Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Health Economics 23 (2004) 353–367
JOURNAL OF HEALTH ECONOMICS
www.elsevier.com/locate/econbase

A new and more robust test of QALYs

Jason N. Doctor[a,*], Han Bleichrodt[b], John Miyamoto[a], Nancy R. Temkin[a], Sureyya Dikmen[a]

[a] *Department of Medical Education, Division Biomedical and Health Informatics (MEBI), School of Medicine, University of Washington, 1959 N.E. Pacific Street, P.O. Box 357240, Seattle, WA 98195-6490, USA*
[b] *iMTA/iBMG, Erasmus University, Rotterdam, The Netherlands*

Received 1 March 2003; received in revised form 1 November 2003; accepted 25 November 2003

Abstract

Previous empirical tests of quality-adjusted life-years (QALYs), the most widely used outcome measure in economic evaluations of health care, generally yielded negative results. These tests were, however, for the most part based on expected utility, which is now widely acknowledged to be descriptively inaccurate. The observed violations might, therefore, have been caused by violations of expected utility. We performed a new test of QALYs, which is valid under expected utility and under the two most influential non-expected utility theories, rank-dependent utility and prospect theory, and found considerable support for the QALY model. Our findings suggest that QALYs may be valid if nonexpected utility formulas are used to compute health state utilities.

JEL classification: D81; I10

Keywords: Quality-adjusted life years; Rank-dependent utility; Prospect theory; Utility measurement

1. Introduction

The quality-adjusted life-years (QALY) model is the most widely used outcome measure in economic evaluations of health care. QALYs are computed by adjusting each year of life by the quality of life in which it is spent. They are intuitively appealing, i.e. easy to explain to doctors and policy makers, and are tractable for decision modeling, which explains their popularity in practical research. A disadvantage of the QALY model is that it represents

* Corresponding author. Tel.: +1-206-616-6640; fax: +1-206-616-3461.
E-mail address: jdoctor@u.washington.edu (J.N. Doctor).

0167-6296/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2003.11.004

individual preferences for health only under restrictive assumptions (Pliskin, Shepard & Weinstein, 1980; Bleichrodt et al., 1997). Empirical tests of the QALY assumptions have generally yielded negative results (McNeil et al., 1978; Verhoef et al., 1994). These findings undermine the credibility of economic evaluations based on QALYs and may call into question the validity of some clinical and health policy decision models. Tests of more general QALY models, in which the utility function over duration can be curved, have faired somewhat better (Bleichrodt et al., 1997; Miyamoto and Eraker, 1988). These more general models are, however, rarely utilized in applied studies.

The above mentioned tests of the validity of the QALY model were typically based on expected utility. It is now widely acknowledged that expected utility is not valid as a descriptive theory of decision under risk. Therefore, the possibility cannot be excluded that the observed violations of the QALY model were due to violations of expected utility. QALYs could be salvaged if they were found to hold under a descriptively accurate theory of choice under risk.

The most influential non-expected utility models are rank-dependent utility (Quiggin, 1981; Yaari, 1987) and prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Rank-dependent utility deviates from expected utility by permitting probability weighting. Prospect theory deviates from expected utility by permitting both probability weighting and loss aversion: outcomes are framed as gains and losses relative to a reference point and people are more sensitive to losses than to gains. Many studies show support for probability weighting (Bleichrodt et al., 1999; Gonzalez and Wu, 1999; Abdellaoui, 2000; Bleichrodt and Pinto, 2000) and loss aversion (Benartzi and Thaler, 1995; Bateman et al., 1997; Herne, 1998; Rabin, 2000).

The only test of QALYs performed hitherto that used a non-expected utility framework rejected the QALY model, but supported a more general QALY model in which the utility for duration was curved rather than linear (Bleichrodt and Pinto, 2001). The theoretical model assumed by Bleichrodt and Pinto (2001) is consistent with rank-dependent utility, but only with prospect theory when all outcomes are either gains or losses, i.e. in decision contexts where loss aversion plays no role. Given the importance of loss aversion, it is worthwhile to use a test of the QALY model that is generally valid under prospect theory.

The aim of the present paper is to perform such a new and more robust test of QALYs. Miyamoto (1999) showed that this test is valid under expected utility and rank-dependent utility. We show (Proposition 1) that the test is also valid under prospect theory. Contrary to previous studies, we find considerable support for the QALY model.

In what follows, Section 2 presents notation and structural assumptions. Section 3 briefly describes expected utility, rank-dependent utility, and prospect theory. Section 4 presents our main theoretical result, that the test we use is valid under all three theories of decision under risk. Sections 5 and 6 describe experimental procedures and results. Section 7 concludes. The proof of our main result is given in an appendix.

2. Notation and structural assumptions

We consider an individual who has to make a decision under risk. Throughout the paper, we consider only situations in which there are at most two possible states of the world.

The individual's decision problem is to choose between *prospects*. A prospect is a pair of health outcomes, one for each state of the world. A typical health outcome is (q,t) denoting t years in health state q. The durations t lie in an interval $\Omega = [0,M]$, where M stands for the maximum life duration. We shall write Ψ for the set of health states. A prospect yielding health outcome (q_1, t_1) with probability p and health outcome (q_2, t_2) with probability $1-p$ is denoted as $[(q_1, t_1), p; (q_2, t_2)]$.

A preference relation $\succcurlyeq$, meaning "at least as preferred as", is defined over the set of prospects Γ. As usual, we denote strict preference by $\succ$ and indifference by $\sim$. Preferences over health outcomes are derived from $\succcurlyeq$ by restricting attention to riskless prospects, i.e. prospects of the type $[(q, t), p; (q, t)]$.

Throughout the paper we shall assume that all prospects are *rank-ordered*. That is, when we write $[(q_1, t_1), p; (q_2, t_2)]$, we implicitly assume that $(q_1, t_1) \succcurlyeq (q_2, t_2)$. This assumption is not a restriction because each prospect can be written in a rank-ordered form by rearranging the outcomes.

We assume that for a given health state people prefer more life-years to less. A real-valued function V *represents* the preference relation $\succcurlyeq$ if for all prospects P, Q in Γ, $P \succcurlyeq Q$ if and only if $V(P) \geq V(Q)$.

If for some health states q_1, q_2, q_3 in Ψ, and for some life durations t_1, t_2, and t_3 in Ω, people are asked to state the probability p so that they are indifferent between the risky prospect $[(q_1, t_1), p; (q_3, t_3)]$ and the riskless prospect (q_2, t_2), we say that p is the *probability equivalent* of health outcome (q_2, t_2) with respect to health outcomes (q_1, t_1) and (q_3, t_3). The *standard gamble*, a widely used technique to elicit health state utilities, is based on the determination of probability equivalents. Standard gamble measurements typically use $t_1 = t_2$, $t_3 = 0$, and $q_1 =$ full health.

3. Expected utility, rank-dependent utility, and prospect theory

Expected utility holds if preferences over prospects $[(q_1, t_1), p; (q_2, t_2)]$ can be represented by $pU(q_1, t_1) + (1 - p)U(q_2, t_2)$, where U is a utility function over health outcomes, which is unique up to unit and location. In expected utility probabilities are evaluated linearly.

Rank-dependent utility generalizes expected utility by allowing probability weighting. Rank-dependent utility holds if preferences over prospects $[(q_1, t_1), p; (q_2, t_2)]$ can be represented by $w(p)U(q_1, t_1) + (1 - w(p))U(q_2, t_2)$, where U is a utility function over health outcomes which is unique up to unit and location and w is a probability weighting function that has $w(0) = 0$, $w(1) = 1$, and that is strictly increasing, (i.e. $p > q$ if and only if $w(p) > w(q)$). Empirical studies have found that the probability weighting function is inverse S-shaped, overweighting small probabilities and underweighting large probabilities (Tversky and Kahneman, 1992; Gonzalez and Wu, 1999; Abdellaoui, 2000; Bleichrodt and Pinto, 2000). The most widely used parametric specification of the probability weighting function is that given by Tversky and Kahneman (1992):

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{1/\gamma}} \tag{1}$$

This function has an inverse S-shape for $0.27 \leq \gamma \leq 1$. Probability weighting can explain several of the violations of expected utility commonly found in empirical studies. Rank-dependent utility reduces to expected utility when $w(p) = p$. If $w(p)$ is described by Expression (1), this occurs when $\gamma = 1$.

Like rank-dependent utility, *prospect theory* (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) posits a nonlinear probability weighting function. In addition, prospect theory assumes *sign-dependence*: outcomes are perceived as gains and losses relative to a reference point. People are assumed to be more sensitive to losses than to gains, a phenomenon known as *loss aversion*. The perception of outcomes as gains, losses, or the reference point may be susceptible to framing effects (Hershey and Schoemaker, 1985; Bleichrodt et al., 2001). There is a separate probability weighting function for gains, w^+, and for losses w^-. Using Expression (1), Tversky and Kahneman (1992) found that γ was equal to 0.61 for gains and to 0.69 for losses.

Prospect theory leads to three types prospects: (1) *pure gains prospects*, in which all outcomes are perceived as gains, (2) *pure losses prospects*, in which all outcomes are perceived as losses, and (3) *mixed prospects*, in which one outcome is perceived as a gain and the other as a loss. In our experiment, reported in Section 5, we asked subjects standard gamble questions. Empirical evidence suggests that standard gamble questions are subject to a framing effect and are evaluated as mixed prospects (Hershey and Schoemaker, 1985; Bleichrodt et al., 2001). Therefore, our focus with prospect theory is on the mixed case only.

Studying monetary outcomes, Hershey and Schoemaker (1985) were the first to hypothesize a framing effect for standard gamble questions, where subjects take the riskless outcome as the reference point, perceive the better outcome in the risky prospect as a gain and the worst outcome as a loss. Bleichrodt et al. (2001) formalized this hypothesis under prospect theory. They showed that in standard gamble questions, comparisons between $[(q_1, t_1), p; (q_3, t_3)]$ and (q_2, t_2) are evaluated as

$$w^+(p)(U(q_1, t_1) - U(q_2, t_2)) - \lambda w^-(1-p)((U(q_2, t_2) - U(q_3, t_3)) = 0 \tag{2}$$

The equation reflects the assumption that outcomes are evaluated as deviations from the reference point, (q_2, t_2), through terms $U(q_i, t_i) - U(q_2, t_2)$. The utility function U is unique up to unit and location. The parameter λ reflects loss aversion. Tversky and Kahneman (1992) estimated λ to be 2.25. Expression (2) shows that rank-dependent utility is the special case of prospect theory where $\lambda = 1$ and $w^-(1-p) = 1 - w^+(p)$. Expected utility is the special case of prospect theory where $\lambda = 1$ and $w^+(p) = w^-(p) = p$.

4. The main result

The focus of this paper is on the most widely used QALY model, the linear QALY model. Whenever we refer to the QALY model we will mean the linear model. Under the *QALY model* the utility function U in expected utility, rank-dependent utility, and prospect theory takes the form

$$U(q, t) = H(q)t, \tag{3}$$

where H is a health utility function that assigns a positive index to every conceivable health state. Often H is scaled so that H (full health) $= 1$. In many applications QALYs are discounted at a constant rate to reflect that people do not attach equal weight to different years of life. Such a discounted utility model is conceptually similar to Expression (3), except of course that the model is linear in discounted life-years rather than in undiscounted life-years.

To give a preference foundation for the QALY model, the following three conditions are used.

Definition 1. Preferences satisfy *solvability* with respect to survival duration if for all life durations t_1, t_2, t_3 in Ω, and for all health states q_1, q_2 in Ψ, if $(q_1, t_1) \succcurlyeq (q_2, t_2) \succcurlyeq (q_1, t_3)$, then there exists a life duration t_4 in Ω such that $(q_1, t_4) \sim (q_2, t_2)$.

Solvability guarantees that utility is a continuous function of survival duration. Continuity of utility with respect to survival duration is commonly assumed in medical decision-making and is implied by the QALY model.

Definition 2. Preferences satisfy the *zero-condition* if for all health states q_1, q_2 in Ψ, $(q_1, 0) \sim (q_2, 0)$.

The zero-condition is self-evident in the medical context, because for any h and h' in Q, $(q_1, 0)$ and $(q_2, 0)$ are indistinguishable under the interpretation of time as survival duration (Miyamoto and Eraker, 1988; Bleichrodt et al., 1997; Miyamoto et al., 1998).

Definition 3. *Constant proportional coverage* holds if for all health states q in Ψ, and for all life durations t_1, t_2, t_3, t_1', t_2', t_3' in Ω with $t_1 > t_2 > t_3$ and $t_1' > t_2' > t_3'$, if $[(q,t_1), p_1; (q, t_3)] \sim (q, t_2)$, $[(q, t_1'), p_2; (q, t_3')] \sim (q, t_2')$ and $(t_2 - t_3)/(t_1 - t_3) = (t_2' - t_3')/(t_1' - t_3')$ then $p_1 = p_2$.

Constant proportional coverage is somewhat similar to Pliskin et al.'s (1980) condition of constant proportional tradeoffs but with respect to risky decisions. Constant proportional tradeoffs says that if an individual considers health outcome (q, t) indifferent to health outcome (q', t') then this indifference should still hold if we multiply the durations t and t' by some common positive number α. Constant proportional coverage implies that if the individual is indifferent between $[(q, t_1), p_1; (q, t_3)]$ and (q, t_2) then he should also be indifferent between $[(q, \alpha t_1), p_1; (q, \alpha t_3)]$ and $(q,\alpha t_2)$ with $\alpha > 0$ and both prospects in Γ. Constant proportional coverage is, however, more than just a translation of constant proportional tradeoffs to decisions under risk, for it also implies that if the individual is indifferent between $[(q, t_1), p_1; (q, t_3)]$ and (q, t_2) then he will also be indifferent between $[(q, \alpha + t_1), p_1; (q, \alpha + t_3)]$ and $(q, \alpha + t_2)$ with α real and both prospects in Γ. The above analysis shows that constant proportional coverage implies both constant proportional risk aversion and constant absolute risk aversion with respect to duration risk. The only utility function that is consistent with both constant proportional risk aversion and constant absolute risk aversion is the linear one.

We can now state the main theoretical result of this paper, a proof of which is given in Appendix A.

Proposition 1. *Suppose that the structural assumptions given in* Section 2 *hold and that prospect theory holds. Then the following two statements are equivalent*:

(i) *The QALY model holds*.
(ii) *Solvability* (Definition 1), *the zero-condition* (Definition 2) *and constant proportional coverage* (Definition 3) *hold*.

Because expected utility and rank-dependent utility are special cases of prospect theory, Proposition 1 also holds under rank-dependent utility and expected utility. If we substitute discounted life-years for undiscounted life-years in the definition of constant proportional coverage, i.e. we require constant proportional coverage for discounted life-years, then Proposition 1 gives a preference foundation for the QALY model with constant rate discounting.

As noted above, solvability and the zero-condition are widely accepted and will be assumed in our empirical analysis. Proposition 1, therefore, shows that, under prospect theory, constant proportional coverage is the central assumption of the QALY model. As long as prospect theory holds, it is possible to identify persons that satisfy the QALY model by examining if constant proportional coverage holds. This is the empirical test that we performed.

Several authors have provided preference foundations for the QALY model. Pliskin et al. (1980) were the first to do so under expected utility assumptions. They showed that the QALY model holds if and only if utility independence, constant proportional tradeoffs, and risk neutrality with respect to life duration hold. Later, Bleichrodt et al. (1997) simplified Pliskin et al.'s (1980) characterization of the QALY model by using the zero condition to show that only risk neutrality needs to be imposed. Under expected utility, risk neutrality implies constant proportional coverage. Bleichrodt et al. (1997) axiomatized the QALY model under rank-dependent utility. Their axiomatization hinged on the assumption of constant marginal utility for life duration. Miyamoto (1999) showed that constant proportional coverage is the key axiom for QALYs under rank-dependent utility. Proposition 1 extends Miyamoto's result to prospect theory. Finally, Bleichrodt and Miyamoto (2003) gave a preference foundation for several types of QALY models under prospect theory using different axioms than we use here.

5. Experiment

5.1. Participants

Participants were 48 patients that had sustained a head injury 6 months prior to the interview. They completed the study as part of a larger randomized clinical trial examining the effectiveness of magnesium sulfate administered in the emergency department as a neuroprotectant after head injury. The average subject was male (80.9%), 27.2 years of age (S.D. $\pm$ 13.2 years), and had at least a high school education.

Table 1
The four standard gamble questions[a] varying in survival duration, and presented each in full health and in current health but constrained to have the same proportional coverage, such that for all t_1, t_2 and t_3, $(t_2 - t_3)/(t_1 - t_3) = 5/8$

Standard gamble question	Risky gamble (t_1, p; t_3)		Riskless gamble
	t_1 (years)	t_3 (years)	t_2 (years)
1	20	4	14
2	16	0	10
3	15	7	12
4	10	2	7

[a] The questions involved having the subject examine a chance board with a probability wheel with a pink and a white partition. In a typical question, the examiner would say, "If you choose Choice A, I would spin this imaginary wheel. There is a 90% chance the pointer will land on pink, in which case you would live 20 years in full health. There is a 10% chance it will not land on pink, in which case you would live 4 years in full health. Choice B shows a 100% chance of living 14 years in full health. Which do you prefer, Choice A, spinning the wheel, Choice B, or are the choices about equal?".

5.2. Research design

Table 1 displays the four standard gamble questions that were asked. Probability equivalents were elicited both in full health and in current health. Full health and current health were described by the EQ-5D (The EuroQol Group, 1990), a widely used instrument to describe health states. Observations were replicated once for both full health and for current health yielding 16 probability equivalents per subject.

The standard gamble questions were constructed such that for each standard gamble question $[(q, t_1), p; (q, t_3)] \sim (q, t_2)$, $(t_2 - t_3)/(t_1 - t_3) = 5/8$. Because $(t_2 - t_3)/(t_1 - t_3)$ is constant, constant proportional coverage predicts that the probability equivalents should be the same across questions. Our empirical test was to examine whether this was indeed true. To avoid order effects in the group analysis, we varied the order in which the stimuli were administered. Half the subjects first faced the "full health" choices, the other half first faced the "current health" choices. This counterbalancing was extended also to the replication choices.

In addition to the stimuli in the primary research design, an additional stimulus was given to subjects to check the validity of the responses. A probability equivalent was elicited for the choice [(Full Health, 20 years), p; (Full Health, 1 year)] versus (Full Health, 2 years). Under all theories tested, this standard gamble question was predicted to produce a lower probability equivalent than all other standard gamble questions, since the riskless prospect is highly unattractive relative to the risky prospect. If subjects did not produce a lower probability equivalent for this question, their data were not analyzed. In such cases, it is possible that subject response patterns were repetitive from stimulus to stimulus, which would generate artificial support for the QALY model because these subjects will always satisfy constant proportional coverage. A subject might use such a response pattern to avoid consideration of the questions and end the exercise quickly.

Finally, subjects were asked to compare [(Full health, 20 years), p; (Full Health, 0 years)] versus (Current Health, 20 years). This question was included to test whether subjects considered their current health as worse than full health. If they considered their current health equivalent to full health then it did not make sense to collect replication data, because the

questions with current health were already equivalent to the questions with full health. Hence, for these subjects we had just four observations replicated once, totaling 8 observations. The number of observations on these subjects was too low to be able to detect an effect with a reasonable chance. Including these subjects in the analysis would inflate support for the QALY model and, therefore, they were excluded from the analysis.

5.3. Procedure

A trained interviewer, blind to the hypothesis, tested subjects. Subjects were first trained in the use of the standard gamble method for prospects over different survival durations. The examiner also gave subjects sample choices with feedback on their responses. All probability equivalents were elicited using a sequence of choices, starting with extreme probabilities, first a question with probability 1 of the best outcomes then a question with probability 1 of the worst outcome, and "ping-ponging" to the probability equivalent. The sequence of values offered in the ping-pong approach was the same for all subjects.

Expected utility, rank-dependent utility and prospect theory each satisfy *elementary stochastic* dominance, i.e. the requirement that if we increase the probability of the most desirable health outcome, then utility should increase: if $p > q$ and $(q_1, t_1) \succ (q_2, t_2)$, then $[(q_1, t_1), p; (q_2, t_2)] \succ [(q_1, t_1), q; (q_2, t_2)]$. Using practice exercises it was examined whether subjects satisfied elementary stochastic dominance. Subjects whose preferences were not consistent with elementary stochastic dominance received additional training. If after the additional training, their preferences still were inconsistent with elementary stochastic dominance, the assessments would be discontinued. Replication of the standard gamble questions took place after a 30 min intervening task. The time to complete the full questionnaire, i.e. including the intervening task, averaged 65 minutes.

5.4. Statistical analysis

The hypothesis that the elicited probability equivalents were equal across questions was tested for each subject by analysis of variance and by the nonparametric Kruskal–Wallis rank sum test. For the ANOVA we used a significance level of 5%. For the Kruskal–Wallis test we used a significance level of 10% to compensate for the lower statistical power of nonparametric tests.

A problem of single subject analysis may be low statistical power. This is because subject fatigue and memory for previous responses prohibit the large number of observations needed to achieve high levels of statistical power. To assess the problem of statistical power for the single subject statistical tests we performed a power analysis. We assumed that subjects' true utility function for duration was not linear, but a power function, $U(q, t) = H(q)t^r$. We analyzed the statistical power, i.e. the probability of correctly rejecting the null hypothesis that the utility function for duration is linear, for three values of the power coefficient $r : r = 2/3$, $r = 1/3$, and $r = 1/6$. When the power coefficient r equals 1 the utility function is linear. As $r < 1$ gets smaller the utility function becomes more and more concave indicating a larger deviation from linearity. Table 2 shows the proportion of the sample for which the linear QALY model could be correctly rejected at three different levels of power. For example, the upper left entry of the table shows that for 47% of the

Table 2
The proportion of the sample for whom the linear QALY model could be rejected at three different levels of statistical power, $1-\beta$, assuming one of three values of the power coefficient, r

		$1-\beta$		
		0.50	0.80	0.95
r	2/3	0.47	0.32	0.32
	1/3	0.76	0.59	0.59
	1/6	0.86	0.76	0.62

sample the probability of correctly rejecting the linear QALY model when their true utility function was a power function with coefficient 2/3 exceeded 50%. The table shows that the power was generally reasonable, except perhaps for relatively small deviations from linearity ($r = 2/3$). Technical details of the power analysis are given in Appendix B.

In addition to the individual analyses, we performed a group analysis on the data to get an indication of whether individuals on average deviate systematically from the QALY model. The group analysis consisted of a repeated measures analysis of variance using a mixed effects model (Maxwell and Delaney, 1990). The group analysis modeled the effect of subject, lottery, replication, and interactions. Because of the far greater degrees of freedom in the group analysis, statistical power is improved.

We assessed the test–retest reliability of the standard gamble method by examining how responses at the two time points correlate within each subject across the eight stimuli and by examining whether significant differences occurred between mean responses at the two time points. Note that we used the replication data both to test whether constant proportional coverage is satisfied and to test consistency. This double use of the data is not problematic. A subject can at the same time be perfectly consistent and also violate constant proportional coverage.

6. Results

Of the 48 subjects, nine were dropped for not having a lower probability equivalent on the prospect designed to lower their probability equivalent. Five subjects were dropped because they reported being in full health at the time of the interview. Data on the remaining 34 subjects were analyzed. Results of the probability equivalent exercise are summarized in Table 3, which gives means and standard deviations by experimental choice condition.

As is clear from this table, means and standard deviations are fairly similar across questions. The analysis based on medians and interquartile ranges was similar. We found a median of 0.85 for seven of the eight conditions (Table 1, Question 4s median = 0.88 for full health) and interquartile ranges of 0.20 for five of the eight conditions, 0.25 for two of the eight conditions (Table 1, Question 3 Full and Current Health) and 0.21 for one condition (Table 1, Question 1 Full Health).

Table 3 shows that most subjects were risk averse with respect to duration, i.e. they preferred the expected value of a risky prospect to the prospect itself. Risk neutrality with respect to duration would have meant a probability equivalent of 0.625 in all questions. The

Table 3
Mean (M) and standard deviation (S.D.) by prospect for probability equivalent data among subjects in the data analysis ($N = 34$)

Choice equivalence	Health state	
	Full health	Current health
(20 years, PE; 4 years) ∼ 14 years		
M	0.81	0.84
S.D.	0.18	0.13
(16 years, PE; 0 years) ∼ 10 years		
M	0.81	0.83
S.D.	0.19	0.14
(15 years, PE; 7 years) ∼ 12 years		
M	0.80	0.81
S.D.	0.20	0.17
(10 years, PE; 2 years) ∼ 7 years		
M	0.84	0.82
S.D.	0.14	0.17

finding of risk aversion with respect to duration is consistent with the findings from other studies (McNeil et al., 1978; Verhoef et al., 1994).

The replication results were satisfactory. The mean and median within-subject correlation coefficients between test and retest were 0.75 and 0.82, respectively. No significant differences occurred between mean responses at the two time points. The fact that we find good consistency suggests that the data are not too noisy.

We retained the null hypothesis that constant proportional coverage holds, and hence that the QALY utility model accurately describes preferences for health outcomes, for 27 of the 34 subjects (79%) with ANOVA and for 29 of the 34 subjects (85%) with Kruskal–Wallis. The five subjects who violated constant proportional coverage under the Kruskal–Wallis test also violated the QALY model under the ANOVA test. The group analysis also showed support for the QALY model. As predicted by the QALY model, there was no significant effect for lottery across subjects ($F(7, 231) = 0.5813$, $P = \text{NS}$).

7. Conclusion

We find considerable support for the QALY model when we use a test that is robust to probability transformation and loss aversion, two important reasons why expected utility is violated. Like previous studies, we find strong aversion to duration risks. However, under non-expected utility risk aversion is not inconsistent with a utility function for duration that is linear. This observation may be puzzling for economists who are used to think that risk aversion corresponds to concave utility. Let us explain by means of one of our data pairs that this one-to-one correspondence between risk aversion and concave utility no longer holds under non-expected utility.

One of our subjects indicated that he was indifferent between the risky prospect giving a probability of 0.85 of 20 years in good health and a probability of 0.15 of 4 years in good

health and the riskless prospect giving 14 years in good health. Clearly, this subject was averse to duration risks: to be risk neutral he should have stated an indifference probability of 0.625. If we scale utility so that U(20 years) = 1 and U(4 years) = 0, then it follows under expected utility that U(14 years) = 0.85 and the utility function is clearly concave. Under rank-dependent utility, if we use Expression (1) with the estimate obtained by Tversky and Kahneman (1992), it follows that U(14 years) = 0.65, which is close to 0.625, the value that corresponds to linearity of utility. Under prospect theory, it follows by Expression (2) with the estimates obtained by Tversky and Kahneman (1992) that U(14 years) = 0.57 and the utility function is slightly convex. Hence, the deviations from expected utility modeled by rank-dependent utility and prospect theory, can reconcile risk aversion with linear utility.

The joint findings of risk aversion with respect to duration and no rejection of the hypothesis that the QALY model holds, suggest that our subjects did not behave according to expected utility. Under expected utility, these two findings are mutually exclusive. The common approach in cost-effectiveness analysis is to measure effectiveness by the expected change in QALYs and to compute the QALY weights, the health state utilities, by assuming expected utility. Previous studies have generally criticized the common approach for assuming the QALY model. Our findings suggest that the QALY assumption may be defensible, but that the expected utility assumption is problematic. Instead, health state utilities should be computed under rank-dependent utility or prospect theory. How this can be done is explained in Wakker and Stiggelbout (1995) for rank-dependent utility and in Bleichrodt et al. (2001) for prospect theory.

The recommendation to use rank-dependent utility or prospect theory in health utility measurement means that probability weighting and loss aversion should be taken into account in the computation of health state utilities. The question then arises whether it is appropriate to do so, given that cost-effectiveness analysis is a prescriptive exercise and expected utility is still the dominant prescriptive theory of decision under risk. We believe that it is appropriate. Elicitation of a utility is essentially a *descriptive* activity because it concerns observed behavior. To elicit utilities we should, therefore, use the theory that is descriptively most accurate. Basing utility elicitation on a theory that is descriptively inaccurate, such as expected utility, will lead to biased utilities and, consequently, to inaccurate treatment recommendations.

Our findings are consistent with other studies that have shown that utility is less curved under non-expected utility than under expected utility. See Wakker and Deneffe (1996) and Bleichrodt et al. (1999) when the outcome domain consists of life durations. See also Edwards (1955), Fox et al. (1996), Selten et al. (1999), Luce (2000), Rabin (2000), and Diecidue and Wakker (2002), when the outcome domain consists of moderate amounts of money. It is important to emphasize that these findings of linear utility are empirical observations and not theoretical restrictions; nothing about prospect theory or rank-dependent utility guarantees that the utility for duration is linear. Therefore, in no way were the findings of our study predetermined.

Our empirical study has several limitations. A first limitation is that we were unable to examine the responses from all 48 subjects. However, limiting the sample the way we did insured that we only analyzed data for subjects that put forth a good effort and clearly understood the procedure. Second, the ‘ping-pong’ procedure used, although common in

health utility measurement, may have led to potential anchoring toward the first probability in the choices. If anchoring occurs it is unlikely to have eliminated differences between probability equivalents, but may have adjusted them upwards. Hence, anchoring is unlikely to have affected our conclusions. A third, and perhaps most important, limitation of our study is its statistical power. For only one in three subjects, there was adequate power to detect violations of linearity when the power coefficient in the utility function over duration equals two-thirds. There is some empirical evidence that the power coefficient in the utility function over duration is around 0.7 (Bleichrodt and Pinto, 2000), which suggests that we may have been unable to pick up some deviations from the QALY model. On the other hand, it is arguable how important such deviations from linearity are for practical research in the sense that they lead to different treatment recommendations. Moreover, it is encouraging that the group analysis, a test conducted under greater statistical power than the individual tests, also failed to indicate that the average subject deviated systematically from the QALY model.

The QALY model offers important advantages, such as intuitive appeal and tractability. It has often been argued that these important advantages conflict with descriptive accuracy. Our study shows that this is not necessarily true and that QALYs, when nonexpected utility is used to compute health state utilities, may be a better description of individual preferences for health than is commonly thought.

Acknowledgements

David Cutler (the editor), Peter Wakker, and two anonymous referees provided helpful comments. Jason Doctor's research was made possible by a grant from the United States Department of Health and Human Services, National Institutes of Health, National Center for Medical Rehabilitation Research (NIH-NCMRR: K01HD01221). Han Bleichrodt's research was made possible by a grant from the Netherlands Organisation for Scientific Research (NWO).

Appendix A. Proofs

First we state a lemma that is used in the proof of Proposition 1.

Lemma 1. For all t_1, t_2, t_3, t_1', t_2', t_3' in Ω with $t_1 > t_2 > t_3$ and $t_1' > t_2' > t_3'$, $(t_2 - t_3)/(t_1 - t_3) = (t_2' - t_3')/(t_1' - t_3')$ if and only if $(t_2 - t_3)/(t_1 - t_2) = (t_2' - t_3')/(t_1' - t_2')$.

Proof.

[i] $\frac{t_2 - t_3}{t_1 - t_3} = \frac{t_2' - t_3'}{t_1' - t_3'}$ implies
[ii] $(t_2 - t_3)(t_1' - t_3') = (t_2' - t_3')(t_1 - t_3)$, implies
[iii] $t_2 t_1' - t_2 t_3' - t_3 t_1' = t_2' t_1 - t_3' t_1 - t_3 t_2'$, implies
[iv] $t_2 t_1' - t_2 t_3' - t_3 t_1' - t_2 t_2' = t_2' t_1 - t_3' t_1 - t_3 t_2' - t_2 t_2'$, implies
[v] $(t_2 - t_3)(t_1' - t_2') = (t_2' - t_3')(t_1 - t_2)$, implies

[vi] $\frac{t_2-t_3}{t_1-t_2} = \frac{t'_2-t'_3}{t'_1-t'_2}$.

To see that $\frac{t_2-t_3}{t_1-t_2} = \frac{t'_2-t'_3}{t'_1-t'_2}$, implies $\frac{t_2-t_3}{t_1-t_3} = \frac{t'_2-t'_3}{t'_1-t_3}$, follows steps [vi]–[i] in descending order. □

Proof of Proposition 1. (ii)⇒(i). Suppose the prospect theory functional represents preferences over mixed prospects. Solvability implies that the utility function is continuous with respect to survival duration. Fix p in (0,1). Constant proportional coverage implies that for all q in Ψ, for all t_1, t_2, t_3 in Ω with $t_1 > t_2 > t_3$ and for all real numbers s such that $t_1 + s$, $t_2 + s$, $t_3 + s$ are in Ω, $[(q, t_1), p; (q, t_3)] \sim (q, t_2)$ if and only if $[(q, t_1 + s), p; (q, t_3 + s)] \sim (q, t_2 + s)$. That is, constant absolute risk aversion holds. Let $U_q(t) = U(q, t)$ and let $f(t) = t + s$. Both U_q and f are continuous. By constant absolute risk aversion, $U_q \circ f$ and U_q both represent preferences and thus, because the utility function in prospect theory is an interval scale, $U_q \circ f(t) = U_q(t + s) = \alpha(s)U_q(t) + \beta(s)$. This is a functional equation with as solution that U_q is linear or exponential (Aczel, 1966, p. 150).

Constant proportional coverage also implies that for all q in Ψ, for all t_1, t_2, t_3 in Ω with $t_1 > t_2 > t_3$ and for all real numbers s, such that st_1, st_2, st_3 are in Ω, $[(q, t_1), p; (q, t_3)] \sim (q, t_2)$ if and only if $[(q, st_1), p; (q, st_3)] \sim (q, st_2)$. That is, constant proportional risk aversion holds. Let $g(t) = st$. Clearly, g is continuous. By constant proportional risk aversion, $U_q \circ g$ and U_q both represent preferences over prospects and thus $U_q \circ g(t) = U_q(st) = \alpha(s)U_q(t) + \beta(s)$. This functional equation has as solution that U_q is a power or logarithmic function (Aczel, 1966, p. 150). Hence, U_q must be linear, $U_q(t) = \alpha t + \beta$.

Because people prefer more life-years to less, α is positive. Hence, $U(q, t) = H(q)t + G(q)$ with $H(q)$ positive. By the zero-condition $G(q)$ is independent of q and by the uniqueness properties of U it can be set equal to zero. The QALY model follows.

(i)⇒(ii). Suppose that the QALY model holds. It immediately follows that the zero condition and solvability hold. It remains to show that constant proportional coverage holds. Suppose that for some $t_1, t_2, t_3, t'_1 t'_2 t'_3$, in Ω with $t_1 > t_2 > t_3$, $t'_1 > t'_2 > t'_3$, and $(t_2 - t_3)/(t_1 - t_3) = (t'_2 - t'_3)/(t'_1 - t'_3)$ we find that $[(q, t_1), p_1; (q, t_3)] \sim (q, t_2)$ and $[(q, t'_1), p_2; (q, t'_3)] \sim (q, t'_2)$. We must show that $p_1 = p_2$.

Suppose that prospect theory holds. Lemma 1, Expression (2), and the QALY model give $(w^+(p_1))/(\lambda w^-(1-p_1)) = (t_2-t_3)/(t_1-t_3) = (t'_2-t'_3)/(t'_1-t'_3) = (w^+(p_2))/(\lambda w^-(1-p_2))$. Suppose that $p_1 > p_2$. Then $w^+(p_1) > w^+(p_2)$ and $w^-(1 - p_2) > w^-(1 - p_1)$. Hence $(w^+(p_1))/(\lambda w^-(1 - p_1)) > (w^+(p_2))/(\lambda w^-(1 - p_2))$, which is a contradiction. By a similar argument, $p_1 < p_2$ leads to a contradiction. □

Appendix B. Details of the power analysis

We conducted a power analysis for each subject's data under several different assumptions about nonlinearity of utility. Using methods described in Maxwell and Delaney (1990, p. 113–115) we computed the effect size, ϕ, as:

$$\phi = \frac{\sigma_m}{\sigma_e} n^{1/2}, \tag{A.1}$$

where σ_m is the standard deviation of the nonlinear effect, or the square root of the mean square for effect of treatment, and σ_e is the square root population within-cell error.

Under a nonlinear power utility function, utility would take the form

$$U(q, t) = H(q)t^r, \tag{A.2}$$

where r is a utility curvature parameter and $r \neq 1$ for violations of the QALY model. We computed the standard deviation of the nonlinear effect, σ_m, under three scenarios for Eq. (A.2): (1) $r = 2/3$, (2) $r = 1/3$ and (3) $r = 1/6$. Because deviations from linearity are symmetric about the linear case, we need only test deviations for the concave case to determine statistical power.

Following the methods of Miyamoto and Eraker (1988), we computed the mean square error (MS_{error}) for each subject using results from each individual ANOVA. We then used the square root of this value as an estimate of σ_e, (MS_{error} is an unbiased estimator of population within-cell variance). Finally, with two observations per cell, $n = 2$ in Eq. (A.1) and the computation of the effect size for each subject under the three aforementioned assumptions of nonlinearity was complete.

References

Abdellaoui, M., 2000. Parameter-free elicitation of utilities and probability weighting functions. Management Science 46, 1497–1512.

Aczel, J., 1966. Lectures on Functional Equations and Their Applications. Academic Press, New York.

Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R., 1997. A test of the theory of reference-dependent preferences. Quarterly Journal of Economics 62, 479–505.

Benartzi, S., Thaler, R.H., 1995. Myopic loss aversion and the equity premium puzzle. Quarterly Journal of Economics 110, 73–92.

Bleichrodt, H., Miyamoto, J., 2003. A characterization of quality-adjusted life-years under cumulative prospect theory. Mathematics of Operations Research 28, 181–193.

Bleichrodt, H., Pinto, J.L., 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. Management Science 46, 1485–1496.

Bleichrodt, H., Pinto J.L., 2001. The Validity of QALYs under non-expected utility. Working Paper Erasmus University.

Bleichrodt, H., Pinto, J.L., Wakker, P.P., 2001. Using descriptive findings of prospect theory to improve the prescriptive use of expected utility. Management Science 47, 1498–1514.

Bleichrodt, H., van Rijn, J., Johannesson, M., 1999. Probability weighting and utility curvature in QALY based decision making. Journal of Mathematical Psychology 43, 238–260.

Bleichrodt, H., Wakker, P.P., Johannesson, M., 1997. Characterizing QALYs by risk neutrality. Journal of Risk and Uncertainty 15, 107–114.

Diecidue, E., Wakker, P.P., 2002. Dutch books: avoiding strategic and dynamic complications, and a comonotonic extension. Mathematical Social Sciences 43, 135–149.

Edwards, W., 1955. The prediction of decisions among bets. Journal of Experimental Psychology 50, 201–214.

Fox, C.R., Rogers, B.A., Tversky, A., 1996. Options traders exhibit subadditive decision weights. Journal of Risk and Uncertainty 13, 5–17.

Gonzalez, R., Wu, G., 1999. On the form of the probability weighting function. Cognitive Psychology 38, 129–166.

The EuroQol Group, 1990. EuroQol: a new facility for the measurement of health related quality of life. Health Policy 16, 199–208.

Herne, K., 1998. Testing the reference-dependent model: an experiment on asymmetrically dominated reference points. Journal of Behavioral Decision Making 11, 181–192.

Hershey, J.C., Schoemaker, P.J.H., 1985. Probability versus certainty equivalence methods in utility measurement: are they equivalent? Management Science 31, 1213–1231.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291.
Luce, R.D., 2000. Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.
Maxwell, S., Delaney H., 1990. Designing Experiments and Analyzing Data. Wadworth, Belmont, CA.
McNeil, B.J., Weichselbaum, R., Pauker, S.G., 1978. Fallacy of the five-year survival in lung cancer. New England Journal of Medicine 299, 1397–1401.
Miyamoto, J.M., 1999. Quality-adjusted life-years (QALY) utility models under expected utility and rank dependent utility assumptions. Journal of Mathematical Psychology 43, 201–237.
Miyamoto, J.M., Eraker, S.A., 1988. A multiplicative model of the utility of survival duration and health quality. Journal of Experimental Psychology: General 117, 3–20.
Miyamoto, J.M., Wakker, P.P., Bleichrodt, H., Peters, H.J.M., 1998. The zero-condition: a simplifying assumption in QALY measurement and multiattribute utility. Management Science 44, 839–849.
Pliskin, J.S., Shepard, D.S., Weinstein, M.C., 1980. Utility functions for life years and health status. Operations Research 28, 206–223.
Quiggin, J., 1981. Risk perception and risk aversion among australian farmers. Australian Journal of Agricultural Economics 25, 160–169.
Rabin, M., 2000. Risk aversion and expected-utility theory: a calibration theorem. Econometrica 68, 1281–1292.
Selten, R., Sadrieh, A., Abbing, K., 1999. Money does not induce risk neutral behavior, but binary lotteries do even worse. Theory and Decision 46, 211–249.
Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323.
Verhoef, L.C.G., de Haan, A.F.J., van Daal, W.A.J., 1994. Risk attitude in gambles with years of life: empirical support for prospect theory. Medical Decision Making 14, 194–200.
Wakker, P.P., Deneffe, D., 1996. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. Management Science 42, 1131–1150.
Wakker, P.P., Stiggelbout, A., 1995. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Medical Decision Making 15, 180–186.
Yaari, M.E., 1987. The dual theory of choice under risk. Econometrica 55, 95–115.