Quality of Life Research (2005) 14: 1333–1344
DOI 10.1007/s11136-004-6012-0

Are indirect utility measures reliable and responsive in rheumatoid arthritis patients?

Carlo A. Marra[1,2], Amir A. Rashidi[3], Daphne Guh[3], Jacek A. Kopec[4,5], Michal Abrahamowicz[6], John M. Esdaile[5,7], John E. Brazier[8], Paul R. Fortin[9] & Aslam H. Anis[4]
[1]*Faculty of Pharmaceutical Sciences, University of British Columbia;* [2]*Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Institute, Vancouver, BC, Canada;* [3]*Centre for Health Evaluation and Outcome Sciences, St. Paul's Hospital, Vancouver, Canada;* [4]*Department of Health Care and Epidemiology, Faculty of Medicine, University of British Columbia (E-mail: aslam.anis@ubc.ca);* [5]*Arthritis Research Centre of Canada, Vancouver, Canada;* [6]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada;* [7]*Division of Rheumatology, Faculty of Medicine, University of British Columbia;* [8]*Sheffield Health Economics Group, School of Health & Related Research, University of Sheffield, Sheffield, UK;* [9]*Division of Rheumatology, Toronto Western Hospital, University of Toronto, Toronto, Canada*

Accepted in revised form 4 November 2004

Abstract

Background: Preference-based, generic measures are increasingly being used to measure quality of life and as sources for quality weights in the estimation of Quality Adjusted Life Years (QALYs) in rheumatoid arthritis (RA). However, among the most commonly used instruments (the Health Utilities Index 2 and 3 [HUI2 and HUI3], the EuroQoL-5D [EQ-5D], and the Short Form-6D [SF-6D], there has been little comparative research. Therefore, we examined the reliability and responsiveness of these measures and the Rheumatoid Arthritis Quality of Life (RAQoL) and the Health Assessment Questionnaire (HAQ) in a sample of RA patients. *Major findings*: Test–retest reliability was acceptable for all of the instruments with the exception of the EQ-5D. Using two external criteria to define change (a patient transition question and categories of the patient global assessment of disease activity VAS), the RAQoL was the most responsive of the instruments. For the indirect utility instruments, the HUI3 and the SF-6D were the most responsive for measuring positive change. On average, for patients whose RA improved, the absolute change was highest for the HUI3. *Conclusions*: The HUI3 and the SF-6D appear to be the most responsive of the preference-based instruments in RA. However, differences in the magnitude of the absolute change scores have important implications for cost-effectiveness analyses.

Introduction

Improvement in health related quality of life (HRQL) is one of the most important goals in the management of rheumatoid arthritis (RA) [1]. As such, HRQL and health status measures have often been used as outcomes in clinical trials and studies assessing a variety of interventions in RA [2–5]. A variety of instruments that assess RA-specific HRQL (for example, the Arthritis Impact Measurement Scales (AIMS), the Rheumatoid Arthritis Quality of Life questionnaire (RAQoL)) or generic HRQL or function (such as the Short Form 36 (SF-36)) have been applied to the assessment of RA [2, 6, 7].

Preference-based or indirect utility measures are generic HRQL measures that are often used in clinical and observational studies as the scores that they generate can be utilized to calculate quality adjusted life-years (QALYs) and can thus be integrated into cost-utility analyses [8]. Examples of these instruments include the Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3), EuroQol (EQ-5D), and the Short Form 6D (SF-6D).

All of these instruments have been previously applied in the assessment of patients with RA [9–11].

Responsiveness is often defined as the ability of an instrument to measure change [12]; however, there are multiple definitions of responsiveness that exist in the literature [13, 14]. There has been little work in the evaluation and comparison of responsiveness (using any definition) of the indirect utility instruments. A recent study by Conner-Spady and Surez-Almazor [11], examined the responsiveness of three preference-based measures of HRQL (EQ-5D, HUI3, and the SF-6D) in a sample of patients with at least one of several types of rheumatological conditions. To our knowledge, there have been no evaluations of the responsiveness of the RAQoL in RA in North American populations although one has been published in a Swedish sample [7]. Therefore, there remains a need for more research to assess the responsiveness of these measures, to compare their characteristics, and to determine how their properties compare to disease-specific measures. Finally, since the indirect utility measures are often used as the source of quality weightings used for the estimation of QALYs in cost-utility studies in RA, it is important that they are determined to be reliable, valid and responsive in this disease state.

Therefore, the primary purpose of this study was to examine the reliability and responsiveness of the indirect utility instruments and the RAQoL and the HAQ from baseline to 6 months in a sample of rheumatoid arthritis patients.

Methods

Study sample

To be included, subjects had to have a rheumatologist-confirmed diagnosis of RA (as defined by the American College of Rheumatology diagnostic criteria) [15], receive rheumatology care within the province of British Columbia, consent to and be sufficiently proficient in English to answer the questionnaires and be willing to participate in follow-up surveys. Recruitment of RA patients began in October 2001 and ended in September 2002. Ethical approval for this study was obtained through the University of British Columbia's Behavioural Ethics Committee and informed consent was obtained from each of the participants.

Eight private rheumatologists' offices from the study areas referred subjects into the sample during their interactions in routine clinical practice. In addition, two of the eight rheumatologists' practices sent letters to all of their patients with RA inviting them to participate in the survey. All patient questionnaires were self-administered, self-completed and submitted via mail. The study rheumatologists' offices supplied additional information from the patients' health record.

Measures

Participants were asked to complete a questionnaire at baseline and three and six months thereafter. The questionnaire consisted of sections devoted to socio-economic, clinical and functional status and quality of life assessment instruments.

Clinical

Participants self-reported clinical variables included swollen joint count (SJC) and tender joint count (TJC) (using the mannequin-based 42 joint count methodology) [16], a 10 cm pain visual analogue scale (VAS), a patient global assessment of disease activity (10 cm VAS) [1], and RA severity and RA control (both using a 5 point Likert scale). The attending rheumatologists were asked to complete a physician global assessment of disease activity (10 cm VAS) for each patient [1].

For the 6-month questionnaire, participants were asked to complete a five point Likert scale that assessed changes in their RA since answering the baseline questionnaire. The question asked was 'Overall, how would you describe changes in your rheumatoid arthritis since answering the FIRST questionnaire (i.e. about 6 months ago?'). Response choices included 'Much Worse', 'Somewhat Worse', 'The Same', 'Somewhat Better' and 'Much Better'. These questions are referred to as 'patient transition questions'. To increase the number of patients in each category, responses to these questions were collapsed into three categories as follows: (1) worse (included responses 'much worse and somewhat worse'); (2) the same; and (3) better (included 'much better and somewhat better) which is a similar approach adopted by other investigators [9, 12, 14, 17].

The sample of RA patients in our study experienced 'natural' courses of their disease over time rather than changes associated with a treatment of known efficacy. In group level analyses, average change scores can mask the proportion of patients with follow-up scores that differ (either improved or deteriorated) from those at baseline. Because of this, we carried out separate analyses for each of the distribution-based responsiveness measures according to our collapsed transition question criteria ('worse', 'the same', or 'better'). This is the same approach postulated by other investigators [11, 17].

Health status and HRQL measures

Health Assessment Questionnaire (HAQ) Disability Index

The HAQ is a measure of physical disability that assesses ability to complete everyday tasks in areas such as dressing and grooming, rising, eating, walking, personal hygiene, reach, grip and other activities (such as getting into and out of a car). Each of these areas is assigned a section score that is further adjusted to account for the use of any aids, devices or help from another person. Section scores are then summed and averaged to give an overall score between 0.0 (best possible function) to 3.0 (worst function). A HAQ score difference of 0.25 is said to represent the minimally important difference (MID) [18, 19].

Rheumatoid Arthritis Quality of Life Questionnaire (RAQoL)

The RAQoL consists of 30 questions (answered by yes/no) that assess such aspects of RA as moods and emotions, social life, hobbies, everyday tasks, personal and social relationships, and physical contact. The RAQoL is scored by assigning a point for each affirmative response and no points for negative responses. Thus, scores range from 0 (least severity) to 30 (highest severity). To date, the MID for the RAQoL has been estimated to be approximately 2.00 [20].

Preference based indirect utility assessment instruments

The indirect utility assessment instruments used were the HUI2, HUI3, SF-6D, and the EQ-5D [21]. In a cross-sectional analysis in patients with RA, the MID for the overall utility scores was determined to be 0.03 to 0.04 for the HUI2, 0.06 to 0.07 for the HUI3, and 0.03 to 0.05 for the SF-6D and the EQ-5D [20]. Grootendorst et al. concluded that differences on the HUI3 of 0.03 or more should be considered to be clinically important [22], whereas Samsa et al. [23]. determined, from a small random sample of 160 patients from a Veteran's Administration hospital, that 0.02 (95% confidence interval 0.01–0.05) was a clinically meaningful difference for the HUI2. Based upon these results and the fact that change in one level within any attribute in either system (a clinically important change) generates a change of 0.03 or more in overall score forms the basis for the guideline that differences of 0.03 or more in HUI2 and HUI3 scores are clinically important [24]. In another analysis of seven longitudinal studies examining SF-6D global utility scores, investigators estimated that the MID to be 0.033 (95% CI: 0.029–0.037) [10]. A recent comprehensive review of the similarities and differences across these instruments is available and is beyond the scope of this research paper [21].

Data analysis

Reliability

To determine test–retest reliability, a second questionnaire was sent to a randomly selected group of 50 patients immediately after receipt of their follow-up questionnaire with the instructions to complete and return within 5 weeks. The 5 week period was chosen as this was determined *a priori* to be the time window in which changes (either improvement or deterioration) in their RA would be unlikely. Intraclass correlation coefficients (two-way mixed effect model such that the subject effect was random and the instrument effect was fixed) were calculated for the overall scores from the two time periods (Table 1).

Measures of responsiveness

Our analysis assessed responsiveness to change in RA for the indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D), the RAQoL and the HAQ for the changes between the baseline and six month responses. For each patient who had data on all instruments at each of the pair of visits,

Table 1. Test–retest reliability

Instrument	ICC	95% CI
HUI2	0.77	0.59–0.88
HUI3	0.81	0.66–0.90
SF-6D	0.89	0.79–0.94
EQ-5D	0.46	0.18–0.68
HAQ	0.97	0.93–0.98
RAQoL	0.93	0.86–0.96

Questionnaire results compared to results within 35 days. Results are intraclass correlation coefficients (ICC) with 95% confidence intervals (CIs).

the difference between the two corresponding scores was calculated. In the primary analysis of responsiveness, the results were stratified into patients classified as 'better', 'the same', or 'worse' according to the collapsed transition question. In addition, in a secondary analysis, utilizing the patient global assessment of disease activity (called 'patient global' hereafter), the percentage improvement over baseline (i.e. the relative change) was calculated utilizing the following formula:

$$\frac{(\text{6mos.patientglobal} - \text{baseline.patientglobal})}{(\text{baseline.patientglobal})} \times 100$$

According to this formula and adapting guidelines of response from American College of Rheumatology 20 criteria [1], patients were classified as: (1) 'better' if the patient global had changed by ≥20%, (2) 'the same' if the patient global had changed $> -20\%$ and $<20\%$; and (3) 'worse' if the patient global had changed $<-20\%$. All the indices of responsiveness (as described below) were calculated for the subgroups defined by this criterion.

Three distribution-based approaches were employed to assess responsiveness:

(1) the effect size (ES) [13, 25] using the following formula:

$$\frac{\text{mean}(x_1 - x_2)_{\text{totalgroup}}}{\text{SD}_{\text{totalgroup}}}$$

where x_1 is the mean score at 6 months for the entire group; x_2, the mean score at baseline for the entire group; $\text{SD}_{\text{totalgroup}}$, the standard deviation at baseline for the entire group.

An effect size of 1 indicates a mean change in magnitude equivalent to one standard deviation. We adopted the criteria of Cohen, where absolute values of effect sizes (d) can be categorized as small (<0.5), medium (0.5–0.8), or large (>0.8) [26, 27]. Positive values reflect improvement while negative values reflect worsening for the indirect utility instruments while the converse is true for the HAQ and the RAQOL.

(2) the standardized response mean (SRM) [13] using the following formula:

$$\frac{\text{mean}(x_1 - x_2)_{\text{totalgroup}}}{\text{SD}(x_1 - x_2)_{\text{totalgroup}}}$$

where x_1 is the mean score at 6 months for the entire group; x_2 the mean score at baseline for the entire group; SD $(x_1 - x_2)_{\text{totalgroup}}$, the standard deviation (SD) for the change in scores in the entire group.

The absolute values of the SRM are regarded as either small (<0.5), medium (0.5–0.8) or large (>0.8) and the signs (either positive or negative) are interpreted as for the ES [27].

(3) the relative efficiency statistic (RE) [28, 29] using the following formula:

$$\left[\frac{t_{\text{comparison}}}{t_{\text{goldstandard}}}\right]^2$$

Given the information on the superior responsiveness of disease-specific over generic measures [30], we selected the RAQoL as the 'gold standard' which to compare each of the instruments. The measure with the highest RE has the highest power for a given sample size, or requires fewer patients, to achieve a given level of statistical power [12].

Since the standard errors of the distribution-based approaches are not defined, we used bootstrap methods to estimate 95% confidence intervals (CI) for the ES, and the SRM [31]. Rather than conduct a large number of statistical tests, the 95% CIs were investigated to determine the degree of overlap between the values generated across the HRQL measures.

The distribution-based methods described above do not provide answers to practical questions such as, for example, how likely is a decrease in a specified amount in the utility score (as measured by the indirect instruments) to represent actual deterioration? Thus, we utilized a flexible polytomous regression model [32] to assign probabilities of patient's improvement, status quo, or deterioration (as defined by the

transition question) to different levels of change in the indirect utility and disease specific HRQL measures. The polytomous regression has been adapted to assess responsiveness and the results are presented in a graph of three curves, each of which describes how the estimated probability of a respective outcome (improvement, no change, or worsening as defined by the collapsed transition question or the patient global assessment of disease activity question), changes as a function of the difference in two consecutive scores [17]. Bootstrap sampling with 1000 simulations was performed to obtain the empirical 95% confidence limits of each estimated curve.

Finally, we examined associations between changes in either the unweighted domain scores of the EQ-5D and the SF-6D (as these instruments do not typically calculate single-attribute utility values) or the single-attribute utility scores of the HUI2 and HUI3 with the external criteria. The purpose of these analyses was to investigate which domains/single attributes were most likely to change in response to improvement or worsening in RA (as defined by the external criteria). Statistical analysis using Kruskal–Wallis was employed. Conservatively, we defined a clear association if the statistical tests were significant ($p < 0.05$) for the domain or single attribute with both external criteria.

Results

Demographics and missing values

Of the 320 RA patients who returned the baseline questions, 239 (75%) returned the 6 month questionnaires. Characteristics of our baseline sample have been described in detail elsewhere [20]. Baseline characteristics of those who completed the 6 month questionnaires compared to those who did not were similar between the two groups. However, for all of the instrument scores, those who completed the 6 month questionnaires appeared to have poorer baseline mean HRQL scores than those who did not (with the exception of the HUI2) but this relationship was statistically significant only for the HAQ. Other variables that differed between the subgroups were self-reported severity and proportion who worked outside the home in the past 12 months (both favoring those only completing the baseline questionnaire).

Reliability

The results for the test–retest reliability approach for the generic and disease specific instruments are shown in Table 2. The EQ-5D overall score appeared to be the lowest while the RAQoL and the HAQ displayed the highest reliability.

Responsiveness

For the 0–6 months transition question, 96 (40%) reported improvement, 85 (36%) reported no change and 58 (24%) reported worsening. Of these, 222 patients had pairs of answers on all questionnaires to permit comparisons (89 reporting improvement, 77 reporting status quo and 56 reporting worsening). For the secondary external criterion (as defined by categorization of the patient global assessment of disease severity VAS) for these 222 pairs, results of the patient global scores were available and were classified as follows: 65, 118, and 39 reporting improvement, status quo and worsening using criterion described in the Methods section. The two external criteria had fairly low agreement (weighted kappa 0.30, 95% CI 0.20–0.41).

The indices of responsiveness (ES, SRM, and the RE) and their associated 95% CI for those who responded as better, the same or worse according to the transition question and the patient global rating of disease severity VAS are presented in Table 2. Generally, the results of the various responsiveness statistics tended to agree within each of the instruments and there was little overlap between their 95% CI. Overall, the RAQoL was the most consistently responsive of the instruments tested regardless of which of the external criteria were applied. Depending on whether the change was classified as either 'worse' or 'better' and which of the external criteria were applied, the indirect utility instruments and the HAQ displayed varying degrees of responsiveness. For example, the EQ-5D appeared to be responsive in those who were classified as 'worse' irrespective of which external criteria were applied but less responsive in those classified as 'better'. The HAQ appeared to be relatively responsive in both those classified as better or worse using the patient transition question to

Table 2. Differences and responsiveness statistics from baseline to 6 months stratifying the sample by the transition question and by the patient global VAS categories

Measures	Transition question defined categories						Patient global VAS defined categories					
		Effect Size	95% CI	SRM	95% CI	RE		Effect Size	95% CI	SRM	95% CI	RE
HUI3	Worse	−0.10	−0.31 to 0.13	−0.12	−0.56 to 0.08	0.12	Worse	−0.36	−0.04 to −0.65	−0.46	−0.07 to −0.88	0.78
	Same	0.12	−0.03 to 0.26	0.18	−0.14 to 0.31		Same	0.05	−0.04 to 0.24	0.07	−0.06 to 0.31	
	Better	0.23	0.08 to 0.41	0.29	0.01 to 0.40	0.39	Better	0.60	0.28 to 0.72	0.73	0.29 to 0.80	0.74
HUI2	Worse	−0.14	−0.41 to 0.10	−0.16	−0.39 to 0.16	0.25	Worse	−0.33	−0.63 to −0.07	−0.44	−0.10 to −0.80	0.82
	Same	0.19	−0.05 to 0.27	0.18	−0.05 to 0.39		Same	0.10	−0.08 to 0.18	0.13	−0.12 to 0.24	
	Better	0.30	0.16 to 0.47	0.40	0.10 to 0.52	0.72	Better	0.49	0.39 to 0.83	0.52	0.48 to 1.01	1.02
EQ-5D	Worse	−0.16	−0.44 to 0.06	−0.19	−0.66 to −0.02	0.73	Worse	−0.55	−0.16 to −0.52	−0.63	−0.19 to −0.85	1.14
	Same	−0.11	−0.36 to 0.16	−0.11	−0.41 to 0.02		Same	−0.09	−0.17 to 0.01	−0.10	−0.34 to 0.01	
	Better	0.15	0.01 to 0.31	0.20	0.12 to 0.59	0.24	Better	0.36	0.16 to 0.52	0.43	0.29 to 0.73	0.61
SF-6D	Worse	−0.08	−0.24 to 0.08	−0.13	−0.44 to 0.15	0.21	Worse	−0.24	−0.02 to −0.49	−0.35	−0.04 to −0.87	0.62
	Same	0.36	0.19 to 0.56	0.50	0.31 to 0.70		Same	0.18	0.05 to 0.31	0.26	0.09 to 0.45	
	Better	0.31	0.11 to 0.49	0.36	0.16 to 0.58	0.52	Better	0.54	0.32 to 0.79	0.62	0.41 to 0.85	0.90
RAQoL	Worse	0.19	0.04 to 0.33	0.34	−0.10 to 0.45	1.00	Worse	0.33	0.21 to 0.83	0.56	0.20 to 0.67	1.00
	Same	−0.17	−0.19 to 0.05	−0.33	−0.39 to 0.07		Same	−0.14	−0.08 to 0.28	−0.27	−0.08 to −0.29	
	Better	−0.36	−0.51 to −0.20	−0.51	−0.22 to 0.60	1.00	Better	−0.56	−0.18 to −0.75	−0.69	−0.27 to −1.08	1.00
HAQ	Worse	0.22	0.04 to 0.38	0.33	0.06 to 0.65	1.21	Worse	0.34	0.11 to 0.44	0.50	0.28 to 0.88	0.97
	Same	−0.09	−0.28 to 0.02	−0.20	−0.56 to −0.10		Same	−0.08	−0.06 to −0.25	−0.17	−0.12 to −0.46	
	Better	−0.24	−0.38 to −0.11	−0.39	−0.69 to −0.30	0.71	Better	−0.35	−0.32 to −0.76	−0.50	−0.48 to −0.92	0.72

define the groups, but less responsive (in relation to the other instruments) when the patient global assessment of disease severity criterion was applied. The HUI3 appeared to be poorly responsive except in those classified as 'better' by the patient global assessment of disease severity. The HUI2 was consistently ranked among the middle in responsiveness and the SF-6D appeared to be more responsive in those classified as 'better' (by either criterion) than those classified as 'worse'.

Flexible polytomous regression techniques

Selected results from the flexible polytomous regressions exploring responsiveness are shown in Figures 1 and 2. The curves on each figure correspond to the three types of outcome (worse, same, better) as defined by the each of external criteria (patient transition question or the patient global assessment of disease activity). Each curve shows how the estimated probabilities of a specific

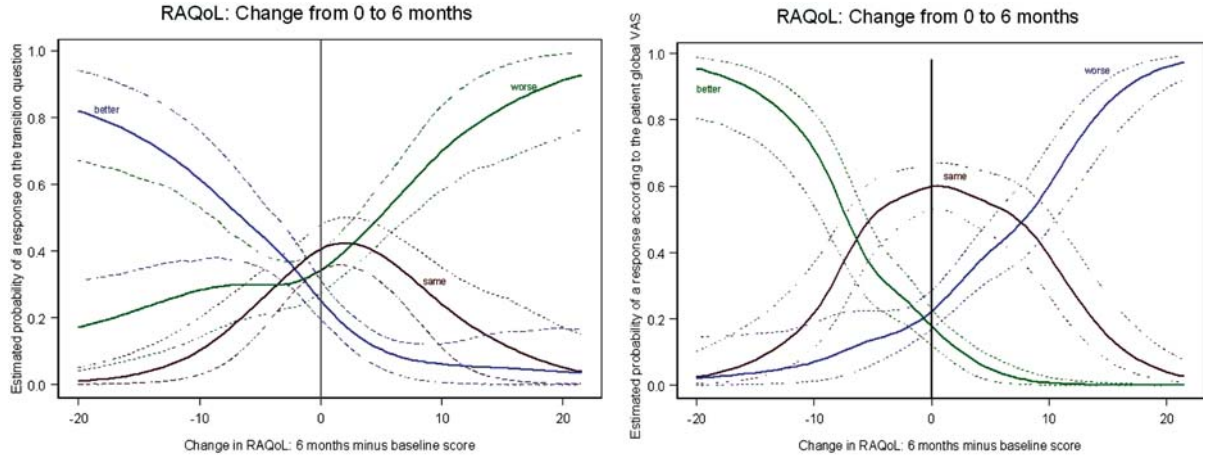

Figure 1. Results of the multi-response model of the association between a change in the RAQoL and the external criteria (transition question on the left hand side and patient global VAS criteria on the right hand side). The solid lines represent the fitted model whereas the dotted lines represent the 95% confidence limits.

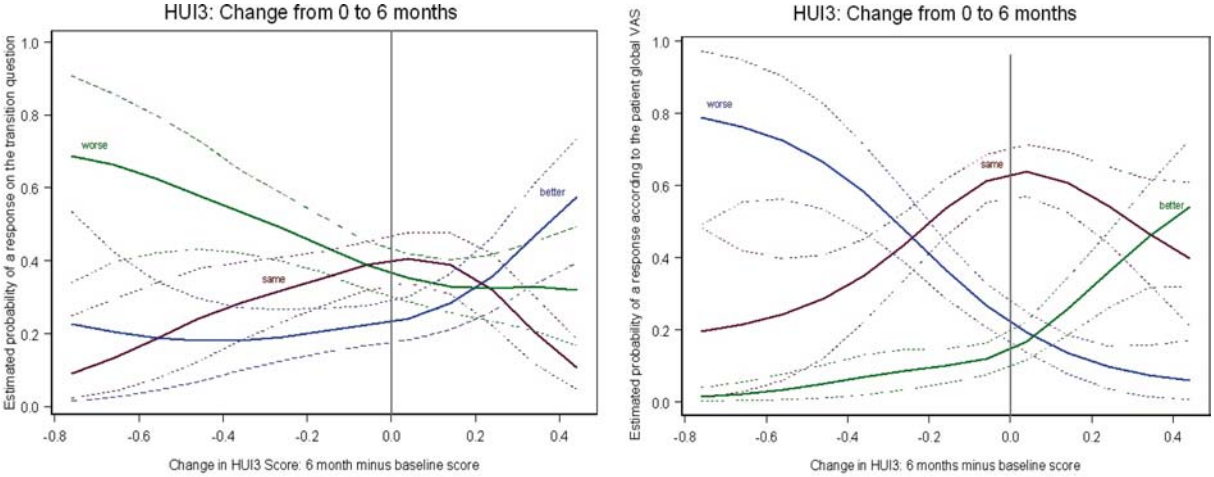

Figure 2. Results of the multi-response model of the association between a change in the HUI3 and the external criteria (transition question on the left hand side and patient global VAS criteria on the right hand side). The solid lines represent the fitted model whereas the dotted lines represent the 95% confidence limits.

response vary depending on the observed change in the scores of the instruments.

In general, the results of using the patient global assessment of disease activity VAS appear to be better able to discriminate between those patients whose RA has improved, worsened or stayed the same than the transition question. This is evident in all of the graphs as there is a sharper delineation between the three curves (worse, better and same) in. Overall, the RAQoL appeared to be most responsive as shown in Figure 1 as compared with the other instruments using the same external criterion. For example, in Figure 1 in the right hand pane, there is very good discrimination between the three curves as shown by their degree of separation. The probability of being classified as 'the same' is high (approximately 60%) if the difference between the two scores is zero. Similarly, this probability decreases as we move in either direction and becomes extremely small when the difference is ± 20. As the difference in the scores gets larger in the positive direction (recall that larger values in the RAQoL reflect worse HRQL), the probability of being classified as 'worse' grows to >80% when the difference in scores is approximately 15 and almost 100% when the difference is 20. These values are similar to those displayed for negative values (reflecting improvement) in the RAQoL and the dashed curve labeled as 'better'.

For the indirect utility instruments, using the patient transition question as the external criteria for change, there was generally fairly poor discrimination between the curves with significant overlap between the probabilities of being classified 'better', 'worse' and 'same' across the range of difference scores. Using the patient global assessment of disease activity VAS criteria, the curves for all the indirect utility instruments showed much better discrimination between those classified as 'better' and 'worse'. However, for those classified as the 'same', there was considerable overlap between these probabilities and the probabilities for 'better' and 'worse'. The HUI3 appeared to be the best able to discriminate in this regard (Figure 2). Thus, it would seem that although these instruments can discriminate change well (according to the external criterion) in those who improve or worsen, those that stay the same yield somewhat problematic difference scores. This finding could be a property of the instruments or may be a reflection of the cut-off values of our external criterion.

Similarly, for the HAQ, the patient global assessment of disease severity VAS criterion appeared to result in better discrimination between the curves; however, as with the indirect utility measures, there was considerable overlap between the 'same' category and the other categories.

Change in unweighted domain scores (EQ-5D, SF-6D) and single attribute utilities (HUI2, HUI3)

For the EQ-5D, pain/discomfort, anxiety/depression and self-care, and, for the SF-6D, physical, and social functioning, role limitations and pain met our criteria for statistical significance. For the single attributes from the HUI systems, ambulation, emotion, and pain (from the HUI3) and mobility, emotion and pain (from the HUI2) met the criteria. Of note, there were more significant associations between the domains/single attributes and the changes defined by the patient global assessment of disease severity categories than the patient transition question responses. For example, with the EQ-5D there was a significant association between the mobility domain in the patient global assessment of disease severity VAS defined changes but not for the other external criterion. For the SF-6D, HUI3, and HUI2 there were significant associations for the vitality domain, the dexterity single attribute, and the sensation single attribute, respectively, using the patient global assessment of disease severity VAS defined changes. Of note, for the self-care single attribute in the HUI2, there was a significant association between the patient transition defined changes but not the other criterion.

Discussion

This study is the first to compare the reliability and longitudinal changes in scores obtained with four indirect utility instruments (HUI3, HUI2, EQ-5D, SF-6D), a disease-specific measure (the RAQoL), and a disability measure (the HAQ) in a sample of patients with rheumatoid arthritis. Our results demonstrate that while the generic, preference-based measures yielded scores that were generally reliable, they had lower responsiveness (as assessed

by multiple methodologies) in RA than the disease-specific RAQoL. The indirect utility measures did, however, yield moderate responsiveness statistics when the patient global assessment of disease severity was applied as the external criterion for change. The domains and attributes of the indirect utility instruments that were commonly associated with the external criteria for change in RA tended to be pain, ambulation/physical functioning, and emotional/mental health.

Using the patient transition external criterion, we found that there were fairly large mean differences in the instruments between the time points for individuals who were classified as being the 'same' from their RA perspective (sometimes the change in this category was of similar magnitude as those classified as 'worse' or 'better'). This point was illustrated in the polytomous regression plots where there was considerable overlap between the 'same' and 'better' or 'worse' curves. While this finding could be the result of shortcomings of the instrument in assessing changes in RA, these findings were not observed when a different external criterion was applied (categories based upon the patient global assessment of disease activity VAS). Also, several single attributes that were expected to have significant associations with changes in RA were significantly associated with changes in the patient global assessment VAS and not the patient transition question changes (mobility (EQ-5D), vitality (SF-6) and dexterity (HUI3)). Therefore, categorization of the patient global assessment of disease activity VAS appears to be a superior external criterion for RA than the patient transition question as it was expected that these domains/single attributes would be associated with changes in RA.

Generally, dividing the sample into 'worse', 'same' and 'better' using the patient global assessment of disease severity VAS categories seemed to more accurately define these groups than the patient transition question. This point is illustrated by the larger responsiveness statistics for all of the instruments, the smaller amount of change in all of the instruments in those classified as having their RA being the 'same' as at baseline, and a greater magnitude of change (either negative or positive) in those classified as having their RA 'worse' or 'better' than baseline. Using the transition question as the external criterion resulted in small ES and SRM statistics for virtually all of the instruments for those who reported to have improved or worsened from baseline for many of the indirect utility measurements (Table 2). Conversely, when applying the classification according the patient global assessment of disease severity VAS (Table 2), many of the responsiveness statistics for those classified having their RA improved or worsened over baseline can be interpreted as moderate or large, and all of the paired *t*-tests for those who improved or worsened were significant for all of the instruments.

The indirect utility instruments displayed different properties in this study. Reliability was acceptable for all of the scores except for the EQ-5D. This finding is considerably lower than previously reported in rheumatoid arthritis (ICC of 0.73 using the stable groups approach and 0.78 using test–retest reliability) [9]. The differences in these two findings may be due to the 5 week window for resubmission of the reliability questionnaires in our study compared to two weeks in the other analysis. In the longer time frame, it is possible that there was a higher probability for change. This change may have penalized the EQ-5D much more than the other scales as there is a term in the EQ-5D scoring function (N3) that subtracts 0.269 if a score of the lowest level (3) occurs on at least one domain. Thus, a one category change (from '2' to '3') in response in a single domain can have profound implications for reducing the EQ-5D utility score. However, other instruments which were found to be more responsive than the EQ-5D were stable (the RAQoL and the HAQ) over this time frame.

The HUI2 and the HUI3 generally had low responsiveness statistics utilizing the patient transition question as the external criteria and moderate responsiveness statistics when the categories of the patient global assessment of disease activity VAS were applied. Their relative rankings were towards the middle or bottom for all of the instruments regardless of the external criteria applied accept for the 'better' category as defined by the patient global assessment of disease activity. For this category, the HUI3 had the highest responsiveness statistics in two categories (the ES, and the SRM). This was likely due to the observation that the mean change in this category was quite large (0.17) which was almost half of the

baseline score. In the polytomous regression plots, the HUI3 appeared to have less overlap between the same and the better or worse curves than the other indirect utility instruments (i.e. Figure 2) which may make it more responsive in RA. As expected, the sensation attribute (HUI2), the vision, hearing and speech attributes (HUI3) and the cognition attributes (both scales) were not associated with the external criteria defined change in RA. Of note, although one would have expected dexterity (HUI3) and self-care (HUI2) to be consistently associated with changes in RA, each was only significant for only one of the external criteria.

The SF-6D generally had low responsiveness statistics utilizing the patient transition question as the external criteria and moderate responsiveness statistics when the categories of the patient global assessment of disease activity VAS were applied. This latter finding was especially true for the 'better' category. One of the problems with the responsiveness of the SF-6D when using our external criteria was the amount of change experienced by those categorized as the 'same'. Using each of the external criteria, there was mean change of similar magnitude in those classified as the 'same' and 'better'.

As anticipated, the RAQoL was the most responsive to changes in both positive and negative directions which are in agreement with other research comparing disease-specific to generic HRQL instruments [30]. The responsiveness statistics were generally moderate to large irrespective of the external criteria of change applied. In addition, the results of the polytomous regressions reveals well delineated curves for same, better and worse without a large degree of overlap (Figure 1).

Results for the HAQ revealed that this instrument performed approximately equivalently for both of the external criteria with responsiveness statistics of similar magnitude. However, when compared to the other instruments, the HAQ rankings were among the highest for responsiveness statistics calculated from categories defined by the patient transition question but were either in the middle (for those categorized as worse) or at the bottom (for those categorized as better) for responsiveness statistics calculated from categories defined by the patient global assessment of disease severity VAS. Although the reason for this finding is not obvious, perhaps the patient transition question is capturing mostly changes in elements of disability (as measured by the HAQ) rather than other aspects/domains of RA which are being captured by the other instruments.

In summary, the RAQoL was consistently the most responsive of the tested instruments. Among the indirect utility instrument's overall utility scores, the EQ-5D appeared to be the most responsive to worsening but not to improvement. Conversely, the HUI3 and SF-6D were superior in detecting improvement but the SF-6D detected changes in those classified as the 'same'. Thus, in RA clinical trial situations where a known effective intervention is to be applied and there is a large probability of positive change, the SF-6D and the HUI3 would be superior to the other instruments. However, changes in the SF-6D might be larger as many patients classified as the same by other criteria would, in fact, improve using this scale. The HUI2 appeared to be fairly non-responsive in RA in comparison to the other measures.

We have characterized the responsiveness of the scores of the instruments but, for economic evaluation, the absolute change size (and not just the effect size) matters the most. For example, when used as quality weightings in the estimation of QALYs, the magnitude of the change in the instrument score determines the size of the denominator in the determination of the incremental cost-effectiveness ratio. As such, in our study, it would appear that when applied to a study examining mostly improvement (ie. a study of a new drug therapy), the HUI3 would yield the largest change compared to the SF-6D which was the smallest (0.17 and 0.06 using the patient global assessment criteria). Obviously, these findings have important ramifications for cost-effectiveness analysis and could result in substantial differences in incremental ratios when used within the same model.

We conclude that the reliability of the scores from all the instruments (with the exception of the EQ-5D) was acceptable. Categories defined by the patient global assessment of disease severity appeared to perform better as an external criterion for change in RA than a patient transition question. The RAQoL was the most responsive although all the instruments were capable of detecting change to some degree. The HUI3 and

the SF-6D may be the best indirect utility instruments to use in clinical trials of RA where a known effective intervention is to be applied. The differences in the magnitude of the absolute change scores have important implications for cost-effectiveness analyses.

Acknowledgements

Dr. Carlo Marra was supported by a Canadian Institute of Heath Research/Arthritis Society Fellowship and a Michael Smith Foundation for Health Research Studentship. Project supported by a grant from the Canadian Arthritis Network (a National Centre of Excellence). Dr. Kopec is supported by a Michael Smith Foundation for Health Research Senior Scholar Award.

References

1. American College of Rheumatology Subcommittee on Rheumatoid Arthritis Guidelines for the Management of Rheumatoid Arthritis: 2002 Update. Arthritis Rheum 2002; 46: 326–348.
2. Lipsky PE, van der Heijde DM, St Clair EW, et al. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. N Eng J Med 2000; 343: 1594–1602.
3. Blumenauer B, Cranney A, Clinch J, Tugwell P. Quality of life in patients with rheumatoid arthritis: Which drugs might make a difference? Pharmacoeconomics 2003; 21: 927–940.
4. Scott DL. Leflunomide improves quality of life in rheumatoid arthritis. Scand J Rheumatol Suppl 1999; 112: 23–29.
5. Zhao SZ, Fiechtner JI, Tindall EA, et al. Evaluation of health-related quality of life of rheumatoid arthritis patients treated with celecoxib. Arthritis Care Res 2000; 13: 112–121.
6. Hammond A, Young A, Kidao R. A randomised controlled trial of occupational therapy for people with early rheumatoid arthritis. Ann Rheum Dis 2004; 63: 23–30.
7. Eberhardt K, Duckberg S, Larsson BM, Johnson PM, Nived K. Measuring health related quality of life in patients with rheumatoid arthritis – reliability, validity, and responsiveness of a Swedish version of RAQoL. Scand J Rheumatol 2002; 31: 6–12.
8. Drummond MF, O'Brien B, Stoddart GL, Torrance GW (eds). Methods for the Economic Evaluation of Health Care Programmes. 2nd ed. Oxford Medical Publications, Oxford, 1997.
9. 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: 551–559.
10. 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; 11: 4–12
11. Conner-Spady B, Surez-Almazor ME. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. Med Care 2003; 41: 791–801.
12. Blanchard C, Feeny D, Mahon JL, et al. Is the Health Utilities Index responsive in total hip arthroplasty patients? J Clin Epidemiol 2003; 56: 1046–1054.
13. Terwee CB, Dekker FW, Wiersinga, Prummel MF, Bossuyt PMM. On assessing the responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. Qual Life Res 2003; 12: 349–362.
14. Liang MH, Lew RA, Stucki G, Fortin PR, Daltroy L. Measuring clinically important changes with patient-oriented questionnaires. Med Care 2002; 40 (Suppl): II-45–II-51.
15. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988; 31: 315–324.
16. Wong AL, Wong WK, Harker J, et al. Patient self-report tender and swollen joint counts in early rheumatoid arthritis. Western Consortium of Practicing Rheumatologists. J Rheumatol 1999; 26: 2551–2561.
17. Fortin PR, Abrahomowicz, Clarke AE, et al. Do lupus disease activity measures detect clinically important changes? J Rheumatol 2000: 27; 1421–1428.
18. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements – an illustration in rheumatology. Arch Intern Med 1993; 153: 1337–1342.
19. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: The patient's perspective. J Rheumatol 1993; 20: 557–560.
20. Marra CA, Woolcott JC, Shojania K, et al. An assessment of the construct validity of four indirect utility measures in rheumatoid arthritis. Social Science and Medicine (in press).
21. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003; 56: 317–325.
22. 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: 290–299.
23. Samsa G, Edelman D, Rothman M, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures. A general approach with illustrations to the Health Utilities Index Mark II. Pharmacoeconomics 1999; 15: 141–155.
24. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): Concepts, measurement properties

and applications. Health and Quality of Life Outcomes 2003; 1: 54 (http://hqlo.com/content/1/1/54).
25. Norman GR, Wridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med Care 2001; 39: 1039–1047.
26. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life. The remarkable universality of half a standard deviation. Med Care 2003; 41: 582–592.
27. Cohen J. A power primer. Psychol Bull 1992; 112; 155–159.
28. 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.
29. Cohen J. Statistical Power Analysis for the Behavioural Sciences. 2nd ed. Hillsdale, (NJ): Lawrence Erlbaum Assoc., 1988.
30. Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. J Clin Epidemiol 2003; 56: 52–60.
31. Chang E, Abrahamowicz M, Ferland D, Fortin PR, for CaNIOS Investigators. Comparison of the responsiveness of lupus disease activity measures to changes in systemic lupus erythematosus activity relevant to patients and physicians. J Clin Epidemiol 2002; 55: 488–497.
32. Abrahamowicz M, Ramsay JO. Multicategorical spline model for item response theory. Psychometrika 1992; 57: 5–27.

Address for correspondence: Aslam H. Anis, MHA Program, Department of Health Care and Epidemiology, Faculty of Medicine, University of British Columbia, 620-1081 Burrard Street, Vancouver, B.C., Canada V6Z 1Y6
Phone: +1-604-806-8712 , Fax: +1-604-806-8778
E-mail: aslam.anis@ubc.ca