MEDICAL CARE
Volume 33, Number 9, pp 922–937

Methodological Issues of Patient Utility Measurement

Experience From Two Clinical Trials

MAUREEN P.M.H. RUTTEN-VAN MÖLKEN, MSC,* CARLA H. BAKKER, MSC,†
EDDY K.A. VAN DOORSLAER, PHD,‡ AND SJEF VAN DER LINDEN, MD, PHD†

This article explores various methodological issues of patient utility measurement in two randomized controlled clinical trials involving 85 patients with fibromyalgia and 144 with ankylosing spondylitis. In both trials one baseline and two follow-up measurements of the patients' preferences for their own health state and several hypothetical states were performed using the rating scale and the standard gamble methods.

It was confirmed that standard gamble scores are consistently higher than rating scale scores for both the experienced and the hypothetical states. The 3-month test-retest reliability for hypothetical states measured by intraclass correlation coefficients ranged from 0.24 to 0.33 for the rating scale and from 0.43 to 0.70 for the standard gamble. Although the reproducibility is not high, the group mean scores are fairly stable over time. Mean standard gamble scores tend to differ depending on the way the measurements are undertaken. Utilities elicited with chained gambles were significantly higher than utilities elicited with basic reference gambles. At the individual level some inconsistent responses occurred. However, more than 70% of these fell within the bounds of the measurement error, which ranged from 0.11 to 0.13 on the standard gamble (0–1 scale) and from 8 to 10 on the rating scale (0–100 scale). The large number of negative utilities for the severe hypothetical state, which was used as an anchor point in the chained gambles, and the magnitude of these negative utilities (down to -19) calls for intensified research efforts to handle these responses in utility calculations. Key words: utility; methodology; standard gamble; rating scale; trials; outcome. (Med Care 1995;33:922–937)

Decisions regarding the allocation of resources to health care interventions ideally should be based on the relative costs and benefits of the alternatives. This requires an assessment of the societal value of the outcomes (i.e., various health states) achieved by these interventions, which can be done by means of utility measurement. When utility measurement is applied in clinical trials, patients are asked to assign a single value to a health state on a scale ranging from 0 (usually death) to 1 (usually perfect health), by

*From the Department of Health Economics, University of Limburg, Maastricht, The Netherlands.

†From the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, The Netherlands.

‡From the Institute for Medical Technology Assessment, Erasmus University Rotterdam, The Netherlands.

This study was supported by grants from the Netherlands Health Research Promotion Program and the National Rheumatic Disease Foundation of the Netherlands.

Address correspondence to: Maureen P.M.H. Rutten-van Mölken, University of Limburg, Department of Health Economics, PO Box 616, 6200 MD Maastricht, The Netherlands.

balancing the positive treatment effects against the negative side effects.[1] A utility can be seen as an inclusive, generic quality of life measure, which reflects the net effect of treatment. It is designed to allow a broad comparison of the effects of health care interventions across patient populations. When such comparisons are made in the context of cost-utility analyses, utilities often are used as weights to compute "quality adjusted life years." Years of life are multiplied by utility weights for the health status during those years, thus adjusting these life years for their quality.

In general, two different approaches to utility measurement have been used in clinical trials.[2] In the first approach, patients are asked a number of questions about their functioning. Their answers are used to rate them on the various quality of life dimensions of a particular utility measurement instrument. Combining these dimensions results in descriptions of patients' overall health states, to which preference values are assigned. These values are obtained in a different population, usually the general population. The majority of utility analyses following this approach have used the health state preference values that were obtained by the original developers of the utility instruments.[3] The most commonly used prepackaged utility measurement instruments for this approach are Rosser's Disability Distress scale,[4] Kaplan's Quality of Wellbeing scale,[5] and Torrance's Health Utility Index.[6]

In contrast to the use of prepackaged systems, the second approach to utility measurement is to ask patients to assign a single preference value to their overall quality of life. This self valuation by patients has not been widely undertaken in the past, but appears to be increasingly used.[3] The Auranofin trial in rheumatoid arthritis patients is a well-known example of this approach.[7,8] It has yet to be established whether one of these two approaches or any of the available utility instruments is superior.

In the two studies in patients with fibromyalgia and ankylosing spondylitis, described in this article, we opted for the second approach. It seems very appealing to us to incorporate patients' preferences in evaluating a therapy, because only they know the true implications of a particular health state from firsthand experience, and their preferences reflect the relative desirability of different health states to those who should benefit from the services provided.[9]

Utilities can be elicited by a number of techniques of which the rating scale, standard gamble, and time trade off are the major ones. A comprehensive description of these techniques is given by Torrance.[10] In our study only the rating scale and the standard gamble were used: the first because it is a very simple technique in which a subject provides his preference values explicitly by placing health states on a scale with clearly defined endpoints (eg, best imaginable health state and worst imaginable health state); the second because it is the only technique that is well-founded on an economic theory, the Von Neumann and Morgenstern expected utility theory.[11] We did not consider using the time trade off technique because, on the one hand, it seems more difficult than the rating scale and, on the other hand, it is not founded on a particular theory. Furthermore, for practical reasons the number of techniques had to be limited. A fuller description of the rating scale and standard gamble techniques is given in the methods section.

A review by Froberg and Kane of measurement issues related to obtaining utilities, indicated there was a considerable lack of knowledge about the accuracy of utility measurement techniques.[12–15] This article attempts to make a contribution to overcoming parts of this deficiency by focusing on those methodological issues of patient utility measurement that became apparent in our studies. In the methods section we will provide a brief description of the patients and the studies in which the utility

measurements were incorporated, and the methodology of utility measurement that was used. The main body of this article contains four separate sections on methodological issues. In the first section, a comparison is made between rating scale and standard gamble utilities, whereas in the second section the reliability of both techniques is compared. The third section concerns the internal consistency of the standard gamble, and the fourth section addresses health states valued worse than death. Some issues, such as the observed differences between the rating scale and standard gamble methods,[16–19] inconsistent responses, and the differences between basic reference gambles and chained gambles[20,21] have been addressed before, but usually not in the context of clinical trials. Reliability data are scarcely reported in the existing literature and little has been reported empirically on the occurrence of negative utilities.

Methods

Patients and Studies

Utility measurements were performed in the context of two randomized controlled clinical trials in rheumatic patients. These trials and their results are described in more detail elsewhere.[22,23] Patients were recruited for both studies from outpatient clinics. In the first study, 85 women with fibromyalgia (criteria of Wolfe et al)[24] were divided randomly into a standardized fitness training program (n = 35), a biofeedback training program (n = 31), and a control group (n = 19). Their mean ± SD age was 44 ± 8 years; 82% were married, 27% employed, and 66% had a low educational level. At the 6-months follow-up 12 patients dropped out: 6 in the fitness group, 5 in the biofeedback group, and 1 in the control group, all for reasons unrelated to their disease.[23] The baseline characteristics of the dropouts did not differ significantly from those of the patients who completed the study.

In the second study, 144 patients (21% women) with ankylosing spondylitis (modified New York criteria, 1984)[25] were assigned randomly to receive either (1) self-administered unsupervised individual physical exercise at home (n = 68); or (2) weekly sessions of group physical therapy in addition to the same individual physical exercise at home (n = 76). Their mean ± SD age was 43 ± 10 years, and 67% were married, 72% were employed, and 35% had a low educational level. By the 9-months follow-up, 9 patients had withdrawn (8 in the experimental and 1 in the control group): 4 because of the inability to exercise individually, 4 owing to other diseases or pregnancy, and 1 who had moved. Their baseline characteristics did not differ significantly from those who did not withdraw.[19]

Because the focus of this article is on methodological issues, it is not our intention to address the effectiveness of the therapies.[22,23] The trial data are used only to illustrate some methodological issues.

Methods of Utility Measurement

In the fibromyalgia study, patients were seen for utility measurement at an outpatient clinic at baseline and at the 3- and 6-months follow-ups, whereas in the ankylosing spondylitis study utility measurements were scheduled at baseline, 3-, and 9-months follow-ups. The measurements were done by means of the Maastricht Utility Measurement Questionnaire, a translated and slightly adapted version of the McMaster Health Utility Index.[26,27] This instrument is administered by a trained interviewer and takes about 30 minutes to complete.

All patients were asked to provide utilities for three hypothetical reference health states describing mild, moderate, and severe impairment in quality of life as well as for their own health state. The description of each health state covers 6 dimensions: (1) activities of daily living, (2) self-care functions, (3) anxiety and depression, (4) leisure

activities, (5) pain and discomfort, and (6) side effects from treatment. Each patient described his or her own health state by ticking off one of the 5 functional levels within each dimension (1 being the best functional level and 5 being the worst functional level). For all health states, duration was specified as "the rest of your life."The reference states help patients to determine the position of their own health state on the spectrum of possibilities.[28] Because the reference states remain the same throughout the study, they also enable the calculation of the test-retest reliability of the utility measurement instrument at the repeated follow-ups.[28]

As part of the Maastricht Utility Measurement Questionnaire, utilities were elicited using both the rating scale and the standard gamble method. In the first part of the baseline and follow-up interviews, patients rank the reference health states and their own health state by preference on a rating scale with the endpoints of perfect health (100) and the severe reference state (0). They are asked to do this in such a way that the distances between the states represent the differences in their preferences. The baseline interview then continues with standard gamble questions 1, 2, and 3 as shown in Table 1. The first 80 ankylosing spondylitis patients coming to the 9-months follow-up were asked an additional fourth question in which they valued their own current health state using perfect health and death as outcomes of the gamble. At the 3- and 9-months follow-up, all patients with ankylosing spondylitis who found that the severe reference state was worse than death were asked a fifth question to assess the magnitude of the negative utility for the severe reference state. Throughout this article we will refer to the question numbers in Table 1.

The standard gamble sometimes is seen as the gold standard for utility measurement, because it is based directly on the axioms of the Von Neumann and Morgenstern expected utility theory.[11] This theory consists of a number of axioms for rational decision-making under risk. One of these axioms specifies the standard gamble approach to utility measurement. In each standard gamble question, a patient is offered a choice between certain continued life in the health state being evaluated (hi), and a gamble with chance P to gain the best outcome of the gamble (perfect health) and chance 1-P of attaining the worst outcome of the gamble. The health state being valued must be intermediate between the two outcomes of the gamble in terms of preference. Chance P is varied systematically until the patient is indifferent between continued life in state hi and taking the gamble. In our studies P was varied in steps of 10% (P/1-P: 100/0, 90/10, 10/90, 80/20, 20/80, etc). When the indifference point has been found, the utilities of the health states (U_{hi}) are calculated using the following expected utility equation:

TABLE 1. Standard Gamble Questions at Baseline (1, 2, and 3) and at Follow-up (1,2,3,4,5,6)[a]

Standard Gamble Question	Health State Being Valued	Outcomes of the Gamble	
		Best (P)	Worst (1-P)
1	Mild reference state	Perfect health	Severe reference state
2	Own health state	Perfect health	Severe reference state
3	Severe reference state	Perfect health	Death
4	Own health state	Perfect health	Death
5	Death	Perfect health	Severe reference state

[a]Question 4 was only put to 80 ankylosing spondylitis patients at 9 months follow-up; question 5 only at 3- and 9-months follow-up to ankylosing spondylitis patients who found the severe state worse than death.

$$U_{hi} = pU_x + (1 - p)U_z$$

where P is the indifference probability, U_x is the utility of the best outcome of the gamble and U_z is the utility of the worst outcome of the gamble.*

Standard gambles with perfect health and death as potential outcomes are called basic reference gambles. Because by definition the utility of perfect health is 1 and the utility of death is 0, the utility of the health state being valued in a basic reference gamble is equal to $P.1 + (1\text{-}P).0 = P$, the indifference probability. Generally, the more undesirable the health state being valued, the greater the willingness to take a risk to escape that health state, the lower the indifference probability P, and thus the lower the utility for that state. Thus, the standard gamble provides an implicit valuation of a health state relative to the two possible outcomes of the gamble.

The worst (or best) outcome of the gamble can be replaced by any other health state as long as it is worse (or better) than the health state being valued.[29] Such gambles are called chained gambles, because they have to be chained to a basic reference gamble that assesses the utility of that other health state. In our studies, the severe reference state was substituted for death, to avoid using death in a gamble that involved a chronically ill patient's own health state. Including death could upset them. Moreover, in the period covered by our studies, death was unlikely to be a relevant outcome in the rheumatic disease patient groups we studied. When the severe reference state is used as the worst outcome of the gamble (as in standard gamble question 1 and 2), the utility for the health state being valued can be calculated using the same equation as above, where U_z is the utility of the severe reference state, a utility that is measured in basic reference gamble 3.

In general, when the severe reference state was considered worse than death, a utility of 0 was assigned to that state, because the magnitude of the negative utilities was not known. Standard gamble question 5, which assesses the magnitude of a negative utility, was asked of patients with ankylosing spondylitis only at the 3- and 9-months follow-up. When indifference is reached in standard gamble question 5, the magnitude of the negative utility is calculated as $-P/(1 - P)$, where P is again the indifference probability.[10] Although calculated, these negative utilities were not used in the chained gambles.

To facilitate the patients' understanding of the standard gamble questions, a probability wheel was used as a visual aid.[18] This is an adjustable disk with two different colored sectors that reflect the probabilities of getting the two outcomes of the gamble. The outcomes of the gamble are described on cards that have the same color as the sectors. The size of the sectors is changed according to the change in probability.

Statistical Analysis

Results will be presented using means and associated standard errors. However, because the negative utilities of the severe reference state elicited in standard gamble question 6 included a number of extremely negative values, we present the 5% trimmed mean and the 5% trimmed standard deviation for these results. This means that the upper and lower 5% of all observations were excluded when calculating the mean and standard deviation, thus removing the influence of the outlier values that caused the distribution of negative utilities to be skewed heavily to the left.[30] Within-patient analyses by means of paired t tests were performed (1) to compare rating scale with standard gamble scores, (2) to compare

*The indifference probability is defined as the midpoint of the two probabilities of perfect health between which the preference shifts from the gamble to the sure state. For example, if a patient prefers a gamble with probabilities 90/10 to the sure health state, but prefers the sure health state to a gamble with probabilities 80/20, the indifference probability is 0.85.

chained gambles with the basic reference gamble, and (3) to test for differences between reference state scores over time. Pearson product moment correlations are reported as measures of association between rating scale and standard gamble scores. A logistic regression analysis was performed to test for differences between patients who did and those who did not give inconsistent responses. Intra-class correlation coefficients were calculated to examine test-retest reliability.[31] Reliability was assessed further in terms of the precision of an individual measurement. This precision, expressed as σ_e, is the standard deviation of the measurement error, also called the standard error of measurement.[32] It is calculated as the square root of the mean square error (MSE), which is given by an analysis of variance.[31] It also can be calculated as $\sigma\sqrt{(1-R)}$, where σ is the standard deviation of all measurements and R is the test-retest reliability coefficient.[32]

Results

1. Comparing Rating Scale and Standard Gamble Utilities

Although the standard gamble method sometimes is seen as the gold standard, the rating scale method is far more frequently used, probably because it is less time consuming and easier to apply. In our studies Pearson product moment correlations between rating scale and standard gamble scores for various health states were found to range between 0.31 and 0.48 ($P < 0.001$). However, highly significant correlations can coexist with systematic differences between the methods. As can be seen in Table 2, the mean utilities of the patients' own health states assessed via the standard gamble method were significantly higher than the utilities assessed via the rating scale method. This pattern, which also was found for the mild reference health state, is consistent with earlier findings.[16–19]

In ankylosing spondylitis, we found a difference of somewhat less than 5% between the methods, and in fibromyalgia we found a difference of more than 10%. Such differences might considerably affect the results of a cost-utility analysis and alter the conclusions drawn. Whether this happens depends on the sensitivity of the decisions to the observed range of variation. Some cost utility ratios may be very robust to the magnitude of the utility, whereas others may change as a result of only a very small change in utility. The variability of responses among patients is somewhat greater for the standard gamble method than for the rating scale method. Several phenomena might ex-

TABLE 2. Mean (SE) Rating Scale and Standard Gamble Values for the Patient's Own Health State

	N	Rating Scale[a] Endpoints: Perfect Health—Severe (SE)	Standard Gamble Outcomes: Perfect Health—Severe (SE)	P^b
Fibromyalgia				
Baseline	85	0.54 (0.020)	0.67 (0.028)	< 0.001
3 months	76	0.57 (0.023)	0.76 (0.023)	< 0.001
6 months	73	0.60 (0.022)	0.76 (0.025)	< 0.001
Ankylosing spondylitis				
Baseline	144	0.72 (0.013)	0.75 (0.018)	0.095
3 months	137	0.74 (0.012)	0.78 (0.016)	0.009
9 months	133	0.75 (0.013)	0.79 (0.015)	0.002

SE, standard error.
[a]Rating scale preferences were divided by 100.
[b]Paired Student's *t* test.

plain the differences between rating scale and standard gamble preferences. Three of them are discussed below.

Response Spreading. The first is called "response spreading" on the rating scale. This means that patients tend to distribute the health states over the entire scale, even if the true values were bunched at one end.[17,33] The mean baseline rating scale scores of 26.8, 54.1, and 72.6 assigned by the fibromyalgia patients to the moderate reference health state, their own health state, and the mild reference health state, respectively, may indicate a tendency to use the whole scale. In ankylosing spondylitis these baseline scores were 36.7, 80.0, and 76.5. Utilities are cardinal measures, reflecting not only the ranking of various health states relative to perfect health and death, but also the magnitude of the difference between these different health states.[26] If response spreading occurs, then the rating scale gives an indication of ordinal rankings and the intensity of the preferences, but it does not provide interval-scale utilities.

Risk Attitude. A second explanation for the significant difference between mean rating scale and mean standard gamble scores may be the patients' attitudes toward risk itself. Rating scale scores are measured under certainty and do not capture the respondent's attitude toward risk. In contrast, the standard gamble approach incorporates the respondent's risk attitude, which may be risk averse, risk neutral, or risk seeking. If subjects are not risk neutral, differences can be expected between rating scale and standard gamble values. If they are risk averse, their indifference probability increases, thus increasing the utility of the health state being valued. Kahneman and Tversky have shown that people generally acted as if they were risk averse when choices were framed in terms of potential gains and as risk seeking when choices were framed in terms of potential losses.[34] According to their "prospect theory" the displeasure of a loss is generally greater than the pleasure associated with an equivalent gain.[35] If the patients in our studies have focused more on the negative outcome of the gamble, risk-averse behavior may have been stimulated. As a result of this behavior, the standard gamble utilities would be biased upward compared with rating scale values.[20]

Standard gamble utilities also may be biased upward because people tend to overweight sure outcomes relative to outcomes that are highly probable. This is called the certainty effect, but it also is referred to as the Allais paradox.[20,34] Kahneman and Tversky's "prospect theory" assumes a decision weight function that over-weights small probabilities and under-weights moderate and high probabilities. If, in our studies, moderate and high probabilities were under-weighted, this might have contributed to the relative attractiveness of the sure outcome, even when the probability of gaining perfect health was rather high. Overweighting a small chance of ending up in the severe reference state might have reinforced the attractiveness of the sure outcome. Moreover, the fact that patients knew from experience that they had been able to adapt to their illness before may have diminished both the severity of the health state being valued and the relative value of the therapeutic pay-off from treatment gambles.[36]

Cognitive and Emotional Factors. Other explanations for the difference between rating scale and standard gamble values are all related to the previous explanations. Such explanations might include differences in cognitive processes such as recalling and taking account of past events, life goals, family circumstances, and the selection of reference points against which consequences are evaluated.[17]

Cognitive factors play an important role in Loomes and Sugden's alternative to expected utility theory, called "regret theory."[37,38] According to that theory, the value a person assigns to a health state depends not only on that health state but also on how that health

state compares with the health state the person might have had if he or she had made a different choice. If what is obtained is better than what might have been, feelings of rejoice may increase the utility; if what comes is worse than what might have been, regret may reduce the utility. In standard gambles where patients are asked explicitly to make a choice, feelings of regret and rejoice may be anticipated, whereas such feelings are absent in the choiceless rating scale valuation process. Subjects may shy away from the gamble choice in the standard gamble because of regret aversion (regret may occur if they "lose" the gamble and end up with the worst outcome). By means of experiments, Loomes and colleagues have shown that regret theory is able to explain why observed preference reversals may have occurred.[39]

2. Reliability of the Rating Scale and the Standard Gamble

Test-Retest Reliability. The reproducibility of the rating scale and standard gamble methods has been scarcely reported in the literature. O'Connor reported a Pearson product moment correlation of 0.77 and 0.80 for the one-week test-retest reliability of the rating scale and the standard gamble,[40] whereas Torrance reported product moment correlations of 0.49 and 0.53 for the one-year test-retest reliability of these methods.[18] In our studies reproducibility was assessed by calculating 3-month intra-class correlation coefficients (ICC) for the values assigned to the reference health states. The results are given in Table 3.

Although the reproducibility of the standard gamble is somewhat higher than that of the rating scale, the ICCs of the scores assigned to the reference states generally are not very high. This may point at difficulties in valuing abstract, hypothetical health states that have never been experienced. There are good reasons why it may be difficult to envision the well-being associated with a hypothetical health state. One is the inevitable gap between imagination and the actual experience of a health state. Individuals may overestimate or underestimate their ability to accommodate or to cope with adversity.[16]

Although the reproducibility is not high, Table 4 shows that—despite the occurrence of some slight but statistically significant changes in the preferences for the mild reference state in fibromyalgia and the severe reference state in ankylosing spondylitis—the mean scores are fairly stable over time. This stability may point at the usefulness of aggregated scores for group decision making.

A patient's true preference may change over 3 months and the preference at one time may not be representative of the patient's long-term preference.[41] This hypothesis is supported by the higher reliabilities that were found when the first 15 fibromyalgia patients from the control (no intervention) group to report for the 6-months follow-up were asked to return for a 4-week test-retest reliability

TABLE 3. Test-Retest Reliability: Three Months IntraClass Correlation Coefficients

	Rating Scale		Standard Gamble	
Reference States	Fibromyalgia	Ankylosing Spondylitis	Fibromyalgia	Ankylosing Spondylitis
Mild	0.33	0.26	0.43	0.50
Moderate[a]	0.24	0.29	—	—
Severe[b]	—	—	0.70	0.65

[a]The moderate reference health state was not valued by means of the standard gamble.
[b]The severe reference health state was not valued by means of the rating scale.

TABLE 4. Mean (SE) Utilities of the Reference Health States

	0 months (SE)	3 months (SE)	6/9 months[a] (SE)	p^b 0–3 (SE)	p^b 3–6/9 (SE)
Fibromyalgia					
Rating scale					
Moderate	27.3 (1.55)	29.1 (1.68)	30.4 (1.54)	0.382	0.388
Mild	72.6 (1.54)	79.7 (1.26)	76.8 (1.28)	0.000	0.035
Standard gamble					
Mild	0.83 (0.02)	0.85 (0.02)	0.85 (0.02)	0.366	0.851
Severe	0.42 (0.04)	0.35 (0.04)	0.31 (0.04)	0.083	0.095
Ankylosing Spondylitis					
Ratings scale					
Moderate	36.7 (1.14)	39.0 (1.28)	39.1 (1.37)	0.116	0.909
Mild	76.5 (0.83)	77.3 (0.70)	77.5 (0.90)	0.346	0.888
Standard gamble					
Mild	0.84 (0.02)	0.85 (0.01)	0.86 (0.01)	0.777	0.453
Severe	0.33 (0.03)	0.31 (0.03)	0.39 (0.03)	0.449	0.014

[a]The second follow-up measurement was scheduled at 6 months in fibromyalgia and at 9 months in ankylosing spondylitis.
[b]Paired Student's *t* test.

assessment.[42] In this assessment, rating scale ICCs of 0.56 and 0.67 were found for the patient's own health state and the mild reference state, respectively. The ICC of the standard gamble utilities was 0.66 for the patient's own health state, 0.74 for the mild, and 0.94 for the severe reference state. The generally higher ICCs for the severe reference state are partly owing to the fact that negative utilities for the severe reference state were recoded to zero.

Standard Error of Measurement. Another way of looking at the reliability of our utility measurements is to look at the measurement error.[18,31,32] Each single patient's preference measurement contains some measurement error, which causes part of the variance among the scores. The standard error of measurement, which is the standard deviation of the measurement error, was calculated to be 0.13 for the standard gamble (scale from 0–1) in fibromyalgia and 0.11 for the standard gamble in ankylosing spondylitis. For the rating scale method (scale from 0–100), the standard deviation of the measurement error was 10 in fibromyalgia and 8 in ankylosing spondylitis. These figures suggest that both methods contain considerable measurement error, implying relative instability of an individual patient's preferences. This supports the notion that utilities may be less useful for individual decision-making than for group decision-making, as was indicated by the relative stability of preferences over time shown in Table 4.

3. Internal Consistency of the Standard Gamble

Basic Reference Gambles Versus Chained Gambles. According to the axioms of expected utility theory, the outcomes of the gamble should not influence a patient's utility for a particular health state. The patients are supposed to adjust their indifference probability to allow for alterations in the gamble outcomes. At the 9-months follow-up, 80 consecutive ankylosing spondylitis patients were asked to value their own health state both in comparison with perfect health and death and in comparison with

perfect health and the severe reference state. The latter was the first of a chained pair of gambles. According to expected utility theory, there should be no difference in utilities elicited by a basic reference gamble or a chained gamble. However, the mean utility value of 0.83 in the case of the basic reference gamble was statistically significantly lower than the mean utility of 0.87 when the severe reference state was used as a gamble outcome in a chained pair of gambles (paired t test; $P = 0.018$). This is in accordance with earlier findings.[21,43] A majority of the patients (75%) assigned lower utilities in the basic reference gamble than in the chained gambles. Table 5 gives an example of this phenomenon for one patient.

The fact that the chained gambles resulted in higher utilities than the basic reference gamble does not imply that patients took a smaller risk when death was replaced by the severe reference state. On the contrary, generally, patients took a greater risk in the gamble where the severe reference state was the worst outcome than in the gamble where death was the worst outcome. The eventually higher utilities in the chained gamble resulted from a combination of relatively small differences between the indifference points in the chained and basic reference gamble and the relatively high utilities assigned to the severe reference state (see example in Table 5). This finding is in contrast with the findings of Llewellyn-Thomas et al, who found that raters were prepared to take a greater risk in gambles when death was the worst outcome.[21] This difference may be explained by the fact that Llewellyn-Thomas et al elicited utilities from cancer patients, to whom death usually is a real risk, whereas death to fibromyalgia or ankylosing spondylitis patients is not or is only a remote issue.

Because the standard gamble method seems to be susceptible to the characteristics of the worst outcome of the gamble, Hellinger and Llewellyn-Thomas concluded that this method is internally inconsistent.[20,21] However, when a change of focus or a change of reference point occurs as a result of a change in gamble outcomes, the preference shifts are not necessarily illogical.

Inconsistent Responses. Respondents are expected to provide preferences that are consistent with the natural underlying order of our health state descriptions. In other words, dominance violations should not occur. For the standard gamble as applied in our studies, dominance implies that when a patient's own health state is compared with the mild reference state and all 6 dimensions indicate a better (worse) or equal functional level, the utility the patient assigns to his or her own health state should be higher (lower) or equal to the utility of the mild reference state. However, seven fibromyalgia patients—each only once—did not provide preferences in accordance with this expectation. In ankylosing spondylitis, dominance was violated by 17 patients on the standard gamble. On the total number of questions this number of dominance violations is

TABLE 5. Difference Between a Chained and a Basis Reference Gamble

SG Question	Health State Being Valued	Gamble Outcomes[a]	P^b	U_{hi}^c
2 (chained)	Own	Perfect health (1) severe ref. state (0.55)	0.75	0.89
5 (basic)	Own	Perfect health (1) death (0)	0.85	0.85

SG, standard gamble.
[a]Utilities for outcomes in parentheses.
[b]Indifference probability for gamble.
[c]Utility for a patient's own health state.

rather low. Moreover, in 17 of the 24 (71%) inconsistent answers, the difference between the utilities of the two compared health states was smaller than the standard deviation of the measurement error reported in the previous section. Thus, most of the inconsistent responses fell within the bounds of the measurement error of 0.13 in fibromyalgia and 0.11 in ankylosing spondylitis.

It would also be expected that, if death is regarded as worse (better) than or equal to the severe reference state, the indifference probability of the patient's own health state when compared with perfect health and death should be higher (lower) than or equal to its indifference probability when compared with perfect health and the severe reference state. This was checked for the 80 ankylosing spondylitis patients who were asked the additional fifth standard gamble question. Twelve of them (15%) gave an inconsistent response. Almost all of these inconsistent responses fell within the bounds of the measurement error.

Extreme Risk Averse Behavior: Assigning a Utility of 0.95 to Each Health State. Eight of 85 (9%) patients with fibromyalgia reached the same indifference probability of 0.95 for all health states being valued in the first three standard gamble questions on at least one of the measurement times. Eleven of 144 (8%) patients with ankylosing spondylitis assigned a value of 0.95 to all three health states at least once. One fibromyalgia patient and one ankylosing spondylitis patient did this consistently at baseline and at each follow-up measurement. When controlling for other patient characteristics, fibromyalgia patients in which this phenomenon was found tended to be somewhat older than the patients in which this was not found. However, this difference was not statistically significant (logistic regression; Wald statistic = 2.975; P = 0.08).

Apparently these patients were never willing to take a larger than 10% risk of getting the worst outcome, irrespective of the severity of illness in the health state being valued. This behavior can be explained by a general aversion to gambling.[16] Such a reluctance to comply with the standard gamble questions reflects either reluctance to face the reality of the decision problem, reluctance to bear decision-making responsibility or inability to grasp hypothetical, unrepresentative experiments presented in a necessarily simple and abstract way. It may have been too difficult for these patients to imagine well-being associated with a hypothetical health state, or they may have underestimated their ability to cope with the severe reference state.

It is also possible that the probability steps of 10%, from 100/0 to 90/10, etc, have been too large. Changes of 5% (100/0, 95/5, 90/10, etc) or even 1% (100/0, 99/1, 98/2, etc) could have produced a difference in utilities between the different health states. However, findings by Kahneman and Tversky suggest that probabilities of less than 0.1 and greater than 0.9 are difficult for people to handle.[34]

4. A Health State Worse Than Death

In standard gamble question 3, the severe reference health state is valued against perfect health and death. Some 41 of the 85 fibromyalgia patients (48%) and 78 of 144 (54%) ankylosing spondylitis patients indicated at least once that the severe reference state was worse than death. This means that they were prepared to accept a 100% risk of dying to avoid this state. Eighteen fibromyalgia patients and 20 ankylosing spondylitis patients preferred death to the severe reference state at all 3 measurements. When asked explicitly, all these patients confirmed that they would rather die than live in the severe reference state.

As mentioned before, when patients indicated that the severe reference state was worse than death, a utility of zero was assigned to this state. To actually measure the magnitude of a negative utility, Torrance has

suggested a slight modification of the usual standard gamble question.[10] In this modified question, patients are offered a choice between a sure death or a gamble with chance *P* of perfect health and chance 1-*P* of living irreversibly in the severe reference state. This question is presented by asking patients to imagine that they suffer a rapidly progressing terminal disease, which—if left unattended—will lead to death. However, if treated there is a chance of gaining perfect health or of becoming like the severe reference health state for the remainder of their life. The utility of the severe reference state is calculated, from the indifference probability, as $-P/(1 - P)$.

At the 3- and 9-months follow-ups all the ankylosing spondylitis patients who indicated that the severe reference state seemed worse than death (51 at 3 months and 36 at 9 months) were asked this standard gamble question. At the 3-months follow-up, the 5% trimmed mean utility of the severe reference state was -0.16 with a 5% trimmed standard deviation of 0.19; at the 9-months follow-up the 5% trimmed mean utility of the severe reference state was -0.18 with a 5% trimmed standard deviation of 0.34. The mean and standard deviation were trimmed 5% because the distribution of the answers to standard gamble question 3 was skewed heavily to the left. The standard gamble to assess the magnitude of a negative utility can result in utilities smaller than -1, thus causing the upper end of the utility scale (from 0 to 1) to be shorter than the lower end of the scale (from 0 to -19 or less, depending on the size of the probability increments in the measurement instrument). At the 3-months follow-up, 1 patient assigned a utility of -3.00, and at the 9-months follow-up 3 patients assigned a utility lower than -1 (-1.86, -3.00, and -19.00) to the severe reference state.

The finding that for so many patients death was not the worst imaginable outcome has led some authors to conclude that death is not the logical zero point for a utility scale.[44] To some extent, the descriptive validity of negative utilities may be questioned, because probably few people would act in accordance with their statement that they would be better off dead. However, we certainly recognize the existence of health states valued worse than death. At the same time, we recognize the limitations of the standard gamble method in handling these states. Particularly, the occurrence of utilities below -1 is problematic. Thus, there is a need to intensify the search for methods to calculate and handle these negative utilities. Given the current methodological problems, we think that for chained gambles it is better to use an anchor point that is not considered worse than death, to avoid unnecessary complication of the chained gambles. Using death as the worst outcome of the gamble may increase the comparability of utilities measured in different patient groups and in different studies.

Conclusions and Discussion

Table 6 provides an overview of the most important findings, possible explanations, and preliminary implications reported in the previous sections.

In the introduction of this article, two different approaches to utility measurement in clinical trial settings were distinguished. One can either use prepackaged systems (eg, the Quality of Wellbeing scale, Rosser's Disability Distress scale, or the recently developed EuroQol[45]) or directly elicit preferences from patients. When using prepackaged systems, one needs to acquaint oneself with the method on which the underlying utilities were based. As has been reported before and reconfirmed in our studies, standard gamble preference scores for a particular health state are significantly higher than rating scale preference scores for that same health state. Hence, because the Quality of Wellbeing scale, the EuroQol, and Rosser's Disability Distress scale are based on rating scales, these prepackaged instru-

TABLE 6. Summary Table

Most Important Findings	Possible Explanations	Implications
1. Comparing RS and SG utilities		
SG utilities significantly higher than RS utilities	Response spreading on the RS Risk attitude incorporated in the SG Cognitive and emotional factors (eg, regret theory)	Prepackaged utility measurement instruments based on the SG are bound to produce higher utilities Costs per QALY will be sensitive to the method of utility measurement used
2. Reliability of RS and SG		
The ICC's of the reference health states were higher for the SG than for the RS, but low for both methods Mean scores of both methods were rather stable The standard error of measurement was about 0.12 for the SG (0–1 scale) and about 9 for the RS (0–100 scale)	Instability of intra-individual valuations due to difficulties in valuing hypothetical states, but stability of mean values at the group level Substantial measurement error inherent to utility measurement Precision decreases as choice is introduced	Limited use of utilities for individual decision making in clinical decision analysis More confidence in the use of utilities for program evaluation Need to undertake repeated measurements
3. Internal consistency SG		
Chained gambles resulted in higher utilities than basic reference gambles About 21% of all patients gave at least 1 inconsistent response About 70% of the inconsistent responses fell within the bounds of the standard error of measurement Almost 10% of all patients assigned a utility of 0.95 to three very different health states; they were not willing to take any risk	Outcomes of the gamble influence an individual patient's utility for a particular health state Inconsistent responses are to a large extend due to measurement error General aversion to gambling Inability to understand hypothetical, abstract questions Risk may seem higher from an individual than from a group perspective Probability increments of 10% are too large Difficulty of SG method Health state descriptions cover too many dimensions	Costs per QALY depend on the outcomes used in the SG Repeated measurements may reduce inconsistency SG not suited for everyone
4. A health state worse than death		
About 50% of the patients valued the severe reference state used in the chained gamble as worse than death on at least 1 occasion Negative utilities smaller than -1 occurred	Recognition of the existence of health states worse than death Low validity of the response	An appropriate way to incorporate negative utilities is needed Searching for alternative ways to calculate negative utilities The anchor point in chained gambles should not be worse than death

RS, rating scale; SG, standard gamble; QALY, quality adjusted life year.

ments would be expected to produce lower scores than the McMaster Health Utility Index, which is based on the standard gamble. When measuring preferences directly, one again has the choice between preference rankings or choice-based methods. Moreover, if one decides to use the standard gamble one also can choose between various ways of taking the measurements, ie, between basic reference gambles or chained gambles. It was found in our studies that these different versions of the same method resulted in statistically significant utility differences. Chained gambles resulted in higher utilities than the basic reference gamble.

Utilities are proposed as a decision aid in two different contexts: (1) where choices have to be made between alternative therapies for the same individual,[29] and (2) where choices have to be made between alternative ways of allocating limited resources among different health care activities serving the same or different patient groups.[46] There is cause for misgivings regarding the use of utilities for clinical decision analysis in the first context. For example, a number of inconsistencies were found in the responses of single patients to different standard gamble questions. It is likely that these reflect the underlying measurement error in taking a single preference measure. Our study indicates that the standard deviation of the measurement error ranged from 0.11 to 0.13 for the standard gamble and from 8 to 10 for the rating scale. Overall, more than 70% of the inconsistent responses fell within the bounds of the measurement error. The relatively poor stability of measurements from an individual patient limits the use of utilities based on a single measurement only for individual decision-making. This increases the need to undertake the measurements repeatedly to average out the measurement noise within the individual.

Even though the 3-months test-retest reliability was not very high, the relative stability of the mean utilities over time on the group level gives some confidence in the use of utilities for program evaluation in cost-utility analysis. However, because generally only a single estimate of utility is used in cost-utility analysis, this analysis may be very sensitive to the method used to elicit utilities. Recently Hornberger has shown that the effect of different methods on the final cost-utility may be considerable.[47] It is as yet premature to suggest a preferred method of utility measurement. Neither the rating scale nor the standard gamble method seems superior. As to the rating scale, there remain fundamental doubts about the interval properties of the scale.[48] Moreover, the repeatability of this method was found to be lower than that of the standard gamble. As to the standard gamble, it would be interesting to determine to what extent the observed difference between basic reference gambles and chained gambles influences the final results of cost-utility analysis.

Many of the inconsistencies and responses that seem to violate expected utility theory, which we found, may be associated with the description of the severe reference state used as the worst outcome in the first and as the certain outcome in the second of a chained pair of standard gambles. The difficulties patients had in valuing that health state is reflected in the fact that many patients changed their opinions as to whether the severe state was better or worse than death. Overall about 50% of the patients indicated at least once that the severe reference state was worse than death. When attempting to measure how much worse than death, a small number of highly negative utilities occurred that greatly influenced the mean utility. These methodological problems, particularly with regard to handling extreme negative utilities may argue in favor of using an anchor point in a chained pair of gambles that is not considered worse than death. However, in an era where the public tolerance of self-determination at the end of the life is growing, the problem of negative utilities calls for

intensified research efforts to handle these responses in utility calculations.

Because preference scores elicited by the standard gamble seem to be susceptible to the way questions are presented and to the endpoints used, doubts may be raised about the validity of expected utility theory. However, because there is no evidence that less formal procedures to guide individual therapy decisions and resource allocation decisions are any less susceptible to the effects of different methods of presentation and various biases, it is not appropriate to reject utilities as useful outcome measures. Moreover, as Torrance and Feeny point out, expected utility theory may be regarded as normative as opposed to behavioral.[28] This theory describes how an individual should behave if he or she wished to act rationally to maximize the expected utility. It does not describe how an individual actually makes a decision under uncertainty. Perhaps patients would have made more rational responses if they had been better informed about the meaning of their responses and the consequences of their choices. Furthermore, many of the observed inconsistencies or preferences that do not seem to fit expected utility theory are not necessarily illogical. Some of the alternatives to expected utility theory, such as prospect theory and regret theory, may help to explain several of the seemingly inconsistent answers from a more behavioral perspective. The challenge is to explore the potential contribution of such theories to utility measurements in health care decision settings.

Acknowledgments

The authors thank the anonymous reviewer, as well as George Torrance, Han Bleichrodt, Grant Rhodes, Silvia Evers, Margreet Janssen, Linda Heurman, and Mariëlle Goossens for their valuable comments on an earlier draft of this article.

References

1. Feeny D, Labelle R, Torrance GW. Integrating economic evaluations and quality of life assessments. In: Spilker B, ed. Quality of life assessment in clinical trials. ltd. New York, NY: Raven Press, 1990.

2. Bell MJ, Bombardier C, Tugwell P. Measurement of functional status, quality of life and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591.

3. Gerard K. Cost-utility in practice: A policy maker's guide to the state of the art. Health Policy 1992;21:249.

4. Rosser R, Kind P. A scale of valuations of states of illness: is there a social consensus? Int J Epidemiology 1978;7:347.

5. Kaplan RM, Bush JW. Health status: Types of validity and the index of well-being. Health Serv Res 1976;478.

6. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res 1982;30:1042.

7. Bombardier C, Ware J, Russell IJ, Larson M, Chalmers A, Read JL, and the Auranofin Cooperating Group. Auranofin therapy and quality of life in patients with rheumatoid arthritis. Results of a multicenter trial. Am J Med 1986;81:565.

8. Thompson MS, Read JL, Hutchings HC, Paterson M, Harris ED. The cost effectiveness of Auranofin: Results of a randomized clinical trial. J Rheumatol 1988;15:35.

9. Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and health-years equivalents. Med Decis Making 1989;9:142.

10. Torrance GW. Measurement of health state utilities for economic appraisal. A review. J Health Econ 1986;5:1.

11. Neumann Von J, Morgenstern O. Theory of games and economic behavior. 2nd ed. Princeton, NJ: Princeton University Press, 1947.

12. Froberg DG, Kane RL. Methodology for measuring health-state preferences-I: Measurement strategies. J Clin Epidemiol 1989;42:345.

13. Froberg DG, Kane RL. Methodology for measuring health-state preferences-II: Scaling methods. J Clin Epidemiol 1989;42:459.

14. Froberg DG, Kane RL. Methodology for measuring health-state preferences-III: Population and context effects. J Clin Epidemiol 1989;42:585.

15. Froberg DG, Kane RL. Methodology for measuring health-state preferences-IV: Progress and a research agenda. J Clin Epidemiol 1989;42:675.

16. Mulley AG. Assessing patients' utilities. Can the ends justify the means? Med Care 1989;27:S269.

17. Read JL, Quinn RJ, Berwick DM, et al. Preferences for health outcomes: comparisons of assessment methods. Med Decis Making 1984;4:315.

18. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. Socio Econ Plan Sci 1976;10:129.

19. Wolfson AD, Sinclair AJ, Bombardier C, McGreer. Preference measurement for functional status in stroke patients: Interrater and intertechnique comparisons. In: Kane RL, Kane RA, eds. Values and Long-Term Care. Lexington, KY: Lexington Books, D.C. Heath and Company, 1982.

20. Hellinger FJ. Expected utility theory and risky choices with health outcomes. Med Care 1989;27:273.

21. Llewellyn-Thomas HA, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. Med Decis Making 1982;2:449.

22. Hidding A, van der Linden Sj, Boers M, Gielen X, Kester A, de Witte L, Dijkmans B, Moolenburgh D. Is group physical therapy superior to individual therapy in ankylosing spondylitis. Arthritis Care Res 1993;6:117.

23. Van Santen-Hoeufft M, Bolwijn P, Van der Linden SJ. Effectiveness of fitness and biofeedback training in fibromyalgia. Submitted.

24. 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;31:1135.

25. Linden van der SJ, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modificration of the New York criteria. Arthritis Rheum 1984;27:361.

26. Bennett K, Torrance G. Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. Controlled Clin Trials 1991;12:118S.

27. Bakker CH, Rutten-van Mölken M, Van Doorslaer E, Bennett K, Van der Linden SJ. Health related utility measurement in rheumatology: An introduction. Patient Education and Counseling 1993;20:145.

28. Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care 1989;5:559.

29. Weinstein MC, Fineberg HV. Clinical Decision Analysis. Philadelphia, PA: W.B. Saunders Company, 1980.

30. Norusis/SPSS Inc. Advanced Statistics SPSS/PC+. Chicago, IL: Norusis/SPSS Inc, 1988.

31. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Controlled Clin Trials 1991;12:142S.

32. Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. New York, NY: Oxford University Press, 1989;83.

33. Kaplan RM, Bush JW, Berry CC. Health status index: category ratings versus magnitude estimation for measuring levels of well being. Med Care 1979;17:501.

34. Kahneman D, Tversky A. Prospect theory. An analysis of decision under risk. Econometrica 1979;47:263.

35. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 1981;211:453.

36. O'Brien BJ, Elswood J, Calin A. Willingness to accept risk in the treatment of rheumatic disease. J Epid Comm Health 1990;44:249.

37. Loomes G, Sugden R. Regret theory: an alternative theory of rational choice under uncertainty. Econ J 1982;92:805.

38. Loomes G, Sugden R. Some implications of the more general form of regret theory. J Economic Theory 1987;41:270.

39. Loomes G, Starmer C, Sugden R. Observing violations of transitivity by experimental methods. Econometrica 1991;59:425.

40. O'Connor AM, Boyd NE, Till JE. Methodological problems in assessing preferences for alternative therapies in oncology: The influence of preference elicitation technique, position order and test-retest error on the preferences for alternative cancer drug therapies. Nursing research: science for quality care. Proceedings of the 10th National Nursing Research Conference. Toronto, Ontario: University of Toronto, 1985:49.

41. Christensen-Szalanski JJJ. Discount functions and the measurement of patients' values. Women's decisions during childbirth. Med Decis Making 1984;4:46.

42. Bakker C, Rutten-van Mölken M, Van Doorslaer E, Bennett K, Van der Linden SJ. Feasibility of utility assessment by rating scale and standard gamble in ankylosing spondylitis or fibromyalgia. J Rheumatol, 1994. 1994;21:269.

43. Bleichrodt H. Testing the validity of expected utility theory in health state valuation: Some experimental results. Institute for Medical Technology Assessment paper no 93.23. Rotterdam, the Netherlands: Erasmus University of Rotterdam, 1993.

44. Haig THB, Scott DA, Wickett LI. The rational zero point for an illness index with ratio properties. Med Care 1986;24:113.

45. The EuroQol Group. A new facility for the measurement of health related quality of life. Health Policy 1990;16:199

46. Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Can Med Assoc J 1992;146:473.

47. Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J Clin Epidemiol 1992;45:505.

48. Nord E. The validity of a visual analogue scale in determining social utility weights for health states. Int J Health Planning and Management 1991;6:234.