The Responsiveness of Generic Health Status Measures As Assessed in Patients with Rheumatoid Arthritis Receiving Infliximab

ANTHONY S. RUSSELL, BARBARA CONNER-SPADY, ALISA MINTZ, CATHERINE MALLON, and WALTER P. MAKSYMOWYCH

ABSTRACT. ***Objective.*** We used a variety of health status measures in 2 groups of patients with rheumatoid arthritis (RA) to assess both the smallest distinguishable difference and the relative responsiveness to change of these measures, when used in clinical practice.

Methods. Two groups of patients were studied. Group 1: 24 patients with stable RA tested on 2 occasions; Group 2: 60 patients receiving methotrexate tested before and 14 weeks after treatment with infliximab. Assessments were made with self-completed questionnaires: the modified Health Assessment Questionnaire, Medical Outcomes Study Short Form-36 [SF-36 (SF-6D)], EuroQol, and, in some, the standard gamble. Group 2 also had joint counts, and measures of erythrocyte sedimentation rate, C-reactive protein, and hemoglobin.

Results. The limits-of-agreement (Bland-Altman) approach had greater confidence intervals (CI) than did CI based on ± 2 standard errors of the measurement. Improvement with infliximab could be determined with all measures, however, but the standard gamble seemed least responsive to change.

Conclusion. The various measures had different degrees of responsiveness, but with all it was possible to show improvement in Group 2 compared to Group 1. There was a closer association of the patient centered measures of improvement with changes in pain score than with joint counts. (J Rheumatol 2003;30:941–7)

Key Indexing Terms:
QUALITY OF LIFE RHEUMATOID ARTHRITIS INFLIXIMAB HEALTH STATUS

Subjective patient based health outcome measures are now widely used in rheumatology, partly because they reflect aspects of disease that are most important to the patient, but also because they have been validated and appear to provide outcome data at least as reliable as the more traditional physician reported data such as joint counts[1,2]. Some are disease-specific and some generic. The modified Health Assessment Questionnaire (MHAQ) has been widely used to assess functional status in patients with rheumatic diseases[3-5], while the Medical Outcomes Study Short Form 36 (SF-36) is an often-used generic health status measure[3,4,6-11]. These generic measures have become critical to assess and describe cost effectiveness, and are often known as measures of health related quality of life, HRQOL, or more simply QOL. They allow comparison of outcomes across disease states and can be used in policy decisions. A further development introduced by medical technology assessment has increased the demand for preference based health status measures for application in cost utility analysis[12,13]. Attaching economic value to quality of life rests on game theory, so that a patient preference based measure of "utility" of health states is intrinsic to an assessment of life quality[12-14]. Other measures have been devised to simulate this, but the standard gamble remains the classic format[12]. Here, a patient is asked to choose between their current health state and a lottery that offers, say, a 90% chance of complete health, but a 10% chance of immediate death. The odds may be changed in either direction to find an equivalence point between 0 and 1. Other techniques of utility measurement are the time tradeoff (not used here directly)[15] and the rating (or feeling) thermometer. It is important to know if treatment interventions that have been shown to be efficacious using conventional approaches result in meaningful improvement as reflected by generic quality of life measures[8,4].

The SF-36 has been validated in healthy and diseased populations[6] and in patients with rheumatoid arthritis (RA)[7,9], although one report suggested it was not as sensitive to change in patients with RA as other similar, generic measures[16], and we found it insensitive in revealing improvement after isolated treatment of carpal tunnel syndrome in patients with RA[17]. Although the SF-36 was not designed specifically for use in economic evaluation, a

From the Rheumatic Disease Unit and Department of Psychology, University of Alberta, Edmonton, Alberta Canada.

A.S. Russell, FRCPC, Professor of Medicine; B. Conner-Spady, PhD; A. Mintz, MSc, Department of Psychology; C. Mallon, RN; W.P. Maksymowych, FRCPC, Professor of Medicine.

Address reprint requests to Dr. A.S. Russell, Room 562, Heritage Medical Research Centre, University of Alberta, Edmonton, Alberta, Canada T6G 2S2.

Submitted June 17, 2002; revision accepted October 31, 2002.

preference based algorithm for the SF-36 has recently been derived to form the SF-6D[18]. The EuroQol (EQ-5D) is a widely used preference based generic tool designed for evaluating health and can be used in cost-utility analysis[16,19]. There are many intrinsic assumptions, especially in calculations of derivatives based on these, i.e., quality adjusted life years (QALY), that are still controversial[13,20]. However, these measures are now used in clinical trials, although not yet routinely in clinical practice.

These generic measures do not assess functional changes directly, but rather how these changes affect a patient in their daily life[3,13,21]. We assessed the variability of some of these measures in patients with clinically stable RA, and compared this to the changes seen in patients successfully treated with an anti-tumor necrosis factor agent, infliximab. This should allow us, under optimal practice conditions, to assess the smallest detectable difference (SDD) and to see if improvement is adequately reflected in QOL or utility measures in clinical practice[22].

MATERIALS AND METHODS

Patients. The samples consisted of 2 groups of patients attending a rheumatology clinic at the University of Alberta between 1999 and 2001. Group 1 included clinically stable patients taking gold or methotrexate (MTX) treatment attending the clinic for blood test monitoring. All had English as their first language. Patients volunteered as a result of a notice that was displayed. This group was used to estimate the test-retest reliability for each outcome measure at 2 consecutive visits roughly 3 weeks apart. Prior to the second assessment it was confirmed that no change in therapy had been introduced. This factor was used as a pragmatic definition of both stable and acceptable disease control. Group 2 comprised patients with RA, all of whom had been taking MTX for at least 6 months, but who were to start taking infliximab because of persistent active disease. They were assessed before treatment and at 14 weeks of followup, at the time of their fourth infusion (3 mg/kg rounded up to the nearest 100 mg). Face-to-face interviews by an experienced interviewer were used to collect standard gamble utility scores using the rotating wheel chance board (Supplied by R. Zazulak, Health Utilities Inc., Toronto, Ontario, Canada) as described[12]. Questionnaires assessing the following outcome measures were completed by the patients while they attended the clinic. This group of patients therefore had a level of disease activity, deemed unacceptably high, at their first assessment.

Pain. Self-rated pain over the past week was measured on a 100 mm visual analog scale (VAS) with anchors of 0 "no pain" and 100 "pain as bad as it could be."

Health status. The SF-36 consists of 36 items. Eight subscale scores are derived from the summation of item scores and transformed to a 0 to 100 scale, with higher numbers representing better health. The physical health subscales include physical functioning, role physical, bodily pain, and general health. The mental health subscales include vitality, social functioning, role emotional, and mental health. The SF-36 also includes 2 aggregate scores, the physical component summary (PCS) and the mental component summary (MCS)[6,23].

The SF-6D is a 6 dimensional health status classification system based on 11 items from the SF-36[24]. Health states were valued using standard gamble from a representative sample of respondents from the UK[24]. The summative weighted score, the SF-6D index, ranges from 0.26 to 1.0. The EQ-5D consists of 5 items graded 1 (no problem), 2 (some or moderate problems), or 3 (unable or extreme problems). Weights were derived using time tradeoff methods[19,25]. The sum of the weighted items results in a single summary score, the EQ-5D index, which ranges from –0.59 to 1. Both the SF-6D and the EQ-5D provide a measure of health status to which full health is assigned a value of 1 and death a value of 0. The EQ-5D self-rated "thermometer" is a visual analog scale (EQ-VAS) indicating a patient's own assessment of their health state. It ranges from 0 (worst imaginable health state) to 100 (best imaginable health state).

Functional status. The MHAQ consists of questions measuring dressing, arising, eating, walking, hygiene, reach, grip, and activities. Scores range from 0 (no difficulty) to 3 (unable to do). The summary score, the mean of all of the items, ranges from 0 (no difficulty) to 3 (unable to do).

Other measures collected for Group 2 included the number of swollen and tender joints, C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and hemoglobin (Hb).

Statistical analysis. Test-retest reliability was estimated for each tool using the intraclass correlation coefficient (ICC). The ICC combines information from the relative ordering of different cases as well as mean differences. The ICC ranges from 0 to 1, a higher ICC indicating better reproducibility.

Responsiveness was defined[26] as "the ability of an instrument to accurately detect change when it has occurred." It was assessed for each health measure using the paired t test, effect size, and standardized response mean (SRM). Effect size was calculated by dividing the mean difference by the standard deviation (SD) at baseline. The SRM was calculated by dividing the mean difference by the SD of difference scores[27]. As well as assessing responsiveness at the group level, it is also important to assess the SDD at the individual level[28]. The standard error of measurement (SEM) is a function of both the reliability of a score and the SD of scores and gives us an indication of how much we would expect an individual's score to vary from one occasion to another when there is no actual change in clinical function or health status. It was computed by taking the square root of 1 minus the reliability, and multiplying the results by the SD of baseline scores for the stable group. A 95% confidence interval (CI) using 2 SEM was used to define the lower boundaries of potentially meaningful change for individuals[29]. The Bland-Altman limits of agreement have also been used to estimate minimum boundaries for meaningful change[28,30]. Ninety-five percent limits of agreement based on the stable group (1.96 × SD of difference score) were used to define a lower boundary for the SDD. The percentage of cases that improved for each outcome measure was compared using each method.

To test the ability of each measure to discriminate between groups, mean changes in health outcomes were compared in 2 subgroups of patients taking infliximab based on (1) clinical improvement in self-rated pain, and (2) improvement in both swollen and tender joint counts. For the first analysis, patients were split into 2 groups based on improvement of more than 2 SEM on self-reported pain: those that were "better" and those that were "the same or worse." For the second analysis, patients were split into 2 groups based on those that met and did not meet the American College of Rheumatology (ACR) criteria for a 20% or greater improvement in both swollen and tender joint counts. To control for type 1 error, multivariate analysis of variance (MANOVA) was used to assess the overall mean differences between the subgroups for the health outcome measures. Dependent health status measures were divided into conceptually similar groups: physical, mental, and overall health status. If statistical significance was reached, univariate tests were completed on each of the outcome measures. We hypothesized that improvement in health outcomes would be greater in the "better" group than in the "same or worse" group. A 0.05 level of significance was used for all statistical analysis.

RESULTS

Group 1 (stable patients). The test-retest sample consisted of 24 clinically stable patients with scores on all measures. Table 1 shows the ICC and SEM for each measure. ICC ranged from 0.50 to 0.92. The values were not significantly different between the first and second measurements except

Table 1. Baseline descriptive statistics, intraclass correlation coefficients (ICC), and standard error of measurement (SEM) for Group 1 (stable disease group).

Measures*	Mean	SD	ICC	SEM	SD Diff**
Pain	24.75	23.03	0.76	11.28	17.37
MHAQ	0.46	0.41	0.89	0.14	0.20
EQ-5D†	0.70	0.18	0.66	0.10	0.17
SF-6D†	0.70	0.13	0.72	0.07	0.09
Standard gamble†	0.82	0.15	0.73	0.08	0.14
EQ VAS	67.85	19.55	0.57	12.82	16.13
SF-36 PCS	32.10	10.93	0.87	3.94	5.12
SF-36 MCS	56.54	11.90	0.63	7.24	8.46
SF-36 subscales					
BP	58.60	19.52	0.59	12.50	17.78
PF	47.25	25.31	0.92	7.16	10.48
GH	46.40	21.32	0.71	11.48	14.62
RP	38.75	40.94	0.86	15.32	22.80
VT	52.00	23.75	0.80	10.62	14.44
RE	83.33	35.04	0.50	24.78	34.20
SF	76.25	24.64	0.53	16.89	23.04
MH	82.20	15.38	0.79	7.05	9.26

* MHAQ: Modified Health Assessment Questionnaire; PCS: Physical Component Summary Scale; MCS: Mental Component Summary Scale; BP: bodily pain; PF: physical functioning; GH: general health; RP: role physical; VT: vitality; RE: role emotional; SF: social functioning; MH: mental health.
** Standard deviation of difference scores used in the calculation of Bland-Altman limits of agreement.
† Preference based.

for pain, which had decreased by a small, but significant amount. SEM ranged from 0.07 (SF-6D) to 0.10 (EQ-5D) for the preference based measures, and from 7.05 (mental health) to 24.78 (role emotional) for the SF-36 subscales. Ninety-five percent CI were used to estimate the boundaries for potential meaningful change. For example, for the MHAQ, a change greater than 0.27 (1.96 * SEM) would have less than a 5% chance of being change due to chance alone. Bland-Altman limits of agreement were consistently larger than the estimated boundaries based on the SEM. For example, for the SF-6D a difference score greater than 0.14 would be considered a change using 2 SEM, compared with a change greater than 0.18 using the limits-of-agreement approach.

Group 2 (infliximab patients). Eighty-four patients taking infliximab had baseline data. Of these, 77 had time 2 data (i.e., 14 weeks) and 60 had sufficient data elements to calculate all of the health status indices. Standard gamble was collected on only 24 of the patients and was not used in the MANOVA. Baseline measures of function and health status reflected the high degree of disability in this group (MHAQ mean 1.25). They were clearly more severely affected than the stable group, and the MHAQ and pain scores put them in the most severe quartile based on the data from the National Databank for Rheumatic Diseases[31]. Patients reported low mean SF-36 subscale scores on pain, physical functioning, role physical, and vitality (Table 2). For the utility measures, mean EQ-5D index scores (0.43) were 12 points below the SF-6D index (0.55), while mean standard gamble scores were the highest (0.69). The EQ-5D index had a bimodal distribution compared with the more symmetric distribution of the SF-6D and the slightly negatively skewed distribution of the standard gamble. For the

Table 2. Descriptive statistics at baseline for Group 2 (infliximab) (n = 60).

Measures*	Mean	SD	Minimum	Maximum
Pain	57.05	22.63	1.00	100.00
MHAQ	1.25	0.65	0.00	3.00
EQ-5D	0.43	0.30	–0.36	1.00
SF-6D	0.55	0.07	0.38	0.80
Standard gamble†	0.69	0.25	0.10	0.95
EQ VAS	50.12	17.56	5.00	85.00
SF-36 PCS	24.66	8.08	9.86	46.79
SF-36 MCS	46.58	11.16	22.71	63.00
SF-36 subscales				
BP	33.48	15.77	0.00	74.00
PF	26.90	20.01	0.00	85.00
GH	41.86	19.29	6.25	82.00
RP	7.92	18.69	0.00	75.00
VT	28.64	16.40	0.00	60.00
RE	50.00	46.13	0.00	100.00
SF	48.96	23.39	0.00	100.00
MH	67.69	20.01	12.00	96.00

* MHAQ: Modified Health Assessment Questionnaire; PCS: Physical Component Summary Scale; MCS: Mental Component Summary Scale; BP: bodily pain; PF: physical functioning; GH: general health; RP: role physical; VT: vitality; RE: role emotional; SF: social functioning; MH: mental health.
† Only 24 of the 60 patients completed the standard gamble.

EQ-5D index, a gap of 0.16 points existed between those with and without a 3 level on any one dimension. Mean EQ-5D VAS scores, a self-reported global assessment of health status, were 50.12.

For the infliximab group, all mean laboratory values changed significantly in the expected direction: mean CRP and ESR values decreased by 12.97 and 6.38, respectively, and mean Hb increased by 6.40 g/l. All health outcome measures improved significantly from baseline to 14 weeks following initiation of treatment (Table 3). Based on conventional interpretations of effect size of small (0.2), medium (0.5), and large (0.8)[32], large effect sizes were seen with pain, the SF-6D, EQ-VAS, SF-36 PCS, and both physical and mental health SF-36 subscales (Table 3). The EQ-5D improved by 0.20 and the SF-6D by 0.10, while the SF-6D had the larger effect size. SRM tended to be lower than the effect sizes but followed a similar pattern. The percentage of patients that improved using 2 SEM ranged from 18% (SF-36 MCS) to 58% (MHAQ). Ninety-five percent limits of agreement based on the Bland-Altman limits of agreement were generally wider than the ICC based on 2 SEM.

Patients were divided into 2 groups based first on change in their self-reported pain. The better group was defined as those that improved more than 2 SEM from baseline. The "unchanged" group were those that improved less than 2 SEM. Three patients reported worse pain (a decline > 2 SEM from baseline) but were combined with the "unchanged" group. Mean changes for CRP and ESR were significantly different between the 2 groups (Table 4). Analysis of variance showed that the SF-36 PCS and MCS improved significantly more for the "better" group than the unchanged (i.e., the same or worse) group (Table 5). Using MANOVA, patients in the "better" pain group improved significantly more on physical status measures [$F(5, 54) = 5.70$, $p = 0.000$] and the utility measures [$F(3, 56) = 5.29$, $p = 0.003$] than did patients in the "same or worse" group. Univariate tests showed that all physical status and generic measures differentiated between the 2 groups except SF-36 general health. Mean differences between the 2 groups for the 2 utility measures were 0.10 (SF-6D) and 0.17 (EQ-5D). There were no significant differences between groups for the SF-36 mental health subscales, although difference scores were consistently greater for the "better" group.

To compare changes in outcome measures for 2 groups based on a more physician centered measure, patients were split into 2 groups based on a 20% improvement in both their tender and swollen joint counts. The 2 groups of patients were equivalent at baseline in the number of tender and swollen joint counts. Using t tests, mean changes for Hb and ESR were significantly different between the 2 groups (Table 4). Chi-square analysis showed no significant relationship between improvement in joint counts and improvement in pain. Although the physical health status measures

Table 3. A comparison of responsiveness indices for outcome measures for Group 2 (infliximab) (n = 60).

Measures*	Mean Change**	Effect Size	SRM	Percentage Improved by > 2 SEM	Percentage Improved 95%, Bland-Altman Limits of Agreement
Pain	24.39††	1.08	0.93	52	35
MHAQ	0.40††	0.62	0.74	58	48
EQ-5D	0.20††	0.67	0.64	43	27
SF-6D	0.10††	1.40	0.87	35	25
Standard gamble***	0.12††	0.49	0.43	33	21
EQ VAS	17.38††	0.99	0.90	37	25
SF-36 PCS	8.62††	1.07	0.94	48	37
SF-36 MCS	4.69††	0.42	0.42	18	10
SF-36 subscales					
BP	19.63††	1.24	0.84	40	22
PF	17.67††	0.88	1.01	57	40
GH	11.23††	0.58	0.70	27	13
RP	26.25††	1.40	0.68	37	37
VT	21.11†	1.29	0.94	50	40
RE	13.89†	0.30	0.28	23	10
SF	18.96††	0.81	0.68	40	18
MH	6.81††	0.34	0.35	32	17

* MHAQ: Modified Health Assessment Questionnaire; PCS: Physical Component Summary Scale; MCS: Mental Component Summary Scale; BP: bodily pain; PF: physical functioning; GH: general health; RP: role physical; VT: vitality; RE: role emotional; SF: social functioning; MH: mental health.
** Absolute mean changes with results of paired t tests. † $p < 0.05$; †† $p < 0.001$. All mean changes are all in expected direction (improved).
*** Only 24 of the 60 patients completed the standard gamble. SRM: standardized response mean. SEM: standard error of measurement.

Table 4. Mean change in laboratory values (0 to 14 weeks) in patients with RA treated with infliximab.

	Improvement in Tender and Swollen Joint Counts		Improvement in Pain†	
	< 20%	≥ 20%	Same or Worse	Better
CRP	14.85	16.18	5.72*	25.11*
ESR	2.91*	8.66*	5.27*	10.62*
Hemoglobin	2.86*	10.20*	2.54	7.31

† Improvement based on 2 SEM.
* T test significant $p < 0.05$.

showed a consistent trend of greater improvement in the group that improved by 20% or greater, MANOVA showed no significant differences in outcome measures between groups (Table 5).

DISCUSSION

Our objective was to assess the relative responsiveness to change in a clinical study of several patient centered outcome measures in individuals with RA. We used the pragmatic definition of responsiveness of deBruin, *et al*[26]: "The accurate detection of change when it has occurred." We used a group of patients before and 14 weeks after initiation of therapy with infliximab. This was chosen because published experience[33] as well as our own has shown that patients generally respond well to this regime as measured by objective criteria, e.g., joint counts and radiologic stabilization. As reviewed by Beaton, *et al*[29] our aim was to determine if, in practice, the measures could readily detect such differences that were recognized to be clinically important and not merely statistically significant. We used a group of patients with stable and low disease activity for comparison, i.e., to provide a reference point for change as well as to determine smallest detectable differences (SDD). That these "smallest detectable differences" may not be equivalent to "minimum clinically important differences" (MCID) has been reviewed[28] and is currently under study by OMERACT. The stable group had a disease activity deemed acceptable by the treating physician and no changes in therapy occurred. While they could have experienced some clinical changes between the tests, it was felt that this was closer to reality than repeating the test immediately on the same occasion. We decided, along with the treating clinicians, that changes seen within that group on the 2 occasions were not "clinically important." The infliximab group, in contrast, were to have new therapy initiated because of persistent active disease. This was illustrated by the increased severity of the baseline scores in Table 2, compared to those of the stable patients (Table 1). The results of the repeated measurements in the stable group were used to provide a reference point for change. A limits-of-agreement approach[30] (Bland-Altman) was used, as was the SEM, to assess the SDD. Others have pointed out that for many clinical measures, e.g., MHAQ, ESR, joint count, etc., the 95% confidence intervals with the limit-of-agreement approach were much greater than those values considered important and relevant at a clinical level[21]. It has also been pointed out that what is an important degree of change, in contrast to the SDD, may vary with the baseline severity and would be larger in our Group 2 patients than in the stable group[8,34].

The relationship between the physician reported assessment of improvement, i.e., joint scores of 20%, and the other measures was not good, although this has been previously reported[3]. The "subjective" assessment of improvement, i.e., a decrease in pain by at least 2 SEM, was more closely related to physical status measures and the 2 preference based measures, the EQ-5D and the SF-6D. Our results are consistent with those of Kosinski, *et al*[8], who noted changes in the SF-36 and HAQ scores to be more strongly related to changes in patient global and patient pain assessments than to changes in joint swelling and tenderness counts. Prior calculations of minimal important changes in the SF-36

Table 5. Mean difference scores in outcome measures for 2 groups based on improvement in pain, or in joint count.

Improvement	Physical Status†**								Mental Status†			Health Status*		
	PCS	MCS	MHAQ	BP	PF	GH	RP	VT	RE	SE	MH	EQ-5D	EQ VAS	SF-6D
Pain														
Better (n = 31)	12.19	8.65	0.54	32.45	23.91	12.39	37.90	26.99	22.58	27.42	8.59	0.28	23.55	0.15
Same or worse (n = 29)	4.81	3.27	0.25	5.93	11.00	9.98	13.79	14.83	4.60	9.91	4.90	0.11	11.03	0.05
F ratio	11.49**	4.47*	4.30*	27.91**	9.40**	0.34	6.47*		—	—	—	5.02*	6.94*	12.26**
Joint 20%, TJC and SJC														
Yes (n = 34)	9.26	6.65	0.46	22.40	21.61	9.23	28.57	23.45	17.65	19.49	4.98	0.21	16.97	0.09
No (n = 26)	5.80	5.05	0.27	17.65	10.63	9.73	22.83	21.47	19.23	19.23	9.73	0.23	20.90	0.10

PCS: SF-36 Physical Component Summary Scale; MCS: SF-36 Mental Component Summary Scale; MHAQ: Modified Health Assessment Questionnaire; SF-36 subscales; BP: bodily pain; PF: physical functioning; GH: general health; RP: role physical; VT: vitality; RE: role emotional; SF: social functioning; MH: mental health.
* $p < 0.05$, ** $p < 0.005$.

scales[35] and the HAQ[20] disability scores were also similar to those calculated from our stable group.

A disadvantage of this study was that no placebo group was included as a control for the responsiveness, and it is possible that the use of an open label treatment may have the effect of increasing pre- and post-differences, particularly in the subjective assessments.

One could argue that the joint count is the more reliable and biologic measure, and that as the patient preference data do not reflect these changes well, they should be put in second place. However, in outcome studies using ACR criteria or the Disease Activity Score, self-assessed patient data (e.g., HAQ) are a critical area of clinical assessment of change in RA. Patient preference data have also become a measure of importance and are not merely critical to but are the essence of the imputation of cost effectiveness and QALY calculations. Our study did show that while reduction in joint count was associated with improvement in some biological measures (ESR and hemoglobin), so were the patient rated measures (CRP and ESR) (Table 4).

The calculation of QALY via utility measures and the assignation of dollar values to these has become an important technique for resource allocation[12,14,35]. The use of generic QOL measures is critical for this, and there has been significant criticism of their use in rheumatic diseases. We chose some of the more frequent measures, as well as the newer SF-6D, to assess test/retest reliability and responsiveness in patients with RA. While the time tradeoff and standard gamble both determine utilities and the results have tended to be similar[12], the standard gamble was chosen over the time tradeoff as this is the only measure that incorporates uncertainty, an important theoretical part of the "utility" construct. The EQ-5D index relates the scores to utilities obtained via the time tradeoff approaches, whereas the SF-6D relates to the standard gamble. The standard gamble did not perform well, as others have noted[4,13,31,36,37]. We discontinued its use after the first 24 patients in Group 2 primarily because of the time requirements. Although this is a small number on which to base decisions, we also found it to have a relatively poor responsiveness. Changes in EQ-5D utilities were roughly twice those for the SF-6D, and similar to other findings in rheumatology patients[38]. This is partly due to the greater range of scores for the EQ-5D index compared with the SF-6D, but although the mean numerical change is greater, the effect size is actually smaller because of the larger variability of scores (and the difference in the numerical range of the scales). The standard deviation of the EQ-5D at baseline was 0.3 compared with 0.07 for the SF-6D. Like others[21], we found that the use of the Bland-Altman CI appears to overestimate the threshold for a meaningful change.

It appears that the newer patient preference indices, especially that based on the more familiar SF-36 (i.e., the SF-6D) are effective for use in clinical practice in patients with RA, and are responsive to clinically important changes.

REFERENCES

1. Pincus T, Wolfe F, Strand V, Kavanaugh A, Crawford B, Felson D. Patient questionnaire data discriminate between drug versus placebo as effectively as the ACR 20 in a clinical trial of patients with RA [abstract]. Arthritis Rheum 2001;45 Suppl:S186.
2. Wolfe F, Skevington SM. Measuring the epidemiology of distress: The rheumatology distress index. J Rheumatol 2000;27:2000-9.
3. Tugwell P, Wells G, Strand V, et al. Clinical improvement as reflected in measures of function and health related quality of life following treatment with leflunomide compared with methotrexate in patients with rheumatoid arthritis. Arthritis Rheum 2000; 43:506-14.
4. Verhoeven AC, Boers M, Van der Linden S. Responsiveness of the Core Set, response criteria and utilities in early rheumatoid arthritis. Ann Rheum Dis 2000;59:966-74.
5. Pincus T, Summey JA, Soraci JA, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Standard Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346-53.
6. Ware JE, Sherbourne CD. The Medical Outcomes Study, SF-36 item short form health survey (SF-36) conceptual framework and item selection. Med Care 1992;30:473-83.
7. Birrell FN, Hassell AB, Jones PW, Dawes PT. How does the Short Form 36 health questionnaire (SF-36) in rheumatoid arthritis (RA) relate to RA outcome measures and SF-36 population values? A cross-sectional study. Clin Rheumatol 2000;19:195-9.
8. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining minimally important changes in generic and disease specific health related quality of life questionnaires in clinical trial of rheumatoid arthritis. Arthritis Rheum 2000;43:1478-87.
9. Ruta DA, Hurst JP, Kind P, Honter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis. Br J Rheumatol 1998;37:425-36.
10. Wells G, Boers M, Shea B, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis. J Rheumatol 1999;26:217-21.
11. Talance J, Frater A, Gallivan S, Young A. Use of the Short Form 36 for health status in rheumatoid arthritis. Br J Rheumatol 1997;36:463-9.
12. Torrance GW, Feeny D. Utilities and quality adjusted life years. Int J Technol Assess Health Care 1989;5:559-75.
13. Guillemin F. The value of utility: assumptions underlying preferences and quality adjusted life years. J Rheumatol 1999;26:1861-3.
14. Laupacis A, Bourne R, Rorabeck C, et al. The effect of elective total hip replacement on health related quality of life. J Bone Joint Surg Am 1993;75:1619-26.
15. Tijhvis GJ, Jansen SJT, Stiggelbout AM, Ziomderman AH, Hazes JMW, Vlieland TPMV. Value of the time trade off method for measuring utilities in patients with rheumatoid arthritis. Ann Rheum Dis 2000;59:892-7.
16. Hurst NP, Kind P, Ruta D, et al. Measuring health related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of Euroquol (EQ5D). Br J Rheumatol 1997;36:551-9.
17. Vaile JH, Mathers DM, Ramos-Remus C, Russell AS. Generic health instruments do not comprehensively capture patient perceived improvement in patients with carpal tunnel syndrome. J Rheumatol 1999;26:1163-6.
18. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51:1115-28.

19. Rabin R, de Charro F. EQ-5D. A measure of health status from the Euro QOL group. Ann Med 2001;33:337-43.
20. Gabriel SE. Controversies in economic evaluation in the rheumatic diseases. J Rheumatol 1999;26:1859-60.
21. Wolfe F, Pincus T, Fries JF. Usefulness of the HAQ in the clinic [letter]. Ann Rheum Dis 2001;60:811.
22. Wells G, Boers M, Shea B, et al. Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT study. J Rheumatol 1999;26:217-21.
23. Ware JE, Kosinski M, Keller SD. SF-36 physical and mental health summary scales: a user's manual. Boston: The Health Institute, New England Medical Center; 1994.
24. Brazier J, Roberts J, Deverill M. The estimation of a preference based, single index measure for health from the SF-36. J Health Econ 2002;21:271-92.
25. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095-108.
26. DeBruin AF, Diederiks JPM, DeWille LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP 68. J Clin Epidemiol 1997;50:529-40.
27. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53:459-68.
28. Lassere MND, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis. Implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001; 28:392-403.
29. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001;54:1204-17.
30. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
31. Wolfe F, O'Dell JR, Kavanaugh A, Wilske K, Pincus T. Evaluating severity and status in rheumatoid arthritis. J Rheumatol 2001;28:1453-62.
32. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
33. Lipsky PE, van der Heijde DM, St. Clair EW, et al, for the Anti-tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Infliximab and methotrexate in the treatment of rheumatoid arthritis patients receiving methotrexate. New Engl J Med 1999;340:253-9.
34. Wolfe F. The psychometrics of functional status questionnaires: room for improvement. J Rheumatol 2002;29:865-8.
35. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001;33:344-9.
36. Hawthorne G, Richardson J, Day NA. A comparison of the assessment of quality of life (AqoL) with four other generic utility instruments. Ann Med 2001;33:358-70.
37. Bakker CH, Rutten-van Molken MPMH, van Doorslaer EKA, Bennet K, Van der Linden S. Feasibility of utility assessments by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. J Rheumatol 1994;21:269-74.
38. Conner-Spady B, Suarez-Almazor ME. A comparison of preference-based health status tools in patients with musculoskeletal disease. In: Norinder AL, Pedersen KM, Roos P, editors. Proceedings of the 18th Plenary Meeting of the EuroQol Group, Sept. 6–7, 2001. Copenhagen. Lund, Sweden: The Swedish Institute for Health Economics; 2001.