generic utility measures in a single study were preferred. These provide the most useful and informative comparisons of the performance of utility measures in this setting because differences in performance cannot be attributed to between-sample variation.

Feasibility considers the practicality of a measure in terms of time, financial, and interpretational constraints. This includes fees incurred in obtaining and using the measure, time taken to complete the questionnaire, and proportion that are correctly or completely filled out by the patient[17].

Truth relates to content (the extent that a measure covers the full range of aspects of the construct being measured) and construct validity (assessment of the performance of a measure against predictions or expectations based on the theory of the construct under study)[13,18]. We focused on attempts to test plausible hypotheses relating to classification or prognosis, or correlations with RA specific measures, or to identify predictive or component factors of the utility measure.

Discrimination is the ability of a measure to distinguish between states of interest at different points in time (to measure change). This relates to reliability and sensitivity or responsiveness to change. In order to measure change a measure needs to be consistent in stable subjects (reliability) and change appropriately with improvement and deterioration (sensitivity/responsiveness)[15,19].

Reliability (test-retest) is the ability of a measure to provide consistent results when repeated under identical conditions[13,17]. The intraclass correlation coefficient (ICC), which assesses the relative ordering of scores and mean differences in test and retest scores, is often used as a test-retest statistic[11]. The ICC for interexaminer reliability from Fleiss[14] may be preferred, as it gives the most complete information, accounting for correlation, slope, and intercept of agreement[20]. Alternative assessments correlating test and retest scores in unchanged patients describe only the line of best fit, which may differ markedly from the line of perfect agreement and overestimate reliability. In practice, both methods tend to yield comparable estimates[13,15]. Minimum acceptable test-retest reliability coefficients of 0.7 for group comparisons have been suggested[11,15], although cross-instrument comparisons within the same sample may be more appropriate than assessment against an arbitrary threshold. We considered evidence of reliability primarily using the ICC statistics, but supplemented this with evidence using stability coefficients.

Potential to detect change was assessed by the minimum important difference (MID), examining for floor and ceiling effects (percentage of patients occupying the worst/best health states). The MID is the smallest clinically relevant difference defined using self-reported change in health. Floor and ceiling effects are considered small if 1%–15% of patients occupy the worst and best health states, respectively, and serious if > 15% of patients occupy these states[12]. These criteria have been used in other reviews of outcome measures in musculoskeletal disease[21]. Floor and ceiling effects should also be assessed within domains, as overall scores can obscure these effects in any one domain; for example, a pain subscale might have floor effects in patients with RA, but the overall score may not show a floor effect.

Ability to measure change was assessed by examining the effect size for change (ES), standardized response mean (SRM), and relative efficiency (RE). Evidence of responsiveness is preferred where the importance of change is assessed, for example, using self-reported change or an assessment of clinical change, because in isolation the responsiveness statistic does not provide any information of whether the change detected is meaningful or useful. The ES and SRM determine the ratio of signal (size of change) to noise (variability in scores). The ES and SRM differ in the standardization of change. The ES divides mean change/difference between groups by the standard deviation of the control or whole group at baseline. The SRM divides the mean change in a subgroup by the standard deviation of change in that subgroup[16]. The RE is the square of the t-statistic in a measure divided by the t-statistic of a gold standard (although there is no recognized gold standard for quality of life)[20]. These responsiveness statistics give information about the power per given sample size, higher values indicate greater power or smaller required sample sizes to achieve a level of statistical power.

RESULTS

Search results. The search strategy identified 923 articles. Articles using the term "utility" in the sense of "usefulness," published in a foreign language, using a health status measure (without producing a utility value), or not relating to RA were removed (Figure 1). A total of 228 validation or methodological studies of utility measures, or economic evaluations reporting the source of utilities for QALY calculations in RA, were retained. A further 6 articles were identified through reference list checking. The total number of full articles detailing the validation or evidence consistent with studies validating utility in RA was 26 (Table 1).

The EQ-5D, HUI3, SF-6D, QWB, and HUI2 measures are the main generic utilities used in RA (Figure 2). These measures were named in the search strategy, which could explain their apparent dominance in RA. However, the terms were included to maximize the possibility of identifying all studies using utilities, and our results are consistent with other searches[10,22]. Utility measures with evidence satisfying the categories of the OMERACT filter are reviewed below.

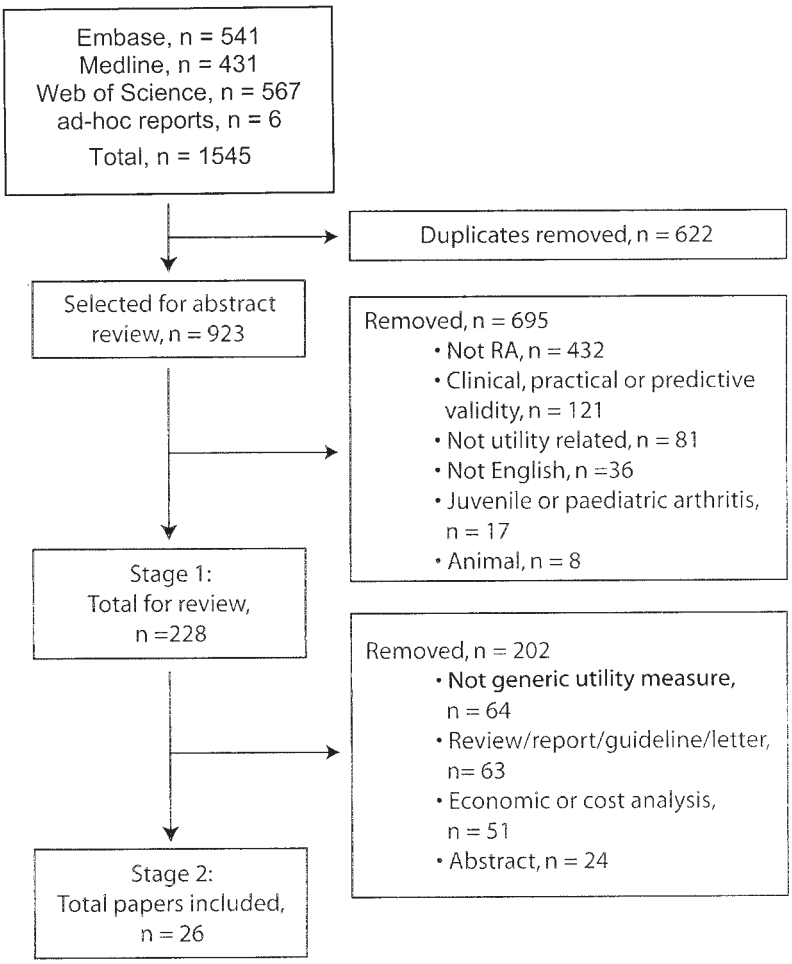

Figure 1. The literature search strategy.

Review

Introduction to measures. The EQ-5D is an extension of the well used and validated EuroQol health profile, comprising 5 questions with 3 levels of response[7]. Health states valuations were derived using a representative UK population sample (n = 3395; Table 1) using TTO methodology[23]. The valuations range from 1 to –0.59. Negative scores imply health states valued as being worse than death. US valuations are now available for the EQ-5D[24]. A supplementary visual analog scale asks the patient to rate their health today from best (100) to worst (0) on a 20 cm scale.

The Health Utilities Index (HUI) system of generic health profiles and preference-based utility measures currently comprises the HUI2 and HUI3[25]. The HUI2 consists of 7 domains that in combination describe 24,000 unique health states. The HUI3 has a greater number and independence of domains and more detailed descriptive ability (972,000 possible states)[8]. Preferences for the HUI2 and HUI3 were derived using RS valuations of single-deficit health states using a population sample from Hamilton, Ontario, Canada, and converted to SG utilities using a power function derived within the studies[26]. The HUI2 and HUI3 include negative valuations for health states rated as being worse than death.

The SF-6D is derived from the widely used and validated SF-36 generic health survey[6]. Eleven questions of the SF-36 are used to create the 6 domains of the SF-6D[27]. Health state valuations were derived using a representative sample of the UK population (n = 836) using the SG technique. The worst possible score on the SF-6D is 0.30, 30% of perfect health.

The Quality of Well-Being Scale (QWB) consists of 3 subscales covering aspects of function — mobility and social and physical. The subscales are converted into a scale from 0 (death) to 1 (asymptomatic/full function)[28], with 0.32 the lowest possible score for a living person. The weighting of QWB components for RA patients was derived using category scaling. About 850 members of a general population survey categorized their relative preference for the states combining function and symptoms[29,26]. A self-

Table 1. Summary of attributes and properties of the main utilities used in RA.

	EQ-5D	SF-6D	HU12	HU13	QWB-SA
Studies identified by	18 studies	7 studies	4 studies	7 studies	5 studies
Review (study)	34,39–43,45,46,52,53,55,60,66–70	34,42,43,54,59,60,69	34,42,43,69	34,42,43,46,54,56,69	28,29,37,57,58
Attributes/domains	5	6	6 (+1)	8	4
Levels	3	4–6	3–5	5–6	4–5
Health states	243	18,000	24,000	972,000	1215
Health states assessed	45	249	—	—	—
Sample size	3395	836	203	504	866
Country	UK	UK	Canada	Canada	USA
Utility method	TTO	SG	VAS-SG transformation	VAS-SG transformation	Category scaling
Assumptions	Additive	Additive	Additive	Multilinear	Additive
Range of scores	–0.59 to 1.00	0.30 to 1.00	–0.03 to 1.00	–0.36 to 1.00	0.32 to 1.00
Domains	Anxiety/depression Pain/discomfort Mobility Usual Activity Self-care	Social Pain Mental health Physical Role limitation Vitality	Sensory/communication Happiness (emotion) Pain/discomfort Learning/cognition Self-care Mobility (+ Fertility)	Vision Hearing Speech Happiness (emotion) Pain/discomfort Learning/cognition Ambulation Dexterity	Mobility Physical Activity Social Activity Physical symptom status (n = 58)
Period of recall for questions	1 day	4 weeks (1 month)/ typical day	General/1, 2 or 4 weeks	General/1, 2 or 4 weeks	3 days

TTO: time tradeoff, SG: standard gamble, VAS: visual analog scale.

assessed version (QWB-SA) is available, with specific weights that have a floor of 0.09[30].

Evidence

Feasibility. The administrative burden of the EQ-5D is low, and the questionnaire is simple and should take no longer than a few minutes to complete[31], while the HUI questionnaires take on average 5 minutes to complete[31]. The SF-36 is more comprehensive, taking about 9 minutes to complete[32], but time could be reduced by asking only the questions required to calculate the SF-6D or by using the SF-12, which estimates similar SF-6D values[33]. A study by Marra, *et al*[34] that included the EQ-5D, SF-6D, HUI-2, and HUI3 found that each measure had less than 4% of data missing, suggesting that these measures are feasible for self-administration. The interviewer-administered QWB version incurs considerable practical constraints. Training interviewers takes one to 2 weeks[35], and interviewing respondents takes 7[36] to 20 minutes[37]. The QWB-SA should take under 7 minutes to complete, and performed comparably with the interviewer-administered QWB in depressed patients[38].

The EQ-5D (www.euroqol.org) and the algorithm to calculate the SF-6D (http://www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d) scores are free to noncommercial users; however, a fee (quoted on request) may be payable to QualityMetric (www.qualitymetric.com) to use SF materials. The SF-36 is commonly collected, thus using the SF-6D may not accrue extra administrative or financial burden, and the algorithm can be applied retrospectively. Use of the HUI questionnaires involves a fee. The guide cost for use of one version is $4000 (http://www.healthutilities.com), with extra fees for additional information and support.

Truth. Distribution. The EQ-5D is the most simplistic of the utility measures, with 3-level scaling of the domains in the health profile (no, moderate, and severe problems), which is thought to limit descriptive ability[39-42]. Studies in RA patients have also shown a bimodal distribution of the EQ-5D (centered around 0 and 0.5–0.75), with few scores between 0.3 and 0.5[39-42], and a gap between 0.883 and 1[39]. The distribution of the HUI3 scores in RA patients is relatively continuous, suggesting potential for patients to progress between states as their health status changes[41,43].

Floor and ceiling effects. The SF-6D has a high floor, at 0.30, and although studies in RA have not shown grouping of scores at this level[41,42], this might limit the potential for use in patients with more severe RA. Floor effects in individual domains of the SF-6D, such as pain, may also exist. Ceiling effects of the EQ-5D in a range of healthcare conditions are well documented[26,44]. However, evidence regarding ceiling effects of the EQ-5D in RA is conflicting[41,42,45], suggesting that the effect may be less common in RA. Ceiling effects were shown to be considerable in 4 of the 5 domains of the EQ-5D (mobility 29%, self-care 57%, usual activities 21%, and anxiety/depression 45%) in patients with rheumatic disorders[39]. Similarly, 21% of patients in the study by Marra, *et al* were found to report no problems on all domains or no problems on 4 domains and some problems on one domain, although more than 97% of patients reported at least mild RA severity[42]. Studies investigating the distribution of the EQ-5D found no evidence of floor

effects[41,42,45], although Wolfe, *et al* reported that 18% of patients scored at the floor on the pain/discomfort domain and 5% scored at the floor of the usual activities and anxiety/depression domains[39,41].

Construct validity. The EQ-5D, SF-6D, HUI2, and HUI3 have been compared simultaneously in the same group of patients in RA[34,40,42,43] and musculoskeletal diseases dominated by RA (Table 2)[41,46]. The evidence of the comparability of scores from the EQ-5D, SF-6D, HUI2, and HUI3 in the same population is conflicting. Correlation of the EQ-5D, SF-6D, HUI2, and HUI3 in RA[34] and rheumatic disease[41,46] have mainly been moderate to strong (0.6 to 0.8; Table 3), although the generic utility measures differ in the domains they assess (Table 1). Luo, *et al*[46] found that although the HUI3 and EQ-5D group scores were similar, individual-level correlations were only moderate, suggesting that different but related aspects of RA were being measured. The HUI2 and HUI3 are influenced by cognition, while emotional and mental aspects of health are important in the EQ-5D and SF-6D[42,43]. Utilities will also differ because of the different methods of valuing health states and the different populations used to generate the valuations. Significant discrepancies between EQ-5D, HUI2, HUI3, and SF-6D utility scores have been found[41-43], particularly in groups with severe RA, despite good correlations between measures[40,42]. The mean scores for the EQ-5D and SF-6D were similar in studies by Russell, *et al*[40] using stable RA patients with mild disease, and Marra, *et al*[43] in a group of RA patients from routine clinical care (Table 2). However, in groups of patients with more severe RA[40] and musculoskeletal disease[41], the SF-6D provided a mean utility score higher than the EQ-5D and the HUI3. The comparability of mean utility values provided by the EQ-5D and HUI3 is unclear. In a group of clinically heterogeneous RA patients the mean EQ-5D utility score (0.66, SD 0.24) exceeded the mean HUI3 score (0.53, SD 0.29), whereas in 2 studies using patients with musculoskeletal disease (about 50% with RA) the EQ-5D and HUI3 provided comparable estimates[41,46]. This suggests that although utilities rank people similarly, differences in the techniques used to estimate utilities lead to different values. This has been demonstrated recently in RA[47], therefore utility measures should not be used interchangeably.

Marra, *et al*[43] demonstrated construct validity of 4 utility measures in correlations with RA–specific measures and against 8 hypotheses relating to severity of RA in a single

Table 2. Patient characteristics in validation studies of utility measures using patients with RA.

Study	n	Age, yrs, mean (SD)	Female, %	RA Duration, yrs	HAQ	EQ-5D, mean (SD)	SF-6D, mean (SD)	HU12, mean (SD)	HU13, mean (SD)	QWB, mean (SD)
Wolfe[39]	537	61.1 (13.8)*	83*	—	—	0.57 (0.25)*	—	—	—	—
Russell[40]	24	—	—	—	0.5 (0.4)	0.70 (0.18)	0.70 (0.13)	—	—	—
	60	—	—	—	1.3 (0.7)	0.43 (0.30)	0.55 (0.07)	—	—	—
Conner-Spady[41]	161**	55 (15.4)**	72**	—	—	0.49 (0.31)**	0.62 (0.14)**	—	0.50 (0.27)**	—
Marra[34,42,43]	313	61.5 (25.9)	78	13.9 (11.4)	1.1 (0.8)	0.66 (0.24)	0.63 (0.13)	0.71 (0.20)	0.53 (0.29)	—
Marra[54]	313	61.5 (25.9)	78	13.9 (11.4)	1.1 (0.8)	—	0.63 (0.13)	—	0.53 (0.29)	—
Hurst[52]	233	56 (14)	81	13 (13)	—	•	—	—	—	—
Hawthorne[51]	139	58.3 (12.5)	80	10.4 (9.1)	—	•	—	—	—	—
Luo[46]	114†	49 (16.4)†	82†	—	—	0.75 (0.21)†	—	—	0.76 (0.17)†	—
Harrison[55]	466	60.6 (11.2)	68	12.5 (6.7)	1.3 (0.7)	0.59 (0.22)	—	—	—	—
Frosch[28]	334	55.1 (16.1)	84	—	0.8 (0.7)	—	—	—	—	0.52 (0.13)

* Whole group (also includes 319 patients with osteoarthritis and 516 with fibromyalgia). ** Whole group (51% had RA). † Whole group (43% had RA). • Measure used but no summary statistic provided.

Table 3. Correlations between generic utility measures in RA.

	Statistic	EQ-5D	SF-6D	HU12	HU13
EQ-5D	Pearson	—	0.66–0.70[41]	—	0.68–0.69[41]
	Spearman	—	—	—	0.45–0.57[46]
	ICC	—	0.59[42]	0.68 [42]	0.66[42]
SF-6D	Pearson	0.66–0.70[41]	—	—	0.61–0.69[41]
	ICC	0.59[42]	—	0.66[42]	0.56[42]
HU12	ICC	0.68[42]	0.66[42]	—	0.79[42]
HU13	Pearson	0.68–0.69[41]	0.61–0.69[41]	—	—
	Spearman	0.45–0.57[46]	—	—	—
	ICC	0.66[42]	0.56[42]	0.79[42]	—
RAQoL	Spearman	0.70[43]	0.80[43]	0.70[43]	0.75[43]

ICC: intraclass correlation coefficient.

group of patients. All measures correlated significantly and strongly (≥ 0.55) with the Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL), pain and patient global visual analog scales, and the Stanford Health Assessment Questionnaire (HAQ)[48], and conformed to at least 6 hypotheses. The EQ-5D, SF-6D, HUI-2, and HUI-3 correlated with self-reported RA severity (≥ 0.45) and control (≥ 0.49) at least as strongly as the disease-specific HAQ and RAQoL[43]. The HUI measures were able to detect differences between patients hospitalized or not in the previous year, whereas the EQ-5D and SF-6D could not do this, but HUI measures were not able to distinguish differences in utility in those having adverse events to drug therapy in the previous 3 months that were identified by the EQ-5D and SF-6D[43].

The HUI2 and HUI3 predominantly measure functional ability and pain[42,43,46,49]. However, the HUI3 weights cognition more heavily than the HUI2, which in contrast weights pain more heavily. The HUI3 also has a greater range of scores for states worse than death. In mild states, the HUI2 and HUI-3 should provide similar values, whereas in conditions with high morbidity and cognitive impairment, values may differ markedly[50]. The EQ-5D and SF-6D scores were predominantly explained by functional ability, and supplemented by aspects of mental and emotional health[42].

The QWB measure has not been compared to other generic utility measures in RA, but demonstrated consistent and significant (although moderate at best) linear relationships with disease-specific measures of physical function (0.48 to 0.55), pain (> 0.40), and swelling (0.28) in a study comparing patients visiting clinics for musculoskeletal diseases (> 50% RA) with those attending for family medicine[28]. The QWB-SA explained roughly 50% of the variance in the HAQ in US patients with RA[28].

The EQ-5D has demonstrated construct validity in distinguishing between patients defined by self-reported health status[51], functional disability (HAQ)[52,53], socioeconomic status (SES)[52,54,55], social support[52], employment status[52], and above/below median SF-36 subscale scores[46]. The HUI3 was able to differentiate between healthy individuals and those reported as having RA and/or stroke, and scores were lower in those with lower education and with higher comorbidity[56]. Elsewhere, patients with lower SES measured by family income had lower HUI3 scores[54]. HUI3 scores correspond to groups defined by above or below median SF-36 subscale scores and tender joint assessments[46].

Patients attending musculoskeletal disease clinics scored significantly worse (about 10% lower) on the QWB-SA than patients at family medicine clinics, after adjustment for differences in age, sex, education, and ethnicity. The severity and type of conditions treated in family medicine clinics were unknown, although the implication is that these were milder cases[28]. A derivation of the QWB was able to detect significant reduction of health-related quality of life[57] and lost utility/QALY due to self-reported arthritis in the US National Health Institute Survey between 1980 and 1994[58].

Discrimination. Reliability. The SF-6D and HUI3 have consistently demonstrated good reliability (ICC from 0.72 to 0.89; Table 4), while the EQ-5D was less reliable (ICC 0.46–0.66) in studies using the same patients[34,40,41,46]. The lowest ICC for the EQ-5D (0.46) was calculated over a 5-week test-retest period, although the HUI2 (ICC 0.77), HUI3 (ICC 0.81), and SF-6D (ICC 0.89) were very reliable over the same period[34]. Hurst, *et al*[52] reported acceptable reliability for the EQ-5D over 2-week (ICC 0.78) and 3-month (ICC 0.73) periods. Alternative tests of reliability based on correlations of test-retest scores in patients reporting no change in health found the EQ-5D, SF-6D, and HUI3 to be highly stable over 3 months (0.81–0.88) and 1 year (0.72–0.86) periods[41].

Minimum important difference. The MID for the EQ-5D[43], SF-6D[43,59], and HUI2 and HUI3[43] measures were 3%–5% of the range of possible values, suggesting that the minimum difference is consistent and reasonable across measures. The MID for the EQ-5D, HUI2, and HUI3 were estimated as roughly 0.05, 0.04, and 0.06–0.07, respectively, by 2 alternative methods, assuming equal MID for improvement and deterioration[43]. The estimated MID for the SF-6D in RA was roughly 0.03 to 0.04 in studies in the UK and Canada[43,59,60]. Walters, *et al* used 11 patient groups from 8 longitudinal studies from a range of conditions (one of which was early RA), finding mean MID of 0.07 for the EQ-5D and 0.04 for the SF-6D, although the estimate for the EQ-5D in the early RA group was 0.13[60]. These studies assumed that improvement and deterioration have a uniform effect, and combined them in the same estimate.

Responsiveness. The EQ-5D, SF-6D, and HUI measures have demonstrated the ability to detect both patient-reported improvement and deterioration over periods of up to one year[34,41]. The EQ-5D was consistently the most responsive measure in detecting deterioration in health and was almost twice as responsive as alternative measures (Table 5)[34,40,41]. Responsiveness in improving patients offers conflicting data. Russell, *et al*, assessing response to treatment over 14 weeks, found the SF-6D was the most responsive[40]. Marra's recommendations[34] varied according to the definition of change and responsiveness statistic. Using self-reported change, the SF-6D and HUI2 appeared most responsive, while according to change in patient global assessment, the HUI3 was most responsive, and the RE indicated that the HUI2 was the most efficient measure by either definition[34]. Only Conner-Spady and Suarez-Almazor reported that the same measure was most responsive in detecting both improvement and deterioration, the EQ-5D[41]. The SF-6D shows relatively small absolute change but has a small stan-

Table 4. Summary of evidence of comparitive reliability of generic utilities in RA. Figures in bold type exceed threshold of 0.7; figures in italics denote most reliable measure.

Study	n	Test statistic	Test-retest	EQ-5D	SF-6D	HU12	HU13
Marra[42]	50	ICC(1)*	5 wks	0.46	***0.89***	**0.77**	**0.81**
Russell[40]	24	ICC(U)*	3 wks	0.66	***0.72***	—	—
Luo[46]	94	ICC(U)*	1 wk	0.64			***0.75***
Conner-Spady[41]	46	Stability coefficient	3 mo	**0.81**	***0.87***	**—**	***0.88***
Conner-Spady[41]	98	Stability coefficient	1 yr	**0.74**	***0.86***	**—**	**0.72**
Hurst[52]	93	ICC(2)*	3 mo	**0.73**	—	—	—
Hurst[52]	31	ICC(U)*	2 wks	**0.78**	—	—	—
Hawthorne[51]	51	Spearman	2 wks	**0.74**	—	—	—

ICC: mixed effects, subject random, instrument fixed. ICC(2): simple linear model (Fleiss[14]). ICC(U): unspecified methodology. * Preferred methodology for assessing reliability.

Table 5. Summary of evidence of comparative responsiveness of generic utilities in RA according to direction of change. Figures in bold type denote the most responsive measure.

	Responsiveness Statistics											
	Effect Size[a]				Standardized Response Mean[b]				Relative Efficiency[c]			
Definitions	EQ-5D	SF-6D	HU12	HU13	EQ-5D	SF-6D	HU12	HU13	EQ-5D	SF-6D	HU12	HU13
Improvement												
Treatment*[40]	0.67	**1.40**	—	—	0.64	**0.87**	—	—	—	—	—	—
Self-report*[41]	**0.53**	0.36	—	0.30	—	—	—	—	—	—	—	—
Self-report*[34]	0.15	**0.31**	**0.30**	0.23	0.20	0.36	**0.40**	0.29	0.24	0.52	**0.72**	0.39
Patient global*[34]	0.36	0.54	0.49	**0.60**	0.43	0.62	0.52	**0.73**	0.61	0.90	**1.02**	0.74
Deterioration												
Self-report*[41]	**–0.58**	–0.24	—	–0.17	—	—	—	—				
Self-report*[34]	**–0.16**	–0.08	–0.14	–0.10	**–0.19**	–0.13	–0.16	–0.12	**0.73**	0.21	0.25	0.12
Patient global*[34]	**–0.55**	–0.24	–0.33	–0.36	**–0.63**	–0.35	–0.44	–0.46	**1.14**	0.62	0.82	0.78

[a] Mean difference divided by standard deviation at baseline; [b] mean difference divided by standard deviation of difference; [c] $({}^{t}\text{statistic}_1/{}^{t}\text{statistic}_2)^2$ where ${}^{t}\text{statistic}_1$ = alternative measure, ${}^{t}\text{statistic}_2$ = gold standard (in this case defined as RAQoL). * Indicates assessment of the importance of change.

dard deviation[40,41], leading to a good effect size in longitudinal studies or studies assessing change/differences.

No studies were found that demonstrated the reliability or responsiveness to change of the QWB in RA.

DISCUSSION

This review of preference-based utility measures in RA has demonstrated that each has a number of strengths, weaknesses, and inconsistencies in their performance and validity (Table 6). The EQ-5D is the most commonly used and extensively validated measure, while the HUI3 is gaining greater acceptance and there is growing positive evidence of its ability to describe, measure, and detect change in the condition. The SF-6D appears to have potential for use in RA patients, although this may be restricted to those with milder disease because of its high floor and preliminary evidence of greater sensitivity in these patients, indicated by the small MID. The questions used to calculate the SF-6D tend to have greater descriptive emphasis on the milder states; 5 of the questions used to calculate the SF-6D have 6 scoring levels and these use roughly two-thirds of the levels to describe the differences between "no" and "moderate" problems — these include all questions regarding the pain, vitality, social, and mental health domains. The SF-6D is also generating interest due to its practicality, allowing utility scores to be estimated in studies that have included the SF-36 or SF-12 questionnaires. However, the SF-6D needs further validation in all aspects of its performance, to confirm whether it truly does have potential for use in RA or whether its properties make it too unreliable. Use of the QWB, and to a lesser extent the HUI2, appears to be in decline. The lack of use of the HUI2 may be due to preference for the HUI3, and QWB because of its preferred but administratively inhibiting interviewer-administered method. The QWB-SA questionnaire has addressed the issue of administrative burden, which may lead to renewed appeal for use in RA.

Although all the instruments aim to measure a societal valuation of health-related quality of life, each measures subtly different aspects and differing levels of detail. The utilities derived from the different measures also vary, leading to large variations in cost-utility analysis conclusions according to the utility measure used[47]. The variation may possibly be due to the differences in the range of scores available, the distributional properties, the domains that are measured and how they are influenced by RA, the recall

Table 6. Summary of evidence of the main utilities in RA using the OMERACT filter.

OMERACT Filter	EQ-5D	SF-6D	HU12	HU13	QWB-SA
Feasibility					
Time taken	†Few minutes[31]	9 min[32]	†5 min[31]	†5 min[31]	< 7 min[38]
Cost	†Free (noncommercial)	†Free (noncommercial)*	††	††	
Completion	†< 4% missing[43]	†< 4% missing[43]	†< 4% missing[43]	†< 4% missing[43]	—
Truth					
Ceiling:utility	†1%[41]	†0%[41]	—	†< 1%[41]	—
Ceiling:domains	††21% within domains[42]	†††Not tested		†††Not tested	
Floor:utility	†0%[41]	†0%[41]	—	†0%[41]	—
Floor:domain	††18% pain/discomfort[39]	†††Not tested		†††Not tested	
Distribution	††Bimodal[39,41,42] ††Gaps[39,41]	†Normal[41,42]	†Continuous[42]	†Continuous[41,42]	—
Disease-specific correlations	†Moderate to strong	†Moderate to strong	†Moderate to strong	†Moderate to strong	†††Poor to moderate
Factor analysis[42]	Functional ability/pain Emotional/mental health	Functional ability/pain Emotional/mental health	Functional ability/pain Cognition	Functional ability/pain Cognition	—
Distinction	†75%[43,46]–100%[55]	†75%[43]–100%[54]	†88%[43]	†80%[43]–100%[54,56]	†100%[28,57,58]
Distinction					
Reliability (ICC)	††0.46[42]–†0.78[51]	†0.72[40]–0.89[42]	†0.77[42]	†0.75[46]–0.81[42]	—
MID (anchor-based)	0.07–0.13[40]	0.03–0.04[59,60]			
MID (distribution-based)	0.05[43]	0.03[43]	0.04[43]	0.06–0.07[43]	—
Responsiveness: improving	†Responsive	†Responsive	†Responsive	†Responsive	
Responsiveness: deterioration	†Most responsive	†††Possibly least responsive	†Responsive	†Responsive	

* Algorithm; † support; †† issues; ††† inconclusive. ICC: intraclass correlation coefficient, MIP: minimum important difference.

period, and the method of valuation and modeling. All seem to measure crucial aspects of RA such as functional disability and pain. The EQ-5D concisely assesses important aspects of health with minimal administrative burden, but may miss important aspects of energy/fatigue, mental health, and cognition. The HUI measures assess cognition, but do not measure the impact of disease on the patient's social life. The differing period of recall between measures, which ranges from 1 day (EQ-5D) to 4 weeks (SF-6D) (Table 1), may also be an important issue leading to different utility scores, although the influence of this is unclear. If patients focus solely on the day on which they complete the questionnaire (e.g., the EQ-5D), the period of recall is minimal and unambiguous; however, the possibility exists that a patient may be having a good or bad day, with responses temporarily influenced by the presence or lack of symptoms for whatever reason. These properties of the EQ-5D may affect its apparent reliability. On the other hand, measures considering a much greater period of time (e.g., 1 month on the SF-6D) allow ambiguity and inaccuracy in assessment of health. It is unclear whether the patient is able to accurately recall health over an extended period of time, and whether responses are influenced by their current health state, the best or worst health state during that period, or some approximation of their "average" health throughout the period.

The EQ-5D appears to be responsive to change in the patient with RA, particularly where the patient's health is getting worse. In this situation, the EQ-5D was consistently the most responsive measure; however, in patients who were improving there was no conclusive evidence to favor any measure over another. The difference in responsiveness by direction of change and the distributional properties of the measures, in particular the EQ-5D, suggest that the assumption of equal MID in improving and worsening patients warrants further study. Despite concerns about the scaling of the EQ-5D potentially limiting its scope to detect change[39], the EQ-5D was able to detect improvement[40] and deterioration in RA patients[34,41]. The non-normal distribution of the EQ-5D is problematic for cross-sectional analysis using techniques that assume normal distribution of data.

A number of the utility measures reviewed have limitations with their scoring properties, which restrict the ability of these measures in more extreme patient samples. The EQ-5D[26,44] and HUI3[26] have been reported to have ceiling effects, while the SF-6D and QWB have high floors[26]. Evidence of the ceiling effects in RA of the HUI3 is absent, and evidence regarding the EQ-5D is mixed. However, QALY gains from RA interventions, obtained using utilities from these measures, will be compared against QALY gains from interventions in other fields. Therefore the ceiling effects remain a limitation and these measures should be used in mild health states with caution. Similarly, the high floor of the SF-6D and QWB may lead to overestimated utility values for severe health states. The validity of the SF-6D

preference weights may also be limited both by the low number of states in relation to the number of potential health states and by the relatively small sample of participant raters.

Of the measures reviewed, only the EQ-5D, HUI2, and HUI3 include negative valuations (state worse than death). However, evidence suggests that most individuals perceive some health states to be genuinely worse than death, for example, coma, or conditions characterized by chronic pain, severe physical dysfunction (e.g., confinement to bed), or mental dysfunction (e.g., inability to reason or communicate)[61,62]. Assuming worse-than-death valuations are valid, then the EQ-5, HUI2, and HUI3 potentially represent a broader range of severe or very poor health states than the SF-6D and QWB. Evidence suggests that worse-than-death valuations may be valid; health states rated as equivalent to death still have potential for improvement or deterioration[63]. Excluding negative valuations may lose important information and underestimate benefits of treatment[61,63], which may be exaggerated in chronic conditions such as RA due to the lengthy time horizon[63]. However, problems with scaling exist, as there are no limits to negative scores and lower limits using TTO methodology as low as –39 have been reported[64]. Consequently, transformations are often used to restrict the lower bound to –1[64,65]. Transformed values are not strictly utilities, although they are often used as such in aggregated mean valuations[64,65].

Although a number of studies in RA using the EQ-5D and HUI3 have included patients occupying states worse than death[39-41,43], the health status of these patients has not yet been investigated.

A limitation of this review is that the methodology will favor the more popular utility measures. However, proof of validity and, to a lesser extent, responsiveness relies on an accumulation of evidence rather than definition by one study. This review has highlighted limitations of each of the currently used measures, which might be addressed by measures yet to be employed in the setting of RA. A possible further limitation of the search strategy was the focus on studies with primary evidence about the validity and performance of utility measures in RA. This led to the exclusion of economic or cost analyses that used modeling techniques to synthesize secondary data rather than analyzing directly-observed data. It is less likely that these studies could have contributed extra information about the performance of utility measures in practice.

There is no conclusive evidence to date as to which measure is the best for use in RA. A number of factors require consideration when choosing an instrument, including the severity of RA in the study sample, resources, patient burden, and the country the utility weights were derived in, as well as evidence of truth, validity, and feasibility discussed here. Considering this last issue, the results of this review suggest that the HUI3 and the EQ-5D have the most supportive evidence, particularly in moderate to severe RA. The EQ-5D does have a number of issues related to its distribution and scaling, yet it appears to have construct validity and to perform well longitudinally in RA patients. The HUI3 has performed well to date and has supportive evidence in all categories of the OMERACT filter, but needs more RA-specific validation to confirm its validity. The EQ-5D has construct validity; however, the problems of ceiling effects and its scaling (no intermediate states between "no problems" and "moderate problems") may limit its descriptive ability and response to change in patients with mild RA. This might mean it is more suitable for use in patients with more severe disease. The SF-6D could be a promising measure; current evidence suggests it may be more suitable for patients with mild RA due to its high floor; however, further work is needed to identify its full potential in RA.

Our review identified the HUI3 and EQ-5D preference-based measures as having the most evidence supporting their validity and responsiveness in the use of patients with RA. However, limitations were identified in both instruments. While less well researched, the SF-6D appears to have potential for use in studies of patients with milder disease. In order to develop understanding of the relevant merits and drawbacks of each measure, more head-to-head comparisons of the measures are required in longitudinal studies across the spectrum of RA disease severity. The lack of interchangeability of measures and recent evidence emphasizing the influence of utility valuations from alternative measures on cost-effectiveness conclusions suggests there is an urgent need to work toward achieving consensus on a single utility measure for use in economic evaluation.

ACKNOWLEDGMENT

We are grateful to the referees for their comments and advice in clarifying issues covered in this review, which we feel improved the report.

REFERENCES

1. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford: Oxford University Press; 2005.
2. National Institute for Clinical Excellence. A guide to NICE. London: National Institute for Clinical Excellence; 2005.
3. Ontario Ministry of Health and Long-Term Care. Ontario guidelines for economic analysis of pharmaceutical products. Internet. 2004. Government of Ontario Central Web Site. http:www.health.gov.on.ca/english/providers/pub/drugs/economic/economic-mn.html. Accessed January 31, 2007.
4. Sullivan SD, Lyles A, Luce B, Grigar J. AMCP guidance for submission of clinical and economic evaluation data to support formulary listing in US health plans and pharmacy benefits management organizations. J Manag Care Pharm 2001;7:272-82.
5. Patrick DL, Erickson P. Health status and health policy. New York: Oxford University Press; 1993.
6. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-form Health Survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
7. The EuroQol Group. EuroQol — a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.

8. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health Qual Life Outcomes 2003;1:54.
9. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for outcome measures in rheumatology. J Rheumatol 1998;25:198-9.
10. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: Bibliographic study of patient assessed health outcome measures. Br Med J 2002;324:1417-9.
11. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979-92.
12. McHorney CA, Tarlov A. Individual-patient monitoring in clinical practice: Are available health status measures adequate? Qual Life Res 1995;4:293-307.
13. Bowling A. Measuring health: a review of quality of life measurement scales. Philadelphia: Open University Press; 1993.
14. Fleiss JL. Reliability of measurement. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986:1-32.
15. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford: Oxford University Press; 2003.
16. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Qual Life Res 2003;12:349-62.
17. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ 1996;15:209-31.
18. Hays RD, Revicki DA. Reliability and validity (including responsiveness). In: Fayers P, Hays RD, editors. Assessing quality of life in clinical trials. Oxford: Oxford University Press; 2005:25-39.
19. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42:403-8.
20. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12:142S-158S.
21. Veenhof C, Bijlsma JWJ, van den Ende CHM, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: A systematic review of the literature. Arthritis Care Res 2006;55:480-92.
22. Brauer CA, Rosen AB, Greenberg D, Neumann PJ. Trends in the measurement of health utilities in published cost-utility analyses. Value Health 2006;9:213-8.
23. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. York: Centre for Health Economics, University of York; 1995.
24. 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.
25. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res 1982;30:1043-69.
26. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003;56:317-25.
27. 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.
28. Frosch DL, Kaplan RM, Ganiats TG, Groessl EJ, Sieber WJ, Weisman MH. Validity of self-administered quality of well-being scale in musculoskeletal disease. Arthritis Rheum 2004;51:28-33.
29. Balaban DJ, Sagi PC, Goldfarb NI, Nettler S. Weights for scoring the quality of well-being instrument among rheumatoid arthritics. A comparison to general population weights. Med Care 1986;24:973-80.
30. Sieber WJ, Groessl EJ, David KM, Ganiats TG, Kaplan RM. Quality of well-being scale self-administered (QWB-SA) scale: User's manual. San Diego: University of California, San Diego, Health Outcomes Assessment Program; 2004.
31. Brazier J, Deverill M, Green C. A review of the use of health status measures in economic evaluation. J Health Serv Res Policy 1999;4:174-84.
32. Kelly S, Jessop EG. A comparison of measures of disability and health status in people with physical disabilities undergoing vocational rehabilitation. J Public Health Med 1996;18:169-74.
33. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851-9.
34. Marra CA, Rashidi AA, Guh D, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333-44.
35. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes: comparisons of assessment methods. Med Decis Making 1984;4:315-29.
36. Kaplan RM. Using quality-of-life information to set priorities in health-policy. Social Indicators Research 1994;33:121-63.
37. Bombardier C, Raboud J. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. The Auranofin Cooperating Group. Control Clin Trials 1991;12:243S-256S.
38. Pyne JM, Sieber WJ, David K, Kaplan RM, Hyman RM, Keith WD. Use of the quality of well-being self-administered version (QWB-SA) in assessing health-related quality of life in depressed patients. J Affect Disord 2003;76:237-47.
39. Wolfe F, Hawley DJ. Measurement of the quality of life in rheumatic disorders using the EuroQol. Br J Rheumatol 1997;36:786-93.
40. Russell AS, Conner-Spady B, Mintz A, Mallon C, Maksymowych WP. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. J Rheumatol 2003;30:941-7.
41. Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Med Care 2003;41:791-801.
42. Marra CA, Esdaile JM, Guh D, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care 2004;42:1125-31.
43. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ5D) and disease specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571-82.
44. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17:13-35.
45. Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H. Validity of Euroqol — a generic health status instrument — in patients with rheumatoid arthritis. Economic and Health Outcomes Research Group. Br J Rheumatol 1994;33:655-62.
46. Luo N, Chew LH, Fong KY, et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. J Rheumatol 2003;30:2268-74.
47. Marra CA, Marion SA, Guh DP, et al. Not all "quality-adjusted life years" are equal. J Clin Epidemiol 2007;60:616-24.
48. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
49. Bansback NJ, Regier DA, Ara R, et al. An overview of economic evaluations for drugs used in rheumatoid arthritis: focus on tumour necrosis factor-alpha antagonists. Drugs 2005;65:473-96.
50. Feeny DH, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the Health Utilities Index Mark

3 system. Med Care 2002;40:113-28.
51. Hawthorne G, Buchbinder R, Defina J. Functional status and health-related quality of life assessment in patients with rheumatoid arthritis. Melbourne, Australia: Centre for Health Program Evaluation; 2000.
52. 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;36:551-9.
53. Kobelt G, Eberhardt K, Jonsson L, Jonsson B. Economic consequences of the progression of rheumatoid arthritis in Sweden. Arthritis Rheum 1999;42:347-56.
54. Marra C, Lynd LD, Esdaile JM, Kopec JA, Anis AH. The impact of low income on self-reported health outcomes in patients with rheumatoid arthritis within a publicly funded health-care environment. Rheumatology Oxford 2004;43:1390-7.
55. Harrison MJ, Tricker KJ, Davies L, et al. The relationship between social deprivation, disease outcome measures, and response to treatment in patients with stable, long-standing rheumatoid arthritis. J Rheumatol 2005;32:2330-6.
56. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000;38:290-9.
57. Anderson JP, Kaplan RM, Ake CF. Arthritis impact on US life quality: Morbidity and mortality effects from National Health Interview Survey data 1986-1988 and 1994 using QBXW1 estimates of well-being. Social Indicators Research 2004;69:67-91.
58. Kaplan RM, Alcaraz JE, Anderson JP, Weisman M. Quality-adjusted life years lost to arthritis: effects of gender, race, and social class. Arthritis Care Res 1996;9:473-82.
59. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4-12.
60. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523-32.
61. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. Med Decis Making 1994;14:9-18.
62. Lockhart LK, Ditto PH, Danks JH, Coppola KM, Smucker WD. The stability of older adults' judgments of fates better and worse than death. Death Stud 2001;25:299-317.
63. Franic DM, Pathak DS. Effect of including (versus excluding) fates worse than death on utility measurement. Int J Technol Assess Health Care 2003;19:347-61.
64. Robinson A, Spencer A. Exploring challenges to TTO utilities: Valuing states worse than dead. Health Econ 2006;15:393-402.
65. Torrance GW. Measurement of health state utilities for economic appraisal: A review. J Health Econ 1986;5:1-30.
66. Suarez-Almazor ME, Conner-Spady B. Rating of arthritis health states by patients, physicians, and the general public. Implications for cost-utility analyses. J Rheumatol 2001;28:648-56.
67. Taylor WJ, Lord S, McPherson KM, McNaughton HK. EuroQol EQ-5D may not adequately describe the health of people with disabilities. Disabil Rehabil 2001;23:281-5.
68. Kobelt G, Eberhardt K, Geborek P. TNF inhibitors in the treatment of rheumatoid arthritis in clinical practice: costs and outcomes in a follow up study of patients with RA treated with etanercept or infliximab in southern Sweden. Ann Rheum Dis 2004;63:4-10.
69. 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.
70. Kobelt G, Lindgren P, Young A. Modelling the costs and effects of leflunomide in rheumatoid arthritis. Eur J Health Econ 2002;3:180-7.