Rating of Arthritis Health States by Patients, Physicians, and the General Public. Implications for Cost-Utility Analyses

MARIA E. SUAREZ-ALMAZOR and BARBARA CONNER-SPADY

ABSTRACT. We elicited preferences for 2 arthritis health states (mild and severe) using visual analog scales, time tradeoff, and standard gamble by interviewing 104 individuals from the general public, 51 patients with rheumatoid arthritis, and 43 health professionals. The health scenarios were based on attributes described in a health status classification instrument, the EuroQol (EQ-5D). In addition, we compared the ratings in our survey with those obtained for the same scenarios by one of the scoring algorithms used for the EQ-5D (York weights). Statistically significant differences were observed in the ratings of the health scenarios, mostly for the severe vignette. Most of the variability was related to the method employed. The cost-utility ratio for a hypothetical intervention varied according to the method employed to determine the utility of the health states, from $15,000 to $111,000 US per quality adjusted life year (QALY). Patient derived weights resulted in cost-utility ratios that ranged from $39,000 to $222,000. Our findings show that the methodology used to elicit and analyze utilities can have substantial implications in the economic evaluation of interventions for patients with RA. (J Rheumatol 2001;28:648–56)

Key Indexing Terms:
COST-EFFECTIVENESS ANALYSIS PREFERENCES UTILITIES QUALITY OF LIFE

Economic evaluations of health care interventions compare 2 or more alternatives by taking into account benefits and costs. In rational decision making, individuals choose an alternative based on their knowledge about the potential gains or losses associated with the available options and their individual preferences. In health, the benefits relate to gains in health status and the risks to losses in health or commodities. Cost-utility analysis incorporates the preferences or values that individuals have for particular health states to compare benefits and costs[1]. Preference elicitation can be used not only to determine preferences for specific health interventions, but also to evaluate quality of life. Health related quality of life is measured in a number of ways. Self-report instruments, such as the Medical Outcomes Study Short-Form (SF-36) or the Sickness Impact Profile, include a number of items weighted according to algorithms that are applied in similar fashion to every individual[2]. These questionnaires do not capture the individual value that a given respondent may assign to a particular health state, and 2 individuals may rate differently the same health state depending on the value they assign to a symptom or impairment and their willingness to accept tradeoffs between benefits and risks.

Several methods have been used to measure health preferences or utilities[1]. The most widely used techniques are rating scales, such as visual analog scales (VAS), time tradeoff (TTO), and standard gamble (SG). All these measures provide a final score where 1 is perfect health and 0 is the worst possible imaginable state or death; occasionally, states considered to be worse than health are rated below zero (negative values). Ratings can be elicited from different groups of individuals such as patients, health professionals, or the public. The values obtained can be used to adjust for the quality of life associated with a particular health state and to calculate measures commonly used in cost-utility analyses such as quality adjusted life years (QALY).

Visual analog scales. VAS are the simplest measures. The individual is asked to place a mark on a line of a given length that reflects the value of a health state with given attributes. The 2 extremes of the line are anchored as 0 and 1, as described above.

Time tradeoff. The TTO method was developed specifically for use in health care. The subject is offered 2 alternatives: (1) living the life of an individual in a health state X for time T1 followed by death, and (2) being healthy for time T2 (with T2 < T1) followed by death. The time is varied until the subject is indifferent between the 2 alternatives.

Standard gamble. With SG the subject is offered 2 alternatives. Alternative 1 is a treatment with 2 possible outcomes:

From Health Services Research, Baylor College of Medicine, Veteran Affairs Medical Center, Houston, Texas, USA.

Dr. Suarez-Almazor was funded by The Arthritis Society and the Alberta Heritage Foundation for Medical Research.

M.E. Suarez-Almazor, MD, PhD; B. Conner-Spady, PhD.

Address reprint requests to Dr. M.E. Suarez-Almazor, Health Services Research, Baylor College of Medicine, Veteran Affairs Medical Center (Station 152), 2002 Holcombe Blvd., Houston, TX 77030.
E-mail: mes@bcm.tmc.edu

(1) the patient regains perfect health and lives for an additional T years (probability p), or (2) the patient dies immediately (probability 1 – p). Alternative 2 has a certain outcome of a particular health status for life. The probabilities are varied until the subject is indifferent between the 2 alternatives. This is the only method that incorporates uncertainty about the outcomes, and captures risk attitudes[1]. In rating scales no choice is made, and in TTO the choices are based on certain outcomes.

Preference based quality of life scores can be elicited directly from patients using any of the techniques described above. Utilities for health states can also be obtained indirectly by asking subjects to rate hypothetical health states using these methods. It has been argued that, for economic evaluations resulting in health policy decisions and allocation of resources, the utility of health states should be determined by the public[3]. Various instruments have been developed resulting in a large number of generic health profiles that have been assigned a utility score by the general population. Typically, these tools include several attributes (e.g., pain, function, etc.) with mutually exclusive ordinal levels of functioning or distress. The health states based on these descriptors have been valued by the public using one of the methods described above. When the questionnaire is administered it results in a descriptive profile of the health state of the respondent. A score is then assigned to each profile based on the utilities obtained from the public. Several preference based classification systems have been developed to measure the quality of life of the respondent on the basis of valuations by other individuals: the Rosser Index, the Quality of Well-Being Scale, the EuroQol (EQ-5D), and the Health Utilities Index[4-7].

Previous studies have shown discrepancies in utility scores according to the group assigning values (e.g., patients vs public) and the technique used (e.g., rating scales vs standard gamble). Further, the findings do not appear to be consistent across diseases. Many studies have been conducted for conditions such as heart disease or cancer, where the value individuals assign to survival may play a role. Although preference elicitation techniques to measure quality of life generally should not incorporate survival (to avoid double counting when combining quality of life with life expectancy), it is unclear whether knowledge about survival prognosis for a given condition influences the ratings. Our objective was to assess the potential effect that utility ratings by different groups using different elicitation techniques would have on hypothetical cost-utility analyses of interventions for rheumatoid arthritis (RA). We compared valuations of arthritis health states by RA patients, health professionals, and the general population. We also obtained direct valuations of the patients' own health status and compared these utilities to those obtained for the same individuals using a preference based classification tool (EQ-5D).

METHODS

We conducted face-to-face interviews of 3 groups of individuals from the Edmonton region (Alberta, Canada). The surveys were conducted in 1999:

General public. One hundred four subjects were recruited through random digit dialing. The sampling was stratified according to sex and age to ensure that the participants were representative of the population of Edmonton with respect to age and sex.

Patients with RA. A convenience sample of 51 patients with RA was identified from medical records of patients attending rheumatology clinics at the University of Alberta. Patients were telephoned or approached at the clinic, and if they agreed to participate, an interview was scheduled at the clinic or their home.

Health professionals. A convenience sample of 43 health professionals participated in the survey: 22 were physicians (rheumatologists and primary care physicians) and 21 were physical and occupational therapists and orthopedic nurses. Health professionals were interviewed at their work site.

Survey

The preference elicitation surveys of the 3 groups were similar, conducted as face-to-face interviews using props developed by McMaster University[8]. Two hypothetical health states were developed using the attributes of the EQ-5D[7]. The EQ-5D is a preference based classification system that includes 5 attributes: mobility, self-care, activity, pain, and depression or anxiety (the latter 2 attributes are evaluated together). Each of these domains has 3 possible levels that correspond to descriptions of no impairment, mild to moderate impairment, and severe impairment. The 2 hypothetical scenarios in this study describe persons with arthritis with severe and mild disease.

Severe. A person with severe arthritis has problems walking about, problems with self-care such as washing or dressing, problems performing daily activities such as work, study, housework, family or leisure activities, has extreme pain or discomfort, is moderately anxious or depressed.

Mild. A person with mild arthritis has problems walking about, but not with self-care such as washing or dressing, and no problems performing daily activities such as work, study, housework, family or leisure activities, has moderate pain or discomfort, but is not anxious or depressed.

Respondents were asked to imagine how it would be to spend the rest of their life like these individuals. For the public and patients we used the general descriptor "arthritis," and for health professionals "rheumatoid arthritis." Each respondent was asked to rate the scenarios in the same order, severe followed by mild, with the following elicitation techniques:

Visual analog scale. We used the EQ-5D feeling thermometer as a VAS. The thermometer is a 20 cm vertical

scale. We anchored the extremes at 0 (death) and 100 (perfect health). A person with perfect health was described as someone with no problems walking about, no problems with self-care such as washing or dressing, no problems performing usual activities such as housework, study, work, family or leisure activities, has no pain or discomfort, is not anxious or depressed.

Standard gamble. Respondents were asked to choose between 2 alternatives. The first one had a certain outcome: to live the rest of their life in the hypothetical scenario. The second alternative was uncertain: to receive an imaginary treatment that offered a chance at *perfect health* with a probability p and a chance of death with a probability 1 – p. We used a ping-pong technique starting at p = 1 and 1 – p = 0, followed by p = 0 and 1 – p = 1, p = 0.95 and 1 – p = 0.05, and so forth.

10-year time tradeoff. Respondents were asked to imagine that they had 10 years to live, after which they would die an immediate painless death, and were given 2 certain choices, live the remaining 10 years of their life with the health depicted in the hypothetical scenario or live a shorter period of time with perfect health. The shorter period of time in perfect health was modified by 6 months each time using a ping-pong technique: 9 years 6 months, followed by 6 months, then by 9 years, etc.

In addition, RA patients and the public were asked to rate their health using the same methods. In this case, instead of imagining that they had the health status of the hypothetical subjects, they gave their preferences for remaining in their current health. They also completed EQ-5D profiles. We imputed values for their health state profiles based on general population valuation weights (York social tariff) for the EQ-5D[9].

York weights (social tariff). The York weights were developed in the UK through a survey of the general population using TTO methods. Two hundred forty-five mutually exclusive states can be derived from combinations of EQ-5D levels for each attribute. Since not all these states have been individually assessed a limited number of profiles, those reflecting the most plausible scenarios, were rated. The final EQ-5D York weights for all 245 states are based on a linear model that includes the TTO valuations as the dependent variable and the attribute levels as independent variables. The coefficients for each attribute/level are "disutilities," which are subtracted from 1. An interaction term is included if any of the attributes is scored at the lowest level of impairment. The range of possible values derived from this linear model is –0.59 to 1. Zero is death, states below zero are considered to be worse than death, and 1 is best imaginable perfect health.

Statistical analysis. All the utility scores elicited from the subjects in our study were scaled from 0 to 1, with 0 representing death and 1 perfect health as defined above. With SG and TTO a few respondents had responses inconsistent with our framework, scoring a scenario better than perfect health, or worse than death. Worse than death scenarios are not necessarily inconsistent, but since one of our objectives was to compare different techniques within the perfect health and death anchors we did not elicit additional utilities to evaluate these states (which would then have been assigned negative values). These values were analyzed in 2 different ways: as missing values, and rescored to the closest anchor (0 or 1). No major differences were observed and the results presented here include rescored values.

Comparisons of hypothetical scenarios across raters and methods. The utilities obtained with these methods are generally assumed to be interval scaled and are often used in linear models. However, we used both parametric and nonparametric techniques since the TTO and SG scores were not normally distributed. Generalized linear models were used to evaluate the combined effects of rater and technique, although some of the parametric assumptions of these models were not fully met. We used a repeated measures model, with technique as a within-subjects factor and rater as between-subjects factors, including a rater by technique interaction term.

Patients' rating of their own health. We compared the patients' direct ratings using VAS, TTO, and SG and the indirect utility scores calculated from their EQ-5D profile using York weights. Parametric and nonparametric repeated measures techniques were used to evaluate statistical differences.

Hypothetical cost-utility analyses. We evaluated the effect of different ratings on a hypothetical cost-utility analysis of an intervention for RA with 2 models. The first one was based on the hypothetical scenarios, evaluating the transition from the severe health state to the milder one (alternative 1) versus remaining in the severe state (alternative 2), using the utilities obtained for the scenarios from the various groups, and with the different methods. We analyzed the data using both the mean and the median values because the distribution of scores was often asymmetric and mean scores are strongly influenced by extreme values. The second analysis was based on the actual health state ratings of the patients with RA. We categorized the patient group across the VAS median rating of their own current health. We used the VAS because it had the most symmetrical distribution of all the ratings. We then assumed a transition from the group below the median (severe) to the group above the median (mild) (alternative 1), using the group means and medians for the various techniques, compared to remaining in the severe state (alternative 2). In both cases we arbitrarily chose a 5 year period, with the health gains linearly accrued at the end of the first year and patients remaining stable thereafter. The additional cost of the intervention required to achieve the mean transition in both examples was estimated at $50,000 US over the 5 years.

Although this hypothetical cost exceeds the expenditures related to most treatments currently used for RA, we modeled our examples for newer interventions such as biologic or cell therapies, which are substantially more costly than traditional second line therapies. The cost-utility ratio represents cost per QALY gained over the 5 year period. Costs and benefits were not discounted in the analysis.

RESULTS

Characteristics of the participants are shown in Table 1. Patients were older and less educated than the other groups, and as expected, had a worse health status than the general public.

Hypothetical scenarios

Illogical paired ratings (rating of severe scenario higher than the milder one) were not observed for health professionals or RA patients. For the public, 3 VAS, 2 TTO, and 5 SG paired ratings were illogical.

Table 1. Characteristics of the groups surveyed.

	General Public, n = 104	Patients with RA, n = 51	Health Professionals, n = 43
Mean age (SD)*, yrs	46.0 (17.1)	60.3 (15.0)	42.0 (8.2)
Female (%)*	55 (53)	37 (72)	28 (65)
Education level (%)*			
< grade 12	15 (15)	11 (22)	—
Completed high school	17 (16)	13 (25)	—
Post-secondary	71 (69)	27 (53)	43 (100)
Mean overall health status (EQ-5D VAS)*	87.5 (15.6)	67.1 (18.5)	N/A

*p ≤ 0.05.

The distribution of the scores was grossly normal for the VAS in all 3 groups. The TTO and SG score distributions were similar in shape for the public and RA patients: for the severe scenario, the distributions were bimodal (U shaped), and for the milder one markedly skewed to the left. Health professional ratings for TTO and SG for both scenarios were markedly skewed to the left. Table 2 shows the utility scores by group and method. Utility scores for the severe scenario derived through TTO and SG varied widely among individuals within the same group as can be seen by the width of the interquartile range. Statistically significant differences between raters (nonparametric tests) were only observed for the VAS and SG utilities for the severe scenario. The major differences were observed between the public and health professionals, with patient ratings usually in between. Significant differences across methods were observed for all groups. In general, the highest utilities were obtained with SG. All the mean ratings in our study were substantially higher than ratings for the same health states derived from the EQ-5D York algorithm. These differences were particularly striking for the severe scenario, where the York score for the severe state is negative, indicating that it is valued as worse than death. All the mean ratings for this state in our study were above 0.4.

In the generalized linear models, significant differences were observed between techniques. The interaction term technique by rater was also statistically significant, indicating that the differences between techniques varied according to the rater group. The overall differences between raters (without considering specific techniques) did not reach statistical significance.

Figure 1 shows the potential results of a cost-utility analysis using the various ratings: Figure 1A shows the

Table 2. Utility ratings of hypothetical scenarios by group and technique.

	General Public, n = 104		Patients with RA, n = 50		Health Professionals, n = 43		
	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	Mean (SD)	Median (IQ range)	p†
Severe scenario							
VAS	0.51 (0.15)	0.50 (0.40–0.65)	0.44 (0.17)	0.42 (0.30–0.60)	0.45 (0.14)	0.45 (0.35–0.60)	0.03*
TTO	0.54 (0.37)	0.70 (0.11–0.85)	0.58 (0.35)	0.72 (0.22–0.86)	0.55 (0.28)	0.60 (0.35–0.75)	> 0.20
SG	0.56 (0.31)	0.65 (0.25–0.78)	0.66 (0.32)	0.82 (0.45–0.95)	0.74 (0.27)	0.85 (0.60–0.95)	0.03*
York weights for scenario	–0.02		–0.02		–0.02		
p†		0.02*		< 0.001*		< 0.001*	
Mild scenario							
VAS	0.77 (0.11)	0.80 (0.70–0.85)	0.76 (0.10)	0.75 (0.70–0.85)	0.76 (0.08)	0.80 (0.70–0.80)	> 0.20
TTO	0.83 (0.23)	0.95 (0.80–0.95)	0.81 (0.28)	0.95 (0.85–0.95)	0.80 (0.23)	0.85 (0.75–0.95)	> 0.20
SG	0.79 (0.22)	0.85 (0.75–0.95)	0.84 (0.20)	0.95 (0.75–0.95)	0.88 (0.18)	0.95 (0.95–0.95)	0.15
York weights for scenario	0.73		0.73		0.73		
p†		< 0.001*		< 0.001*		< 0.001*	

IQ: interquartile.
†p values based on non-parametric tests (not including York weights).
*statistically significant.

results using mean ratings and 1B median ratings. For the ratings in our study the gains in QALY over the 5 year period ranged from 0.45 to 1.58. The cost-utility ratios ranged from about $32,000 to $111,000 US per QALY. Using the York weights, 3.38 QALY were gained, with a cost-utility ratio of about $15,000. The major differences in our data were related to technique, more than to the rater. The lowest cost-utility ratios were observed for VAS, followed by TTO, and then SG. Major differences among raters were only observed for SG, where the ratios substantially increased in this order: public, patients, and professionals. This was largely driven by the ratings of the severe scenario. The differences were more pronounced when using median ratings instead of mean ratings. The most striking differences were observed with the York weights. The cost-utility ratio with this technique was less than half the lowest ratio and one-seventh of the highest one.

Patient ratings of their own health

Table 3 shows the mean and median direct patient ratings of their own health using the 3 methods, and the mean EQ-5D York scores for their individual profiles. Results are shown for the overall group and for the groups categorized by their median score in their VAS rating. Statistically significant differences were observed with both parametric and nonparametric repeated measures analysis. The York scores

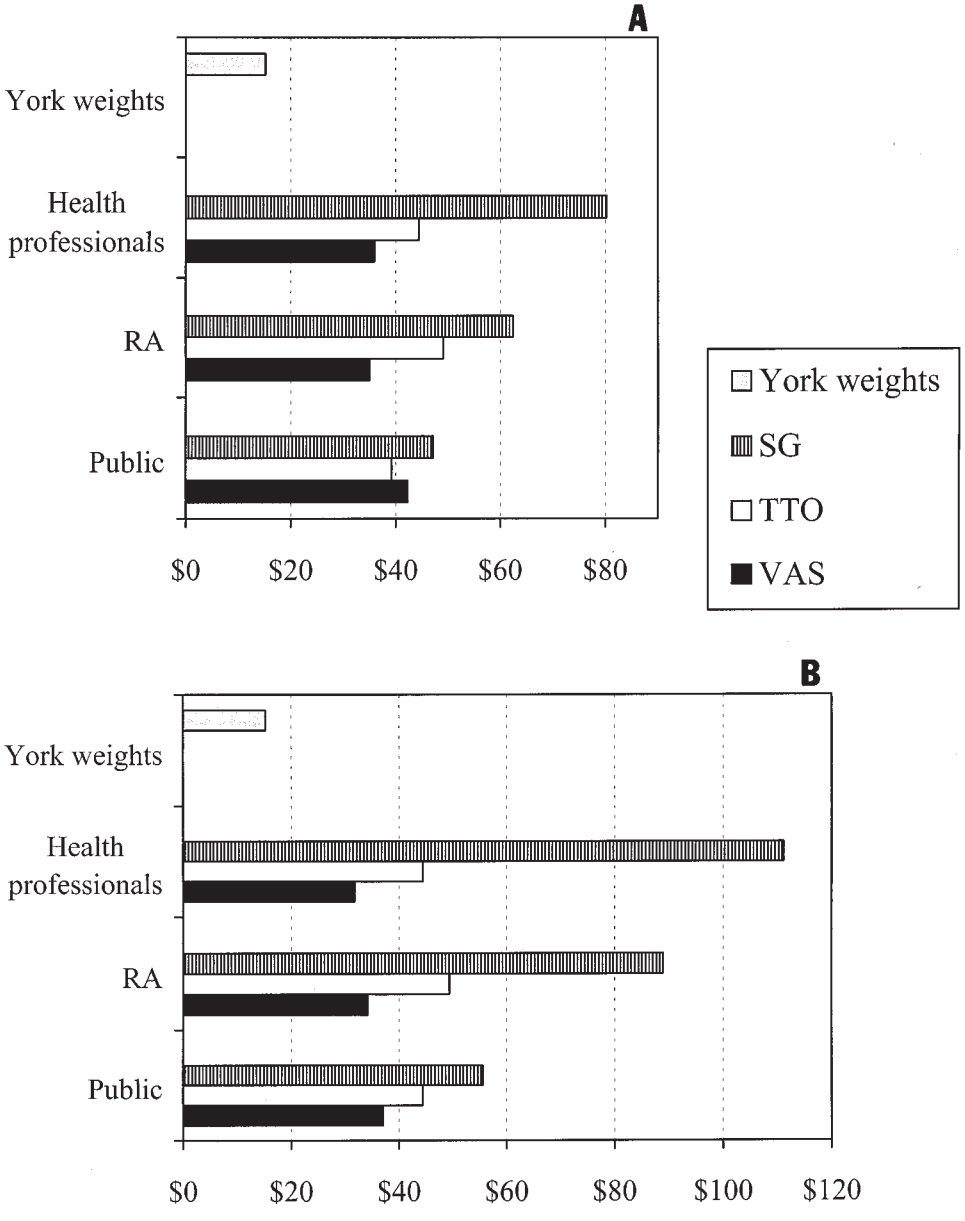

Figure 1. Hypothetical cost-utility analysis based on ratings of the health scenarios by the public, patients with RA, and health professionals. A: using mean ratings; B: using median ratings.

Table 3. Direct and indirect* patient ratings of their own health status.

Method†	Mean (SD)	Median† (IQ range)
VAS	0.67 (0.18)	0.71 (0.55–0.81)
TTO	0.76 (0.32)	0.95 (0.68–0.95)
SG	0.78 (0.26)	0.90 (0.75–0.95)
EQ-5D York weights	0.56 (0.28)	0.62 (0.52–0.69)

* EQ-5D York weights. IQ: interquartile.
†$p < 0.001$, nonparametric test.

were significantly different from all other utilities. VAS scores were significantly different than SG ($p < 0.001$), and had borderline significance compared to TTO ($p = 0.07$). No differences were observed between TTO and SG.

Figure 2 shows the hypothetical cost-utility scenario comparing the transition from the worse state (below median) to the better one (above median) with remaining in the lower health state. The difference between these 2 states was less marked than for the hypothetical scenarios; so overall, the cost-utility ratios were higher in this analysis. No major differences were observed between the York scores and VAS, but substantially higher cost-utility ratios were obtained for SG and TTO. The differences were larger for estimations using median values. Hypothetical gains in QALY ranged from 0.63 to 1.28 when using mean utility scores and from 0.23 to 1.13 when using median scores. These differences were more pronounced for SG and TTO and reflect the skewed distributions of these measures and the large influence of extreme values.

DISCUSSION

The objective of this study was to evaluate the potential empirical implications of using different raters and methods to evaluate quality of life in cost-utility analyses of interventions for patients with RA. We used 2 approaches. For the first one we elicited utilities for hypothetical scenarios from different groups: the public, patients with RA, and health professionals. Ratings of hypothetical scenarios are often used for economic modeling when patient data are not available, or when valuations by the public or a particular group are desired from a policy perspective. The second approach evaluated the differences in the patients' direct ratings of their own health with the various methods. Direct patient derived utility scores are increasingly being incorpo-

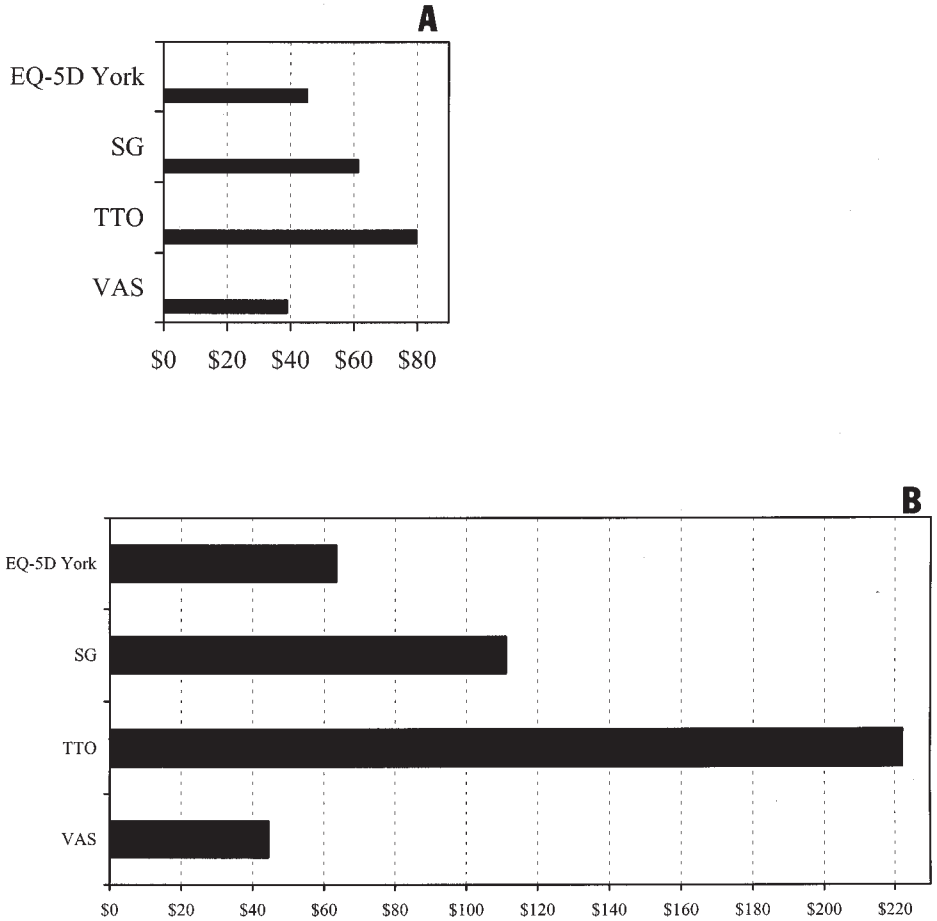

Figure 2. Hypothetical cost-utility analysis based on patient direct ratings of their own health status. A: using mean ratings; B: using median ratings.

rated in clinical trials as an outcome to evaluate quality of life, and to conduct cost-utility analyses of the interventions being investigated.

We found significant differences in the ratings of the health scenarios, mostly in relation to the utility of the severe state. For this state, subjects varied widely in their ratings. For instance, for SG and TTO there was more variation among patients with RA when rating the severe scenario (single health state), than when rating their own health (multiple health states). Overall, the major differences could be attributed to technique more than rater, although a significant interaction between these factors was observed, mostly due to SG, which resulted in higher valuations of health states by professionals compared to the public, with patient ratings in between. A number of studies have compared utilities by different rater groups, with conflicting results[10-27]. Most often, but not always, patients rate health states higher than the public (or patients with milder disease or other disorders) and health professionals tend to assign higher utilities than patients. Our findings suggest that differences between raters are more marked for severe states and also for SG and TTO methods. We are only aware of one study comparing patients with RA and the public, which used rating scales and found no substantial differences, with a high correlation between the scores of $r = 0.97$[12]. Some of the small differences between raters may have been related to the differences in age and education level of the participants; however, our objective was not to identify what determined the valuations, but to examine the empirical consequences of using different raters. We therefore decided not to adjust the results for these differences in an attempt to replicate the common use of these values in research (unadjusted for demographic characteristics).

In our study, the major determinant of differences in ratings was the method used to elicit utilities. VAS scores were consistently lower than TTO or SG ratings, for both the rating of vignettes and the patients' own health. A number of studies have compared elicitation techniques rating hypothetical scenarios or the respondent's own health, with varying results[26,28-47]. Most show differences in ratings, generally, with TTO resulting in higher utilities than rating scales, and SG values higher than those elicited by the other 2 techniques. These findings are in accordance with the theoretical constructs underlying these measures. TTO and SG introduce an additional dimension to rating scales by offering a choice between 2 alternatives (sacrifice). In addition, SG also incorporates tolerance to risk since one of the alternatives is always uncertain with respect to possible outcomes. Only a few of these studies have been conducted in patients with rheumatic diseases, including rheumatoid arthritis[28,29], ankylosing spondylitis[30,31], fibromyalgia[30,32,33], low back pain[33], lupus[34], and osteoporosis[26]. The majority have shown significant differences across methods[29-32]. Studies comparing utility measures to other instruments evaluating disease severity or functioning show that VAS correlates better with these measures than TTO or SG[20,31,32,34,37,48-50]. This would suggest that VAS measures attributes similarly to other generic health instruments (non-utility based), but TTO and SG incorporate preferences in a way that may better reflect decision-making processes. A major limitation of TTO and SG methods is their unusual distribution (skewed or bimodal), which results in substantial weight from extreme values. The implications are not only statistical, limiting the types of analyses that can be performed, but also ethical, since using mean or linearly predicted scores places more value in the utilities assigned by outliers and extreme raters. It is also unclear whether the resulting scale is truly an interval scale, which is a major assumption underlying the estimation of QALY.

Our hypothetical scenarios were based on EQ-5D profiles, and therefore we were able to compare our results with those indirectly derived from this instrument. For this study we used the TTO social tariff scoring system developed in York[9]. The lowest utility in our study was the one derived from the York algorithm for the severe scenario, which resulted in a negative value, worse than death. Zethraeus, *et al* also compared imputed public values and direct patient TTO utilities in women with menopausal symptoms[51]. As in our study, the discrepancies were higher for severe states. The York scoring system is based on linear regression modeling of the individual attribute levels on the TTO scores that the public assigned to the various profiles tested. This modeling process is based on parametric assumptions, linearity, equality of variances across states, and interval properties of the final score. However, our data show that TTO scores are not normally distributed and are markedly skewed to the left, resulting in increased weight in the mean level estimates from the extreme valuations. Furthermore, the variance appears to be larger for severe scores. Using the York approach we regressed our patients' EQ-5D profiles on their TTO utility scores (results not shown). We obtained a rating for the severe scenario that was close to 0 (death); for the milder scenario the utility was 0.90. Because we only included 50 patients and many of the states were not represented in the patient profiles, the results of the regression are not robust, but illustrate how the utility of poor health scenarios can be underestimated using this regression approach. A major difference between the York scores and those in our survey is that they allowed states worse than death, which we did not, since we anchored our scale at 0 and 1. The objective of our study, however, was to conduct an empirical analysis of potential implications in economic evaluations and not a thorough analysis of the EQ-5D.

Comparisons between direct patient utilities and scores derived from other classification systems (Health Utility Index and Quality of Well-Being) have also been reported in selected rheumatic diseases[12,26,34]. Utilities derived from the

Quality of Well-Being questionnaire were lower than direct measures in one study[26] but not in another[12]. The Health Utility Index produced utilities closer to those obtained from patients[26,34]. Our study only examined the EQ-5D York social tariff weights, and we cannot generalize our findings to other classification systems.

Some of the differences we observed can perhaps be attributed to the measurement properties of the utility methods more than to their underlying constructs. The objective, however, was not to perform an in-depth analysis of these properties but to empirically evaluate the implications for cost-utility analyses if different raters or methods are used. Consistent and proportional differences among techniques or raters would not be as important from an economic evaluation perspective since the effect on the cost-utility ratios would be small (for instance if the EQ-5D was consistently lower but with the same absolute difference between states as the other ratings). Our results suggest, however, that the implications for a cost-utility analysis are substantial, particularly in relation to the technique employed and the point estimate (mean or median) of the utility score for a given state. Differences among raters had a major effect for the SG results, but not so much for the other techniques. A previous study found that the costs per QALY in a cost-utility model of a hypothetical intervention to prevent osteoporotic fractures would have ranged from $25,000 to $94,000 depending on whether the women rating the health states had a previous fracture or not[26]. The authors used TTO, which, in our study, did not differ to a great extent between raters. We found that the use of York weights resulted in a considerable underestimate of health costs compared to the other measures. These weights are based on societal ratings and may not be directly compared to patient or health professional ratings. Nevertheless, substantial differences were observed comparing these results to those obtained in our general population sample. This is a concern, since this instrument is increasingly being used in clinical and health services research. Large differences were also observed in the cost-utility analysis using patient ratings of their own health. Depending on the preference elicitation technique and the statistical method used, the incremental ratio per QALY gained ranged from about $40,000 to $220,000. This finding has implications for studies such as clinical trials where direct measures and preference classification systems such as the EQ-5D are obtained from patients.

In summary, our findings show that the methodology used to elicit and analyze utilities can have substantial implications for the economic evaluation of interventions for RA. The discrepancies suggest that the results of different cost-utility studies can only be compared if the same elicitation method and analysis strategy have been used. Our results do not allow us to recommend a method or rater group, but suggest that SG techniques may be less reliable, with higher variability within and between raters. Given the potential policy and clinical consequences of these differences, additional research is necessary to evaluate the measurement properties and underlying constructs of utility tools in the assessment of health related quality of life.

ACKNOWLEDGMENT

The authors acknowledge Donna Fong and the Population Research Laboratory at the University of Alberta for their assistance with the data collection.

REFERENCES

1. Drummond MF, O'Brien B, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. 2nd ed. Oxford: Oxford University Press; 1997.
2. McDowell I, Newell C. Measuring health. A guide to rating scales and questionnaires. 2nd ed. New York: Oxford University Press; 1996.
3. Gold M. Panel on cost-effectiveness in health and medicine. Med Care 1996;34 Suppl:DS197-9.
4. Rosser R, Kind P. A scale of valuations of states of illness: is there a social consensus? Int J Epidemiol 1978;7:347-58.
5. Kaplan RM, Anderson J. A general health policy model: update and applications. Health Ser Res 1988;23:203-35.
6. Brook R, with the EuroQol group. EuroQol: the current state of play. Health Policy 1996;37:53-72.
7. Feeny DH, Torrance GW, Furlong WJ. Health utilities index. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. 2nd ed. Philadelphia: Lippincott-Raven; 1996:239-52.
8. Furlong WJ, Feeny DH, Torrance GW, Barr R, Horsman J. Guide to design and development of health state utility instrumentation. CHEPA Working Paper 90-9. Hamilton, Canada: McMaster University; 1990.
9. Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35:1095-108.
10. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chronic Dis 1978;31:697-704.
11. Patrick DL, Sittampalam Y, Soerville S, Carter W, Bergner M. A cross-cultural comparison of health state values. Am J Public Health 1985;75:1402-7.
12. 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. Medical Care 1986;24:973-80.
13. Froberg DG, Kane RL. Methodology for measuring health-state preferences. III. Population and context effect. J Clin Epidemiol 1989;42:459-71.
14. Kramer MS, MacLellan AM, Ciampi A, Etezadi-Amoli J, Leduc DG. Parents' vs physicians' utilities (values) for clinical outcomes in potentially bacteremic children. J Clin Epidemiol 1990; 43:1319-25.
15. Mazur DJ, Hickam DH. Treatment preferences of patients and physicians: influences of summary data when framing effects are controlled. Med Decis Making 1990;10:2-5.
16. Slevin ML, Stubbs L, Plant HJ. Attitude to chemotherapy: comparing views of patients with cancer with those of doctors, nurses and general public. BMJ 1990;300:1458-60.
17. Llewellyn-Thomas HA, Sutherland HJ, Tritchler DL, et al. Benign and malignant breast disease: the relationship between women's health status and health values. Med Decis Making 1991;11:180-8.
18. Ashby J, O'Hanlon M, Buxton MJ. The time trade-off technique: how do the valuations of breast cancer patients compare to those of other groups? Qual Life Res 1994;3:257-65.

19. van der Donk J, Levendag PC, Kuijpers AJ, et al. Patient participation in clinical decision-making for treatment of T3 laryngeal cancer: a comparison of state and process utilities. J Clin Oncol 1995;13:2369-78.
20. Bosch JL, Hunink MG. The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. Med Decis Making 1996;16:217-25.
21. Molzahn AE, Northcott HC, Dossetor JB. Quality of life of individuals with end stage renal disease: perceptions of patients, nurses, and physicians. ANNA J 1997;24:325-33.
22. Bennett CL, Chapman G, Elstein AS, et al. A comparison of perspectives on prostate cancer: analysis of utility assessments of patients and physicians. Eur Urol 1997;32 Suppl 3:86-8.
23. Jalukar V, Funk GF, Christensen AJ, Karnell LH, Moran PJ. Health states following head and neck cancer treatment: patient, health-care professional, and public perspectives. Head Neck 1998;20:600-8.
24. Lenert LA, Treadwell JR, Schwartz CE. Associations between health status and utilities implications for policy. Med Care 1999;37:479-89.
25. Hallan S, Asberg A, Indredavik B, Wideroe TE. Quality of life after cerebrovascular stroke: a systematic study of patients' preferences for different functional outcomes. J Intern Med 1999;246:309-16.
26. Gabriel SE, Kneeland TS, Melton LJ, Moncur MM, Ettinger B, Tosteson AN. Health-related quality of life in economic evaluations for osteoporosis: whose values should we use? Med Decis Making 1999;19:141-8.
27. Dolan P. Whose preferences count? Med Decis Making 1999;19:482-6.
28. Verhoeven AC, Bibo JC, Boers M, Engel GL, van der Linden S. Cost-effectiveness and cost-utility of combination therapy in early rheumatoid arthritis: randomized comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone. Br J Rheumatol 1998;37:1102-9.
29. Peeters HR, Jongen-Lavrencic M, Bakker CH, Vreugdenhil G, Breedveld FC, Swaak AJ. Recombinant human erythropoietin improves health-related quality of life in patients with rheumatoid arthritis and anaemia of chronic disease; utility measures correlate strongly with disease activity measures. Rheumatol Int 1999;18:201-6.
30. Rutten-van Molken M, Bakker CH, van Doorslaer EKA, van der Linden S. Methodological issues of patient utility measurement. Experience from two clinical trials. Medical Care 1995;33:922-37.
31. Bakker C, Rutten-van Molken M, Hidding A, van Doorslaer E, Bennett K, van der Linden S. Patient utilities in ankylosing spondylitis and the association with other outcome measures. J Rheumatol 1994;21:1298-304.
32. Bakker C, Rutten-van Molken M, van Santen-Hoeufft M, et al. Patient utilities in fibromyalgia and the association with other outcome measures. J Rheumatol 1995;22:1536-43.
33. Goossens ME, Vlaeyen JW, Rutten-van Molken M, van der Linden SM. Patient utilities in chronic musculoskeletal pain: how useful is the standard gamble method? Pain 1999;80:365-75.
34. Moore AD, Clarke AE, Danoff DS, et al. Can health utility measures be used in lupus research? A comparative validation and reliability study of 4 indices. J Rheumatol 1999;26:1285-90.
35. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. Describing health states. Methodologic issues in obtaining values for health states. Med Care 1984;22:543-52.
36. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. Med Decis Making 1984;4:315-29.
37. Tsevat J, Goldman L, Soukup JR, et al. Stability of time-tradeoff utilities in survivors of myocardial infarction. Med Decis Making 1993;13:161-5.
38. Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. Med Decis Making 1994;14:82-90.
39. Rutten-van Molken MP, Custers F, van Doorslaer EK, et al. Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. Eur Respir J 1995;8:888-98.
40. Zug KA, Littenberg B, Baughman RD, et al. Assessing the preferences of patients with psoriasis. A quantitative, utility approach. Arch Dermatol 1995;131:561-68.
41. Stiggelbout AM, Eijkemans MJ, Kiebert GM, Kievit J, Leer JW, de Haes HJ. The 'utility' of the visual analog scale in medical decision making and technology assessment. Is it an alternative to the time trade-off? Int J Technol Assess Health Care 1996;12:291-8.
42. Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: what lies behind the numbers? Soc Sci Med 1997;45:1289-97.
43. Krabbe PF, Essink-Bot ML, Bonsel GJ. The comparability and reliability of five health-state valuation methods. Soc Sci Med 1997;45:1641-52.
44. Hurny C, van Wegberg B, Bacchi M, et al. Subjective health estimations (SHE) in patients with advanced breast cancer: an adapted utility concept for clinical trials. Br J Cancer 1998; 77:985-91.
45. Ubel PA, Loewenstein G, Scanlon D, Kamlet M. Value measurement in cost-utility analysis: explaining the discrepancy between rating scale and person trade-off elicitations. Health Policy 1998;43:33-44.
46. Giesler RB, Ashton CM, Brody B, et al. Assessing the performance of utility elicitation techniques in the absence of a gold standard. Medical Care 1999;37:580-8.
47. Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test-retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. Qual Life Res 1999;8:303-10.
48. Lalonde L, Clarke AE, Joseph L, Mackenzie T, Grover SA. Comparing the psychometric properties of preference-based and nonpreference-based health-related quality of life in coronary heart disease. Canadian Collaborative Cardiac Assessment Group. Qual Life Res 1999;8:399-409.
49. Llewellyn-Thomas HA, Thiel EC, McGreal MJ. Cancer patients' evaluations of their current health states: the influences of expectations, comparisons, actual health status, and mood. Med Decis Making 1992;12:115-22.
50. Revicki DA, Shakespeare A, Kind P. Preferences for schizophrenia-related health states: a comparison of patients, caregivers and psychiatrists. Int Clin Psychopharmacol 1996;11:101-8.
51. Zethraeus N, Johannesson M. A comparison of patient and social tariff values derived from the time trade-off method. Health Econ 1999;8:541-5.