Patient Utilities in Fibromyalgia and the Association with Other Outcome Measures

CARLA BAKKER, MAUREEN RUTTEN, MARIJKE van SANTEN-HOEUFFT, PAULIEN BOLWIJN, EDDY van DOORSLAER, KATHRYN BENNETT, SJEF van der LINDEN

ABSTRACT. ***Objective.*** To compare in patients with fibromyalgia (FM) utilities derived by rating scale and standard gamble methods; to gain insight into construct validity by relating utility values to other outcome measures; to assess the sensitivity to change of utilities.

Methods. A total of 73 patients with FM were randomized into one of 3 groups: low impact fitness training, biofeedback, or controls. At baseline and after 6 mo the Maastricht Utility Measurement Questionnaire was applied. By means of both the rating scale and standard gamble method patients were asked to value their own health status. Construct validity of patient utility measurements was evaluated by Spearman correlation and multiple regression of baseline values with pain, stiffness, patient's global assessment, Sickness Impact Profile (SIP), modified Health Assessment Questionnaire and Arthritis Impact Measurement Scale (AIMS). Sensitivity to change was assessed against changes in these outcomes.

Results. Rating scale utilities correlated significantly ($p < 0.05$) with patient's global assessment ($r_s = 0.53$), pain ($r_s = -0.47$), SIP ($r_s = -0.43$), and with 9 of 11 dimensions of the AIMS (r_s ranging from 0.23 to 0.62). Standard gamble utilities correlated significantly with mobility, pain, and arthritis impact of the AIMS scale (r_s from 0.22 to 0.36) and with pain by visual analog scale ($r_s = -0.24$) and patient's global assessment ($r_s = 0.32$). Multiple regression analysis showed that patient's global assessment explained 41% (rating scale) and 10% (standard gamble) of total variance in baseline utilities. Also, 16% of the variance in change in rating scale utility values was explained by changes in patient's global assessment. In contrast, variance of changes in standard gamble utility values was not explained significantly by changes in other disease outcomes.

Conclusion. Rating scale utilities correlated more strongly with disease outcome measures than standard gamble utilities. Also, construct validity for the rating scale was better than for the standard gamble. In FM, utility measurement is sensitive to the method chosen to elicit patient priorities.

(*J Rheumatol 1995;***22:1536–43**)

Key Indexing Terms:
UTILITY RATING SCALE STANDARD GAMBLE VALIDITY
RHEUMATIC DISEASES FIBROMYALGIA OUTCOME

Fibromyalgia (FM) is a common rheumatic condition. Presenting symptoms are pain, stiffness, and fatigue[1]. It has been suggested that patients with FM may benefit from cardiovascular fitness training[2], whereas another report indicated that myobiofeedback training was successful[3]. In these studies conventional disease oriented endpoints such as pain, stiffness, number of tender points, sleep disturbance, fatigue, psychological, and global assessments were measured.

Recently, utility measurement has been introduced in the evaluation of interventions for arthritis patients[4]. Utility measures of health related quality of life are generic measures of the value or preference that patients attach to their overall health status, i.e., patients have to integrate all positive and negative effects of their disease and its treatment into one single value. In contrast, in disease specific instruments the benefits and risks are measured separately. Therefore, one has no information on the relative weights patients assign to therapeutic improvements and disadvantages such as side effects[5].

In a randomized controlled trial we evaluated the therapeutic effect of low impact fitness training and biofeedback training on patients with FM[6]. During this trial we also elicited utility values. Here we report on aspects of validity of the

From the Department of Internal Medicine, Division of Rheumatology, and the Department of Health Economics, University of Limburg, Maastricht; Institute for Medical Technology Assessment, Erasmus University Rotterdam, The Netherlands; and the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada.

Supported by Het Nationaal Reumafonds of The Netherlands.

C.H. Bakker, PhD, Research Fellow, Department of Internal Medicine, Division of Rheumatology; M.P.M.H. Rutten, PhD, Lecturer, Department of Health Economics; M.H.S. van Santen-Hoeufft, MD, Rheumatologist, Department of Internal Medicine, Division of Rheumatology; P.H. Bolwijn, MSc, Research Fellow, Department of Internal Medicine, Division of Rheumatology, University of Limburg; E.K.A. van Doorslaer, PhD, Health Economist, Institute for Medical Technology Assessment, Erasmus University Rotterdam; K. Bennett, MSc, Associate Professor, McMaster University; S. van der Linden, MD, PhD, Professor of Rheumatology, University of Limburg.

Address reprint requests to Ms C.H. Bakker, Department of Internal Medicine/Division of Rheumatology, University Hospital Maastricht, PO Box 5800, NL 6202 AZ Maastricht, The Netherlands.

Submitted September 7, 1994 revision accepted February 17, 1995.

utility measurement. We address in particular reliability, construct validity, and sensitivity to change relative to improvements in other outcomes.

MATERIALS AND METHODS

Study population. Patients with FM from the outpatient department of rheumatology of Maastricht University Hospital, who had been referred between January 1988 and December 1989, were asked to participate in the study. Altogether, 103 of 174 (59%) patients gave informed consent, of whom 86 met the eligibility criteria (female sex, age 18–60 yrs, criteria of Wolfe, *et al*[1]). Patients were excluded if they had high depression on 6 scales of the symptom checklist (SCL-90)[7,8]. The study was restricted to female patients only for practical reasons. Patients receiving low impact fitness training were allowed to use the sauna and swimming pool once a week.

Study design. After baseline assessment, patients were randomized into 3 groups. One group received low impact fitness training, the 2nd group had biofeedback training, and the last group were controls. Patients in the fitness group performed supervised aerobic and stretching exercises for 60 min twice weekly for 6 mo. Patients in the biofeedback group individually received 20 min relaxation training twice a week for 2 mo[9]. After completing the supervised biofeedback training, patients were encouraged to continue relaxation exercises at home, twice a day for at least 4 more mo. Patients were asked to keep daily records of their exercises. All patients were allowed to continue the treatment they already received before the study.

Utility measures. To elicit utility values the Maastricht Utility Measurement Questionnaire (MUMQ)[10] was administered at baseline and after 6 mo followup by 2 trained interviewers. On both occasions each patient was assessed by the same interviewer, who was blinded to the intervention. The MUMQ is a Dutch translation and adapted version of the McMaster Utility Measurement Questionnaire[11]. Briefly, health is defined by 6 dimensions: activities of daily living, self-care functions, emotions, leisure activities, pain, and side effects of treatment. Each dimension consists of 5 levels of severity: level 1 reflecting the best situation and level 5 the worst (Table 1). Marker states were created as follows: perfect health was the combination of the first levels of all 6 dimensions; a severe case of FM (marker

Table 1. *The 6 dimensions of health and their levels*

I. General daily activities and mobility
Think of limitations caused by tiredness, tightness of the chest or pain while working, housework, shopping, walking, climbing stairs, using public transport, driving a car, cycling, etc.
(1) able to perform all daily activities and duties at a normal level of mobility
(2) able to perform daily activities, but with some difficulties
(3) limited in the performance of daily activities
(4) limited considerably in the performance of daily activities
(5) unable or hardly able to perform daily activities

II. Personal care
Think of eating, washing, taking a shower or a bath, going to the toilet, etc.
(1) completely able to perform all self-care activities
(2) now and then having difficulty in the performance of self-care activities
(3) having difficulty in the performance of self-care activities
(4) considerable difficulty in performing self-care activities
(5) help needed for all self-care activities

III. Anxieties, frustrations and worries related to the course of the disease
(1) no anxieties, no worries, not concerned about the course of the disease
(2) normally no anxieties, sometimes concerned about the course of the disease
(3) depressed because of the inability to function normally
(4) often anxious, often concerned about the course of the disease
(5) depressive, unhappy, and frustrated

IV. Leisure activities
Think of going out, practising sports, hobbies, etc.
(1) able to participate in all leisure activities without difficulty
(2) able to participate in all leisure activities but with some difficulty
(3) ability to participate in leisure activities is limited
(4) no longer able to participate in any leisure activity that requires a certain degree of physical effort or mobility
(5) not able to participate in any leisure activity

V. Pain
(1) no pain
(2) occasional pain
(3) often mild to moderate pain
(4) often severe pain
(5) continuous severe pain

VI. Side effects of treatment
E.g., nausea, vomiting and/or diarrhea, gastrointestinal upset, rash, mouth ulcers.
(1) no side effects
(2) occasional mild side effects
(3) occasional moderate - severe side effects
(4) often moderate - severe side effects
(5) severe side effects

state of severe disease) was described as the combination of all 5th levels of the 6 dimensions. A mild FM marker state was described as the combination of level 1 of dimension 1, level 1 of dimension 2, level 2 of dimension 3, level 3 of dimension 4, level 2 of dimension 5, and level 2 of dimension 6[10].

In the interview the patients were asked to define their own health status by indicating their actual personal levels for each dimension. Then patients were asked to value the provided marker states of disease and their own health status, using both the rating scale and standard gamble method. The rating scale is a numerical scale that looks like a thermometer, with "perfect health" equal to 100 at the top and the "marker state of severe disease" equal to 0 at the bottom. The standard gamble is performed with a probability wheel as a prop[12]. The standard gamble is directly based on the Von Neumann-Morgenstern utility theory and is the original method of measuring utilities[13]. In the standard gamble method, health states are valued under the assumption of risk, as opposed to the rating scale method, where risk is not included in the measurement process.

Usually, patients in the standard gamble utility measurements are asked to value their own state of health in comparison to perfect health (valued as 1) and death (valued as 0) (Figure 1). However, in rheumatic diseases, direct confrontation with the risk of dying may be inappropriate. Therefore, a 2 step utility assessment was performed: patients were first asked to value their own health status in comparison to a gamble with probability (p) of gaining perfect health, and a probability (1 − p) of attaining the marker state of severe disease (Figure 2). Next, they were asked to value the marker state of severe disease in comparison to a gamble with probability (p) of gaining perfect health, and a probability (1−p) of dying (Figure 3). Probability (p) is systematically varied until the patient is indifferent between the 2 alternatives. In our study (p) was varied with steps of 10% [(p)/(1−p): 100/0, 10/90, 90/10, 20/80, 80/20, 30/70, etc.]. When indifference is reached, utilities for the health states in alternative 2 are calculated by:

$$U = pU_{best} + (1-p)U_{worst}$$

where U_{best} is the utility of the best outcome of the gamble (perfect health, and its utility is equal to 1 by definition), and U_{worst} is the utility of the worst outcome of the gamble (severe marker state in step 1 and death in step 2; the utility of death is 0 by definition). The 2nd step is measured between perfect health and death and therefore directly provides a utility value for the severe marker state. The first step provides a standard gamble score, which has to be converted into a utility value for the patient's own health status, using the utility of the severe marker state[14]. Negative utility values for the severe marker state, indicating that this state was considered worse than death, were recoded to zero in the analysis. It is assumed that

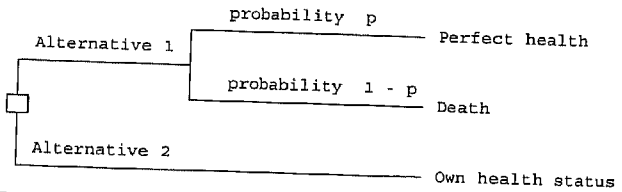

Fig. 1. Standard gamble: Value your own state of health in comparison to perfect health and death.

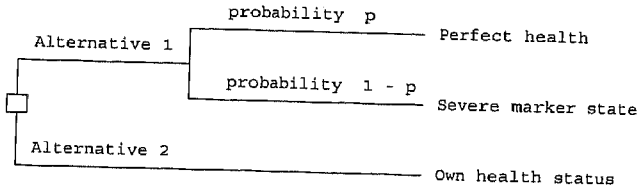

Fig. 2. Two-step standard gamble: First step: Value your own state of health in comparison to perfect health and the marker state of severe disease.

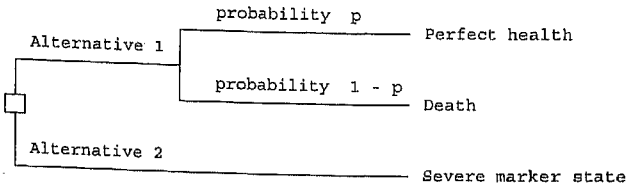

Fig. 3. Two-step standard gamble: Second step: Value the marker state of severe disease in comparison to perfect health and death.

patients with better health accept less risk in order to improve than the more severely affected patients[15].

Six other health status outcome measures were applied at baseline and followup: global health (on a 0–10 numerical rating scale, with 0 equal to very bad health and 10 equal to very good health); a standardized Dutch version of the Sickness Impact Profile (SIP) questionnaire[16,17]; Dutch Arthritis Impact Measurement Scale (Dutch-AIMS)[18]; pain [on a 10 cm visual analog scale (VAS) with 0 equal to no pain and 10 equal to most severe pain imaginable]; duration of morning stiffness (min); and modified Health Assessment Questionnaire (mHAQ)[19]. Note that global health was measured in 2 ways: on the arthritis impact dimension of the AIMS [global health (AIMS)] and on a 0–10 numerical rating scale [global health (NRS)].

Analysis and statistical methods. Reliability was tested in all patients by assessing the stability of marker states after 6 mo.

Construct validity of the MUMQ was tested by Spearman correlation coefficients between baseline utility values and baseline scores for global health, SIP, AIMS, mHAQ, pain, and stiffness. Improvements in utility values were expected to be associated with improvements in these variables. Therefore, p values were tested unilaterally (1-sidedly) at an α level of 0.05. Construct validity indicates whether results do agree with expected results based on *a priori* assumptions of the investigator[20]. In addition, multiple regression analyses (stepwise forward) were performed for these 6 clinical measures and age, disease duration, marital status, and education as independent variables and the rating scale or standard gamble utilities as dependent variables. Independent variables with skewed distributions were analyzed as ln(var+1), the natural logarithm of one plus the variable[21,22].

Discriminant validity or sensitivity to change indicates whether a measure can detect important clinical changes in health status over time[20]. Discriminant validity was tested in 2 ways. First, we calculated Spearman correlation coefficients (1-sided testing) between the changes in utility values and the changes in other outcome measures. Second, we performed multiple regression analyses (stepwise forward) with changes in rating scale or standard gamble utilities as dependent variables, and treatment and changes in the other health status outcomes as independent variables. Both dependent and independent variables were transformed by ln(var+constant), the natural logarithm of a constant just below the minimum of (change in) the variable plus (change in) the variable.

Sensitivity to change of the rating scale and standard gamble method was also assessed by calculation of the efficiency (E) (the mean change of the measure divided by the standard deviation of change), as suggested by Anderson and Chernoff[23].

Within each treatment group mean changes in utility were tested by Wilcoxon signed rank test.

RESULTS

At baseline all 86 eligible patients with FM completed the MUMQ. One patient withdrew before randomization and 12 patients dropped out during the 6 mo followup (6 in the fitness group, 5 in the biofeedback group; 1 from the control group) for the following reasons: illness of husband (2), too busy at job (2), hospitalization (1), no further interest (4), trial too stressful (2), biofeedback makes no sense (1). The

remaining 73 patients (fitness group 29; biofeedback group 26; controls 18) are included in this report. Demographic and clinical characteristics are shown in Table 2. Patients in the fitness group were significantly older than the controls (44.9 compared to 40.1 yrs) ($p = 0.05$). There were no other statistically significant differences in baseline characteristics and utilities between the 3 intervention groups.

Four patients gave inconsistent answers at baseline. These answers were excluded from analysis[10]. No interviews were broken off. The mean duration of the utility measurement at baseline was 10.8 min (SD, 2.9) for the rating scale and 12.5 min (SD, 3.8) for the standard gamble; at the 6 mo followup it decreased to 9.4 min (SD, 7.3) for the rating scale and 11.5 min (SD, 4.5) for the standard gamble method.

Mean values for utilities and other outcomes are shown in Table 3. Patients' utilities improved significantly by rating scale only (Wilcoxon signed rank test: $p = 0.008$). These utilities showed a significant improvement over time for those who had low impact fitness training (mean improvement 11; $p = 0.007$), but were insignificant among patients who had biofeedback training (mean improvement 3; $p = 0.3$), and stayed about the same in controls (mean improvement 0.1; $p = 0.99$). Differences between the 3 groups did not differ significantly (Kruskal-Wallis test; $p = 0.32$). Standard gamble utilities, however, did not change significantly in either the fitness group (mean change 0.06; $p = 0.2$), among controls (mean change 0.01; $p = 0.9$), or in the biofeedback group (mean change -0.01; $p = 0.8$).

Table 2. *Baseline characteristics for the 73 patients with FM*

Age (yrs)	
Mean (SD)	43.3 (8.3)
Duration of complaints (yrs)	
Mean (SD)	11.8 (9.8)
Married (%)	81
Employed (%)	27
Educational level	
High (%)*	37
Low (%)**	63

* Including secondary vocational training and university.
** Including lower vocational training.

Reliability. Test-retest reliability was assessed by the patient's valuation of the marker states at the 6 mo followup visit compared to baseline values. Utilities of marker states should not change over time[24]. However, on the rating scale the mean of the marker state values for mild disease increased significantly (mean change 4.0; Wilcoxon signed rank test: $p = 0.03$). These utilities improved significantly in the

Table 3. *Utilities and other outcomes at baseline and changes at followup among 73 patients with FM*

	Baseline		Change*	
	Mean	SD	Mean	SD
Utilities				
Patient's valuation of own state of health:				
Rating scale utility	55	18	5.5	20.3
Standard gamble utility	0.81	0.20	0.02	0.25
Patient's valuation of marker states of disease:				
Rating scale mild marker	73	13	4.0	13.4
Standard gamble mild marker	0.83	0.21	0.02	0.25
Standard gamble severe marker	0.42	0.38	−0.11	0.35
Outcome measures				
AIMS dimensions				
Mobility	0.5	1.2	0.22	1.42
Physical activity	5.5	2.0	0.03	2.33
Dexterity	3.2	2.7	−0.08	2.56
Social role	0.7	0.8	0.23	0.65
Social activities	4.3	1.7	−0.51	1.16
Activities of daily living	0.3	0.9	0.05	0.72
Pain	7.1	1.5	−0.30	1.52
Depression	3.7	1.6	−0.29	1.55
Anxiety	5.2	1.7	−0.37	1.45
Health perception	4.1	1.9	−0.26	1.60
Arthritis impact or global health	5.4	2.1	0.15	2.38
SIP	14	8.6	−1.36	6.55
mHAQ	0.47	0.37	0.07	0.37
Pain (VAS)**	5.9	1.9	−2.75	17.3
Stiffness	65	57	12.3	46.9
Patient's global health (NRS)***	5.4	1.4	0.85	1.76

* Followup − baseline.
** Visual analogue scale.
*** Numerical rating scale.

patients receiving low impact fitness training (mean change 6.3; p = 0.04), but were insignificant among patients receiving biofeedback training (mean change 1.2; p = 0.6) and among controls (mean change 4.4; p = 0.3). Differences between the 3 groups did not differ significantly (Kruskal-Wallis test: p = 0.4). The mean standard gamble utility for the marker state of mild disease improved insignificantly (mean change 0.02; p = 0.6), but the mean standard gamble utility for the marker state of severe disease deteriorated significantly (mean change −0.11; p = 0.01) (Table 3). In the group receiving low impact fitness training the standard gamble utility of the marker state of mild and severe disease changed insignificantly (mean change 0.05, p = 0.4 and −0.12, p = 0.2, respectively). In the biofeedback group the changes were −0.3 (p = 0.2) and −0.11 (p = 0.10), respectively; in the controls 0.05 (p = 0.4) and −0.10 (p = 0.2), respectively. Differences between the 3 groups did not differ significantly (Kruskal-Wallis test: p = 0.07 and p = 0.7, respectively). It should be noted that on the rating scale the bottom endpoint is the severe marker state, and therefore the test–retest reliability of this state could only be tested by the standard gamble method.

Comparison of rating scale and standard gamble utilities. Utilities did not correlate significantly with age, duration of disease, marital status, or education. Utilities obtained by rating scale did not correlate significantly (r = 0.14, p >0.05) with utilities obtained by standard gamble. The same was found for change scores (r = 0.19, p >0.05) (Table 4).

Construct validity of utility assessment. Spearman correlation. Rating scale scores correlated better with scores on AIMS, SIP, mHAQ, pain, stiffness, and global health than standard gamble scores (Table 4). Rating scale utilities correlated significantly with the arthritis impact, physical activity, social role, pain, depression, and health perception dimensions of the AIMS and also with SIP, mHAQ, global health (NRS), and pain (VAS) (Table 4). Standard gamble utility values correlated significantly with the arthritis impact, mobility, and pain dimensions of the AIMS and with the global health (NRS) and pain (VAS) scales only (Table 4). In interpreting the results, it should be noted that the significance level has not been adjusted for the number of comparisons made.

Multiple regression analysis. Multiple regression analysis was performed to determine which set of variables could best predict the rating scale and standard gamble utilities. For rating scale utilities, patient's global health (AIMS) explained 41% of the variance and the physical activity dimension (AIMS) another 11%. Variance in standard gamble utilities was explained significantly only by patient's global health (AIMS) for 10% (Table 5).

Sensitivity to change of utility assessment. Spearman correlation. Changes in rating scale utilities correlated significantly with changes in 4 dimensions of the AIMS (pain, depression, anxiety, arthritis impact), and with changes in SIP, pain (VAS), and patient's global health (NRS) (Table 4). Note

Table 4. *Spearman correlation coefficients between baseline utilities and health status outcomes at baseline and after followup for 73 patients with FM*

	Rating Scale Utility		Standard Gamble Utility	
	Baseline	Change	Baseline	Change
Age	0.08	–	0.20	–
Duration of disease	−0.01	–	−0.06	–
Marital status	0.01	–	−0.01	–
Education	0.21	–	0.001	–
Rating scale	–	–	0.14	0.19
AIMS dimensions				
Mobility	−0.39***	−0.17	−0.22*	0.03†
Physical activity	−0.52***	−0.18	−0.16	−0.13
Dexterity	−0.17	0.06†	−0.21	0.07†
Social role	−0.28*	0.20†	−0.15	−0.04
Social activities	−0.23*	−0.003	−0.03	−0.07
Activities of daily living	−0.26*	−0.19	−0.08	0.09†
Pain	−0.38***	−0.33**	−0.22*	−0.09
Depression	−0.30*	−0.27*	−0.20	−0.24*
Anxiety	−0.14	−0.29*	−0.16	−0.09
Health perception	−0.29*	−0.07	−0.19	0.07†
Arthritis impact or global health	0.62***	0.41***	0.36***	0.23*
SIP	−0.43***	−0.23*	−0.08	0.06†
mHAQ	−0.28*	0.09†	−0.16	0.003†
Pain (VAS)	−0.47***	−0.25*	−0.24*	0.01†
Stiffness	−0.19	−0.15	−0.09	−0.05
Patient's global health (NRS)	0.53***	0.41***	0.32**	0.24*

* p < 0.05. ** p < 0.01. *** p < 0.001. † Unexpected direction.

Table 5. *Stepwise forward regression analyses with rating scale or standard gamble utility as dependent variable and patient characteristics and other baseline assessments as independent variables*

Step	Variable Entered	Partial R^2	F	p
Dependent Variable: Rating Scale Utility				
1	Global health (AIMS)	0.41	33.44	0.0001
2	Physical activity (AIMS)	0.11	11.11	0.002
Dependent Variable: Standard Gamble Utility				
1	Global health (AIMS)	0.10	5.88	0.02

that Table 4 also indicates which correlations occurred in the "wrong" (paradoxical or unexpected) direction. For example, one would not *a priori* expect a decrease in social functioning to be associated with higher utility values. Changes in standard gamble utilities correlated significantly with changes in depression and arthritis impact dimensions of the AIMS, and with changes in global health (NRS). Again, the significance level has not been adjusted for the number of tests performed.

Multiple regression analysis. Multiple regression analysis with changes in rating scale utilities as dependent variable and changes in the other outcomes and treatment (group) as independent variables showed that 16% of total variance could be explained by changes in a patient's global health (AIMS) (Table 6). Changes in standard gamble utilities could not be explained significantly by changes in any of these variables.

Efficiency. For all patients the efficiency of the rating scale and standard gamble method were 0.27 and 0.08, respectively, indicating the rating scale is more sensitive to change than the standard gamble. For the group receiving low impact fitness training the efficiency of the rating scale and standard gamble were 0.54 and 0.26; for the biofeedback group they were 0.19 and 0.04; and for the controls 0.003 and 0.03, respectively.

DISCUSSION

We evaluated reliability, construct validity, and sensitivity to change of utility measurement by rating scale and standard gamble methods. These aspects of utility measurement differed considerably between both methods. Therefore, patient's utilities elicited by rating scale and standard gamble are not interchangeable.

The test-retest reliability of the MUMQ was assessed by the utilities for the marker states of disease that, of course, should not change over time[24]. The utilities of these marker states were in fact not stable (Table 3). Therefore, either the method itself has poor reliability, or the patient's perception and valuation of the marker states indeed change in the course of 6 mo. Note that both the valuation of the patient's own health status and the valuation of the mild marker state changed in the same direction, i.e., they were at a higher mean level after 6 mo (Table 3). Moreover, as the patient's utility improved, the distance between her own health status and the marker state of severe disease became larger. Patients emphasized this by valuing the marker state of severe disease lower at followup compared to baseline (Table 3). Therefore, possibly a change in the patient's perception of her own health status induces valuations for the reference (marker) states to change too. We suggest that this might be related to a patient's capabilities to adapt to disease, i.e., she might be better able to deal with her disease related limitations, and the perceptions of other patients with the same disease may change too. As marker states are presented as examples of mild and severe FM, the valuation of these reference states may change accordingly. However, in contrast, in patients with ankylosing spondylitis (AS) the median of the mild marker state on the rating scale and on the standard gamble did not change significantly after 9 mo[25]. Future research should clarify this issue.

Construct validity of utilities obtained by rating scale was supported by significant correlations with measures such as

Table 6. *Stepwise forward regression analyses with changes in rating scale or standard gamble utilities as dependent variables and changes in other assessments as independent variables*

Step	Variable Entered	Partial R^2	F	p
Dependent Variable: Change in Rating Scale Utility				
1	Global health (AIMS)	0.16	9.24	0.004
Dependent Variable: Change in Standard Gamble Utility				
1	mHAQ	0.07	3.61	0.06*
2	Physical activity (AIMS)	0.06	3.23	0.08
3	Social activity (AIMS)	0.06	3.16	0.08

* Unexpected direction.

global health, pain, SIP, AIMS, and mHAQ. Standard gamble utility values, however, correlated considerably less with these instruments. Patient's global health explained 41% and 10% of total variance of rating scale and standard gamble utilities, respectively. This suggests that standard gamble utilities reflect different aspects of health status than rating scale utilities, or have indeed considerably lower construct validity. For comparison, in patients with AS construct validity appeared to be higher for the rating scale than for the standard gamble[25]. Clearly, the 2 techniques are not interchangeable. It should be stressed that the standard gamble method incorporates a risk of getting a dispreferred outcome, whereas risk is not an issue in the rating scale procedure. The standard gamble method, therefore, addresses at least elements of uncertainty.

Our findings in patients with FM support the view that utilities obtained by rating scale more closely resemble global assessment. These findings are in accordance with our results in patients with AS[25]. The differences between rating scale and global assessment relate to the endpoints of these scales. Global assessments are measured in many (flexible) ways, i.e., with a variety of different scale endpoints. In contrast, rating scale utilities are measured in a standardized way, with perfect health and (usually) death as endpoints of the scale. Therefore, the methodological advantage of standardized rating scale utility measurement over nonstandardized global assessment is that utilities provide numerical values, which allow patient outcomes (of different diseases or resulting from various health care interventions) to be compared across patients and diseases.

An evaluation of discriminant validity showed that changes in rating scale utilities could be explained to a higher degree than changes in standard gamble utilities. These results are in accordance with our findings in patients with AS[25].

Based upon utility values obtained by rating scale and standard gamble, the therapeutic results of this randomized controlled trial were largely negative. The rating scale showed a trend to improvement in the fitness group; however, it was not statistically significant. This raises the following question: Is this due to ineffectiveness of the interventions offered to the patients with FM, or are the rating scale and standard gamble method not responsive enough to pick up relevant changes? We evaluated the sensitivity to change of the rating scale and standard gamble methods by means of the efficiency measure (E), which is independent of trial size[23]. Efficiency analysis showed that the efficiency of the rating scale was higher than the efficiency of the standard gamble, indicating the rating scale is more sensitive to change than the standard gamble. In the group receiving low impact fitness training the rating scale was moderately sensitive to change as were other outcomes, such as global health and pain assessment. Therefore, it seems more likely that our interventions lack demonstrable effectiveness in these patients. This could be due to the power of this study. The control group consisted of only 18 women, in comparison to 29 and 26 women in the low impact fitness and biofeedback groups. Another possibility is that the effect of the fitness training was in fact too low. A study comparing the effectiveness of high and low impact fitness training is underway[6].

In conclusion, reliability of utility measurement by the rating scale and standard gamble methods assessed by stability of marker states was rather poor in patients with FM. Correlations between utilities and other outcomes showed higher construct validity and sensitivity to change for the rating scale than for standard gamble utilities. Regression analysis indicated that rating scale values are strongly related to global assessment results. Rating scale utilities are better standardized than many global assessments, and they can be compared across patients, treatments, and diseases. Clearly, utility measurement is sensitive to the method chosen to elicit patient well being. This has important implications for decision making and health policy. In our view, more validity testing and standardization are needed before utility measurement can be applied on a larger scale in clinical practice or in health service research.

ACKNOWLEDGMENT

We thank all patients and Rob de Bie, Annemiek Fransen, Nico Groenman, Paul Kubben, Ton Lenssen, Hubert Schouten, Carlo Theunissen, Vicky Verstappen, and Frans Verstappen for their contributions in various stages of the study.

REFERENCES

1. Wolfe F, Smythe HA, Yunus MB, *et al:* The American College of Rheumatology 1990 criteria for the classification of fibromyalgia. Report of the multicenter criteria committee. *Arthritis Rheum 1990;*33:160-72.
2. McCain G, Bell D, Mai F, Halliday P: A controlled study of the effects of a supervised cardiovascular fitness training program on the manifestations of primary fibromyalgia. *Arthritis Rheum 1988;*31:1135-41.
3. Ferraccioli G, Ghirelli L, Scita F, *et al:* EMG-Biofeedback training in fibromyalgia syndrome. *J Rheumatol 1987;*14:820-5.
4. Bombardier C, Ware J, Russell J, Larson M, Chalmers A, Read L: Auranofin therapy and quality of life in patients with rheumatoid arthritis. *Am J Med 1986;*81:565-78.
5. Feeny D, Labelle R, Torrance GW: Integrating economic evaluations and qualify of life assessments. In: Spilker B, ed. *Quality of Life Assessments in Clinical Trials.* New York: Raven Press, 1990.
6. Santen van-Hoeufft M, Bolwijn P, Kleijnen J, *et al:* Is low or high impact fitness beneficial in fibromyalgia patients. Results of 2 randomized controlled trials. (in preparation).
7. Arrindell WA, Ettema JHM: Handleiding bij een multidimensionele psychopathologie-indicator. Lisse: Swets & Zeitlinger BV, 1986.
8. Derogatis LR: SCL-90: Administration, scoring and procedures manual-I for the r(evised) version. Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Research Unit, 1977.
9. Basmajian JV, ed: *Biofeedback, Principles and Practice for Clinicians.* Baltimore: Williams and Wilkins, 1979.
10. Bakker CH, Rutten M, van Doorslaer E, Bennett K, van der Linden S: Feasibility of utility assessment by rating scale and

standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol 1994;*21:269–74.
11. Bennett K, Torrance GW, Tugwell P: Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. *Controlled Clin Trials 1991;*(suppl)12:118–28.
12. Torrance GW: Social preferences for health states: An empirical evaluation of three measurement techniques. *Socioecon Planning Sci 1976;*10:129–36.
13. von Neumann J, Morgenstern O: *Theory of Games and Economic Behavior.* Princeton, NJ: Princeton University Press, 1944 (1st ed.), 1947 (2nd ed.).
14. Torrance GW: Measurement of health-state utilities for economic appraisal: A review. *J Health Econ 1986;*5:1–30.
15. Torrance GW: Utility approach to measuring health-related quality of life. *J Chron Dis 1987;*40:593–600.
16. Bergner M, Bobbitt RA, Carter WB, Gilson BS: The Sickness Impact Profile: Development and final revision of a health status measure. *Med Care 1981;*19:787–805.
17. Luttik A, Jacobs H, de Witte L: *Een Nederlandse versie van de Sickness Impact Profile,* 2nd ed. Vakgroep Huisartsgeneeskunde. Rijksuniversiteit Utrecht, 1987.
18. Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ: Evaluation of the Dutch Arthritis Impact Measurement Scales (Dutch-AIMS) in patients with rheumatoid arthritis. *Br J Rheumatol 1989;*28:487–91.
19. Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP: Assessment of patient satisfaction in activities of daily living using a modification of the Stanford health assessment questionnaire. *Arthritis Rheum 1983;*26:1346–53.
20. Tugwell P, Bombardier C: A methodological framework for developing and selecting endpoints in clinical trials. *J Rheumatol 1982;*9:758–62.
21. Fleiss JL: Reliability of measurement. In: *The Design and Analysis of Clinical Experiments.* New York: Wiley, 1986.
22. Pocock SJ: Further aspects of data analysis. In: *Clinical Trials. A Practical Approach.* New York: Wiley, 1983.
23. Anderson JJ, Chernoff MC: Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. *J Rheumatol 1993;*20:535–7.
24. Torrance GW, Feeny D: Utilities and quality-adjusted life years. *Int J Technol Assess Health Care 1989;*5:559–75.
25. Bakker CH, Rutten 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.