New and robust tests of QALYs when health varies over time

ERASMUS UNIVERSITY ROTTERDAM

Institute of Health Policy and Management

Supervisor: Prof. Dr. Han Bleichrodt
Name: Martin Filko
Student number: 290002
E-mail address: filko@bmg.eur.nl

Abstract

This paper performs new and more robust tests of the QALY model when health varies over time. Our tests require no confounding assumptions and are robust to violations of expected utility. Our results support the use of QALYs at the aggregate level, i.e. in economic evaluations of health care. At the individual level, the support for QALYs is less convincing. The individual data are, however, largely consistent with a more general QALY-type model that remains tractable in applications.

Keywords: QALY, QALYs, utility independence, economic evaluation of health care, decision under risk

JEL classification: I10

1. Introduction

Recently, many methodological improvements have been made to advance the tools of economic evaluation in health care. Nevertheless, some pressing questions remain. Perhaps the most important question that remains unsettled is how the benefits of health care should be valued. The most common approach is to value these benefits in terms of utilities and the utility model that is most widely used is the quality-adjusted life-years (QALY) model. This model makes several simplifying assumptions. For chronic health states these assumptions have been tested experimentally and generally found to be invalid. As an implication, several alternative models have been proposed. Nonetheless, most of the studies assumed expected utility (EU) as a descriptive theory of individual choice. Correcting for some common violations of the EU resulted in support for the QALY model (Doctor et al., 2004), which suggests that violations of expected utility may have confounded other results.

Remarkably, a similar exercise has not been performed for the clinically more realistic case when health varies over time (although chronic health QALYs are a special case of variable health QALY model, therefore violations of the former are also violations of the latter). The few studies that have performed tests of the QALY model when health varied over time made simplifying assumptions which may have confounded their results. The only study that performed an axiomatic test of the QALY model (Spencer 2003a) obtained inconclusive results. As in the chronic case, her tests critically depend on the validity of expected utility, which is known to be invalid as a descriptive theory of decision under risk.

In this paper we examine the preference foundations of the QALY model when health varies over time. Our tests explicitly correct for violations of expected utility and are based on the axiomatization of the variable-health QALYs under general utility model of Bleichrodt and Quiggin (1997). Their model contains EU and many non-EU models as special cases. Importantly, their model includes prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), currently the most popular descriptive theory of decision under risk.

The aim of the paper is to evaluate the descriptive validity of the QALY model when health varies over time, at the aggregate as well as individual level. While aggregate preferences constitute an important input to policy decision making, for example about reimbursement of novel interventions, individual validity of QALYs has important implications for their use as a basis for medical decision making at the point of care and for facilitating informed choice by patients themselves.

The structure of the paper is following. In section 2, we introduce theory and notation behind our experimental tests, based mainly on Bleichrodt and Quiggin's (1997) paper. We review existing – mixed – empirical evidence on the valuation of variable health in section 3. Design of our experiment is described in section 4, while the section 5 presents results and their analysis at the aggregate as well as individual level. The discussion of

the results and possible conclusions which can be drawn from them are presented in sections 6 and 7, respectively.

2. Theory and notation

Let $q = (q_1,\ldots,q_T)$ denote a *health profile* that gives health state q_t at period t. T denotes the last period of life of the decision maker. In our experiment we will only consider health profiles consisting of three periods and, hence, we will take T=3 in what follows. A *binary prospect* (p:q; r) gives health profile q with probability p and health profile r with probability $1-p$. Throughout the paper we will only use prospects involving at most two different outcomes and, consequently, we restrict attention to binary prospects.

A preference relation $\succcurlyeq$ is given over the set of binary prospects. The conventional notation $\succ$ and $\sim$ is used to denote strict preference and indifference. By restricting attention to *constant prospects*, i.e. prospects for which $q = r$ or for which $p = 0$ or $p = 1$, a preference relation over outcomes can be defined, which we also denote by $\succcurlyeq$. It is implicit in the notation (p:q;r) that $q \succcurlyeq r$, i.e. all prospects are *rank-ordered*.

We assume that a prospect (p:q;r) can be evaluated as

$$\pi U(q) + (1-\pi)U(r) \tag{1}$$

and choices and preferences correspond with this evaluation. In Eq.1, π is a decision weight assigned to the outcome that obtains with probability p. This decision weight is entirely general. We will refer to Eq.1 as *general rank-dependent utility* (GRU). Equation 1 is consistent with many theories of decision under risk. For example, if $\pi = p$ then Eq.1 reduces to expected utility and if $\pi = w(p)$, with w a probability weighting function, then Eq.1 reduces to rank-dependent utility and prospect theory for outcomes of the same sign. Eq. 1 was first suggested by Miyamoto, 1988, and was subsequently used by Miyamoto and Wakker, 1996 and Bleichrodt and Quiggin, 1997.

Under the *QALY model* the function $U(\cdot)$ in Eq.1 is additive:

$$U(q) = \sum_{t=1}^{T} V_t(q_t) \tag{2}$$

where the functions V_t can be period-specific. Often a more restrictive QALY model is used where the functions V_t are common for all periods and a constant discount factor is applied to all periods:

$$U(q) = \sum_{t=1}^{T} \delta^{t-1} V(q_t). \tag{3}$$

The focus in this paper is on Eq.2, which captures the essential idea of QALYs of additivity over time. Bleichrodt and Gafni, 1996 showed how Eq.3 can be obtained from Eq.2 by adding one additional preference condition.

Let a_iv_jq denote the health profile q with q_i replaced by a_i and q_j replaced by v_j, $i,j \in \{1,2,3\}$, $i \neq j$. For example, if i =1, j = 2, then $a_iv_jq = (a_1,v_2,q_3)$. Consider the following condition:

Definition 1: the preference relation $\succcurlyeq$ satisfies *generalized marginality* (GM) when for all $i,j \in \{1,2,3\}$, $i \neq j$, and for all health profiles q, health states a,b,c,d,v,w,x,y, and for all p:
$(p:a_iv_jq;\ b_iw_jq) \succcurlyeq (p:c_iv_jq;\ d_iw_jq) \Leftrightarrow (p:a_ix_jq;\ b_iy_jq) \succcurlyeq (p:c_ix_jq;\ d_iy_jq)$.

It is easy to show that under GRU, the QALY model (Eq.2) implies generalized marginality. Let $k \neq i,j$. Under GRU and the QALY model, $(p:a_iv_jq;\ b_iw_jq) \succcurlyeq (p:c_iw_jq;\ d_ix_jq)$ implies that

$$\begin{aligned}\pi\big(V_i(a_i) + V_j(v_j) + V_k(q_k)\big) + (1-\pi)\big(V_i(b_i) + V_j(w_j) + V_k(q_k)\big) \geq \\ \pi\big(V_i(c_i) + V_j(v_j) + V_k(q_k)\big) + (1-\pi)\big(V_i(d_i) + V_j(w_j) + V_k(q_k)\big)\end{aligned} \tag{4a}$$

or

$$\pi V_i(a_1) + (1-\pi)V_i(b_i) \geq \pi V_i(c_i) + (1-\pi)V_i(d_i) \tag{4b}$$

Eq. 4b implies that

$$\begin{aligned}\pi\big(V_i(a_i) + V_j(x_j) + V_k(q_k)\big) + (1-\pi)\big(V_i(b_i) + V_j(y_j) + V_k(q_k)\big) \geq \\ \pi\big(V_i(c_i) + V_j(x_j) + V_k(q_k)\big) + (1-\pi)\big(V_i(d_i) + V_j(y_j) + V_k(q_k)\big)\end{aligned}$$

and thus $(p:a_ix_jq;\ b_iy_jq) \succcurlyeq (p:c_ix_jq;\ d_iy_jq)$.

Bleichrodt and Quiggin (1997, Theorem 4) showed that under GRU, the QALY model not only implies generalized marginality, but generalized marginality also implies the QALY model. Hence, generalized marginality is the central condition of the QALY model.

We next define utility independence. Let J be a subset of {1,2,3} and let q and k be two health profiles. By k_Jq we denote the health profile q with q_j replaced by k_j for all j in J. For example, if J = {1,3} then $k_Jq = (k_1,q_2,k_3)$.

Definition 2: The preference relation $\succcurlyeq$ satisfies *utility independence* (UI) if for all subsets J of {1,2,3}, for all health profiles k,l,m,n,q,r, and for all probabilities p:

$$(p:k_Jq;\ l_Jq) \succcurlyeq (p:m_Jq;\ n_Jq) \Leftrightarrow (p:k_Jr;\ l_Jr) \succcurlyeq (p:m_Jr;\ n_Jr).$$

In other words, time periods during which all health profiles yield the same health state do not affect preferences. Utility independence is less restrictive than generalized marginality: generalized marginality implies utility independence but the reverse is not

true. Miyamoto and Wakker (1996, Theorem 4) showed that if utility independence holds but generalized marginality is violated then

$$U(q) = \prod^{T}_{t=1} V_t(q_t). \quad (5)$$

Equation 5 is still a tractable model. Consequently, not all is lost when generalized marginality is violated. Spencer and Robinson, forthcoming tested for utility independence and found support for it in 6 out of 8 tests. Our tests of utility independence differed in several respects from the tests in Robinson and Spencer as we will explain in Section 4.

Guerrero and Herrero, 2005 further relaxed utility independence and only imposed it for initial health states. They showed that even then a reasonably tractable model results. Their condition is hard to test empirically because it involves dynamic decisions and tests of their model require the comparison of choices made at different points in time. We do not consider their model in this paper.

3. Existing empirical evidence

Empirical evidence provides only mixed support for the QALY model as a descriptive model of choice regarding chronic health states (Tsuchiya and Dolan, 2005). Similar evidence is scarce with respect to the decisions about variable health. In particular, conditions underlying the QALY model for variable health have not been, with a single exception, tested in an axiomatic setting.

Nonetheless, a body of research in the psychology of decision making suggests that people pay attention to some characteristics of sequences that are not readily captured by standard models based on the UI of individual periods. In particular, people care about "improvement and deterioration over time, and peak and end levels", rather than just the sum of time-specific utilities (Ariely and Loewenstein, 2000, p. 508). These so-called sequencing effects have also been documented in health (Ross and Simonson, 1991; Kahneman et al, 1993; Gafni, 1995; Chapman, 2000).

Two types of approaches can be used to evaluate the descriptive validity of the QALY model for variable health. The more common approach compares direct valuation of the health profile with valuation derived from its constituent health states using the additive formula. Assuming zero discounting, additive separability means that the value of a complete health profile would be equal to the sum of the values of its constituent health states, irrespective of the order of the states. If there is nonzero discounting, the value of the health profile depends exclusively on the time-specific weights and values of the health states that make up the health profile.

Nonetheless, the two values may differ not only because the model is violated, but also due to other confounding factors. In particular, it requires specific assumptions about discounting. On the other hand, axiomatic studies test preference foundations of the model directly, i.e. they test necessary and sufficient conditions stated in terms of

observable preferences which hold if and only if a model in question holds. This approach has several advantages. By making its assumptions explicit and behaviorally observable, it allows not only to test the model, but also to identify which of these assumptions are empirically justified. Furthermore, it prevents many of the decision biases due to the fact that all experimental stimuli are similar, i.e. it does not require that both health states and health profiles are used in an experiment. Last but not least, it requires fewer assumptions about specific aspects of individual decision making not related to the model, for example about those related to time preference.

Several studies attempted to determine experimentally whether the sequence of presentation of states in a health profile would affect the valuations assigned to them. In a study by Krabbe and Bonsel (1998), a small effect of the sequence of the tradeoffs was detected at the group level even after accounting for discounting effects. Individual level data suggested that this is due to two groups of respondents who were sensitive to the sequence of events. One group preferred the best years first; the other group preferred the reverse sequence. Nonetheless, the majority of the respondents were indifferent to the sequence.

Richardson and others (1996) interviewed a sample of women who did not have breast cancer to value 4 breast cancer–related health scenarios using VAS, TTO, and SG. Using specific assumptions about discounting, they found that the number of QALYs calculated indirectly from the individual health states was 30% to 50% higher than the number of QALYs calculated from the direct value of the profile.

Kuppermann and others (1997) interviewed pregnant women and let them value (using VAS and SG) 8 health "paths" related to the pregnancy. At the individual level, preferences were not additive, without any pattern emerging from the data. Additivity was also violated at the aggregate level, although it was possible to infer the mean value of the health profiles from individual states under different assumptions. Generally, individual values tend to overvalue the health profiles assessed directly.

MacKeigan and others (1999) presented diabetics with nine health profile covering 30 years and followed by death and asked them to value these by VAS and TTO. They found that the indirect and direct values obtained were not statistically significantly different from one another. On the other hand, the correspondence between the two methods was weak.

According to our knowledge, only a handful of studies performed axiomatic tests of the conditions underlying the QALY model. In a study by Treadwell (2000) testing *preference independence* (a condition stating that if two profiles have the same health state during a certain period, the preference between them does not switch if the health state in that period changes to some other common health state), psychology students were asked to choose between the pairs of health profiles. Each task was accompanied by a similar task in which, should the preference independence hold, subjects should choose a specific profile (i.e. if they chose A over B in one task, they should choose A' over B'

in the other). Although the results were mixed, the independence assumption was more commonly satisfied than it was violated.

Treadwell (1998) tested the preferential independence condition on a sample of 98 subjects. In his first experiment, the condition was satisfied in all 27 tests. Second experiment, designed specifically to be sensitive to violations of preference independence, yielded similar results.

Spencer (2003a) conducted interviews with the sample of 29 residents of York participating in a pilot Health and Safety study. She tested for additive independence in two ways while controlling for risk attitude and time preference. Her results were inconclusive; the first test, using the SG method, rejected additive independence assumption, while the second failed to provide clear-cut results. However, the validity of Spencer tests depend critically upon the validity of expected utility theory, which is known to be a flawed descriptive model of human decision making under risk.

A paper by Spencer and Robinson (2007) tested utility independence in an axiomatic setting. They first conducted 5 tests of utility independence using a standard gamble method on a sample of 64 subjects. Due to the concerns about ordering effects, they conducted a second study in which 3 of the tests were repeated in random order to prevent these effects. In most of the tests, utility independence has been satisfied.

4. Experiment

Participants

Participants in the experiment were students at Erasmus University Rotterdam. They were compensated by a gift certificate worth 10 euros. In total, 60 subjects participated, 30 males and 30 females. The median age of the subjects was 22 years. Three participants had to be excluded from the analysis either for not cooperating, or for apparent use of an extremely simple heuristic (targeting 100% or 0% in all tasks).

Research design

We elicited indifference probabilities between prospects involving health profiles using a standard-gamble method. If the respective condition (UI or GM) being tested holds, the elicited probabilities in the two decision tasks corresponding to the left-hand and right-hand side of formula 3 and 5 should be the same.

Bleichrodt and Quiggin specified the QALYs under uncertainty rather than risk. Although their specification is more general (risk is a special case of uncertainty), it also makes it difficult to manipulate the experimental situation in a way that allows for systematic elicitation of preferences. Therefore, in our experiment decision tasks were presented with specific probabilities, rather than using state-contingent prospects as in

their paper. Probabilities presented in both options of the decision task were the same, to indicate that corresponding states of the world used were identical.

To describe health states, we needed a descriptive system in which it is possible to unequivocally rank health states from best to worse and which is at the same time reasonably realistic. This latter requirement is even more important because our sample consisted of college students, usually with limited experience with more severe health limitations. Instead of using clinically realistic descriptions, we used the generic EuroQol system, which describes states of health along five functional dimensions – mobility, self care, daily activities, pain and anxiety/depression (Dolan, 1997). Outcomes on each of these dimensions can be coded into three levels: no problems, some problems and severe problems. In our study, we used only the former two; our goal was to elicit preferences with regard to moderate health states, which can be imagined more easily by a healthy population.

Four distinct health states were used in the experiment. Each health state was labeled using capital letters from the middle of the alphabet, minimizing potential distortive associations with some other letters (D – death, etc.). The EuroQoL system was introduced in the initial instructions and, throughout the experiment, subjects had the descriptions of the health states available on paper cards in front of them. Health states are summarized in Table 1 and reproduced in Annex 1.

Table 1: Description of health states used in the experiment

Label	Color	EuroQoL code	EuroQoL utility
K	Green	11111	1.000
L	Yellow	11121	0.850
M	Orange	11122	0.722
N	Red	12222	0.551

Due to the fact that we used axiomatization of the QALY model under generalized RDU, it was important to take into account rank-ordering of the prospects. To prevent changes in decision weights due to rank reversals, we maintained rank-ordering of the health profiles throughout the study. Health states were uniquely ordered in terms of utility by using worse or equal rating on each EuroQoL dimension in each of the lexicographically ordered health states (e.g. K was weakly preferred to L on each of the EuroQoL dimensions). Analogously, the ordering of health profiles in a certain decision task was uniquely determined by using worse or equal health states in each of the subsequent health profiles (e.g. KMN was weakly preferred to MMN because in each of the periods, the former allowed to live in the same or better health state than the latter). As a result, the four outcomes (health profiles) satisfied the ordering in Table 2 in each of the choice situations. One prospect (labeled "Option") in a certain task did not dominate the other; otherwise it would not be possible to elicit indifferences.

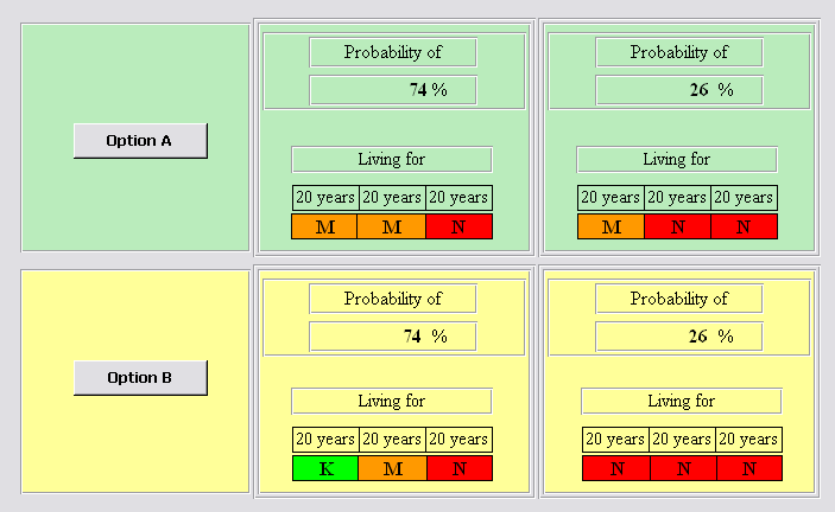
Table 2: Rank-ordering of health profiles in a decision task

	Left column	Right column
Option A	Best option	Worst option
Option B	Second-best option	Second-worst option

Health states were assigned colors (green the best, red the worst, yellow and orange in between) in an order which – based on pilots – was considered intuitive. Although increasing the possibility of invoking idiosyncratic preferences, color-coding was introduced to facilitate decision making in a cognitively demanding situation.

Health profiles were represented by rows of successive boxes labeled by capital letters. Each box represented a health state described on one of the cards. Each health profile consisted of three periods, each lasting for 20 years. A screenshot of a choice situation is in Figure 1.

Figure 1: Screenshot of a decision task



Subjects were asked to make a series of hypothetical choices between two options presented on a computer screen. Each of the options gave the subject a probability to live in a certain health profile, starting today, and a complementary probability to live in a different health profile. Thus, we presented subjects with a standard gamble question with two prospects.

Initial probabilities were determined by chance. Based on a subject's current and immediately preceding choice in a particular decision task, the probabilities were adjusted after each choice according to the *midpoint elicitation procedure*. The procedure consists of a series of binary choices, with probabilities in subsequent steps determined as an average (midpoint) of those presented in the last two steps. Threshold probability difference between two last choices representing indifference was set to 5 percent, i.e. if the difference between the probability to be used in a current step and the one used in the

last was lower than 5, the procedure is terminated and the midpoint between the two probabilities is assumed to represent indifference.

Thus, possible error solely due to the (in)sensitivity of measurement was 8 percentage points (double the error in a single task, because we were comparing indifference probabilities between two corresponding tasks). This might look like a non-negligible difference; however, our pilots suggested that it would be unrealistic to expect subject to have more accurate, consistent cognitive representation of probabilities in decisions tasks of similar complexity. Table 3 presents health profiles that were used in test (decision trees for GM tasks are reprinted in Annex 3). Health states common to all health profiles used in a particular pair of decision tasks are underlined.

Table 3: Health profiles used in GM tasks

		Best outcome	Intermediate outcome	Intermediate outcome	Worst outcome
1	A	KM<u>N</u>	M<u>MN</u>	<u>MN</u>	<u>NN</u>
	B	K<u>KN</u>	<u>MKN</u>	M<u>MN</u>	<u>NMN</u>
2	<u>A</u>	<u>K</u>KM	<u>KL</u>M	<u>KMN</u>	<u>KNN</u>
	<u>B</u>	<u>KKL</u>	<u>KMM</u>	<u>KLL</u>	<u>KNM</u>
3	<u>A</u>	K<u>LL</u>	K<u>LM</u>	<u>LLM</u>	<u>LLN</u>
	<u>B</u>	K<u>LL</u>	K<u>LM</u>	<u>MLM</u>	<u>MLN</u>

UI was tested in a similar manner. Rank-ordered prospects in a standard-gamble setting were based on formula 5. Specific decision tasks used in the experiment are exhibited in Table 4 (again, common outcomes are underlined). Decision trees for UI tasks are reprinted in Annex 4.

Table 4: Health profiles used in UI tasks

		Best outcome	Intermediate outcome	Intermediate outcome	Worst outcome
1	A	LL<u>M</u>	LM<u>M</u>	MM<u>M</u>	MN<u>M</u>
	B	LL<u>N</u>	LM<u>N</u>	MM<u>N</u>	MN<u>N</u>
2	<u>A</u>	<u>KKL</u>	<u>KML</u>	<u>KMN</u>	<u>KNN</u>
	<u>B</u>	<u>L</u>KL	<u>LML</u>	<u>LMN</u>	<u>LNN</u>
3	<u>A</u>	<u>LKN</u>	<u>LMN</u>	<u>LMN</u>	<u>LNN</u>
	<u>B</u>	<u>NKM</u>	<u>NMM</u>	<u>NMM</u>	<u>NNM</u>
4	<u>A</u>	<u>KML</u>	<u>KMM</u>	<u>KMM</u>	<u>KMN</u>
	<u>B</u>	<u>MLL</u>	<u>MLM</u>	<u>MLM</u>	<u>MLN</u>

Procedure

The experiment was computerized; however, both the possibility of misunderstanding the instructions and the level of difficulty of the choice tasks posed significant challenges for individual administration. To prevent unreliable answers, which could have been caused

in the beginning by not understanding the structure of the decision situation properly, and towards the end by the repetitive nature of the tasks, an experimenter was always present. He was not only reading the instructions and explaining the questions, but also operating the computer for the subject. We limited the experimenter's involvement to reading or rephrasing the instructions, and minimized any kind of other communication.

Before data collection, the program was piloted among graduate students at EUR. Improvements aimed at improving cognitive processing of the tasks suggested during the pilot were implemented.

The experiment began with an introduction motivating the experiment, explaining the idea of hypothetical questions in health, showing definitions of health states and giving the opportunity to practice a standard gamble question.

Actual testing consisted of 4 pairs of UI tasks and 3 pairs of GM pairs. The tasks were presented in a random fashion. Initial probabilities of living in different health profiles were also determined at random.

After eliciting indifference, the first choice in the task was repeated as a consistency check. If answered differently than in the beginning, the elicitation process was repeated again. We recorded the number of reversals of the original choices for each of the subjects. If the subject answered consistently, the program moved on to the next task. As a consistency test at the decision-task level, we also repeated one GM question and one UI question (GM1A and UI4B).

The median duration of the experiment was 40 minutes. Furthermore, it consisted of three parts – instructions and practice questions, data collection for the experiment reported here, and data collection for an unrelated experiment in decision making in health. Actual collection of the data for the experiment reported in this paper took approximately 15-20 minutes.

5. Results

Basic descriptives and consistency checks

Basic descriptive statistics of the elicited indifference probabilities can be found in Table 5 and 6. All mean and median values fall in the interval between 53 and 66; interestingly, none of them is bellow 50%.

Table 5: Descriptives for GM tests

	Profile 1	Profile 2	Profile 3	Profile 4	Mean	Median	IQR
GM1A	MMN	MNN	KMN	NNN	53.70	54	32-76
GM1B	MKN	MMN	KKN	NMN	55.53	56	44-67
GM2A	KLM	KMN	KKM	KNN	62.26	61	52-75

GM2B	KLL	KMM	KKL	KNM	61.30	62	48-74
GM3A	KLM	LLM	KLL	LLN	61.19	61	47-74
GM3B	KLM	MLM	KLL	MLN	62.58	61	50-78
Consistency check (GM1A)	KLM	MLM	KLL	MLN	58.74	58	44-70

Table 6: Descriptives for UI tests

	Profile 1	Profile 2	Profile 3	Profile 4	Mean	Median	IQR
UI1A	LLM	MNM	LMM	MMM	59.56	58	51-70
UI1B	LLN	MNN	LMN	MMN	62.75	62	52-76
UI2A	KKL	KNN	KML	KMN	63.40	66	52-74
UI2B	LKL	LNN	LML	LMN	55.91	54	44-70
UI3A	LKN	LNN	LMN	LMN	58.67	61	48-70
UI3B	NKM	NNM	NMM	NMM	55.49	57	39-68
UI4A	KML	KMN	KMM	KMM	64.28	62	54-75
UI4B	MLL	MLN	MLM	MLM	61.39	62	50-72
Consistency check (UI3B)	NKM	NNM	NMM	NMM	62.02	60	50-71

At the aggregate level, consistency appears to be a problem. For the GM task, we rejected the consistency check, using a non-parametric test but not a parametric test (which suggests that the difference was driven by outliers; difference between means was 5.04, medians 4 percentage points). Similarly, consistency check for the UI task was rejected parametrically and non-parametrically at the .05 level, but not non-parametrically at .01 level (difference between means was 6.53, medians 3 percentage points).

Subjects exhibited only moderate tendency to reverse their initial choices (median subject reversed it only once out of 16 tasks, and the mean number of reversals was 1.81 per subject). We recorded reversals of initial choice in 11.29% of the decision tasks (103 reversals out of 16 tasks x 57 subjects = 912 tasks). It suggests that understanding of the tasks was accurate. It also implies that mistakes were rare when probabilities differed significantly from indifference, as was usually the case with initial choices.

Aggregate-level analysis

To perform parametric tests of the difference between the corresponding pairs of tasks at the aggregate level, normality of the distribution of elicited values has to be satisfied. After performing both Skewness/Kurtosis and Shapiro-Wilk W tests for normality (Table 7 and 8), we conclude that with a possible exception of the first GM task, we cannot reject the null hypothesis that distributions are normal. The case of the GM task may appeared by chance due to the fact that we tested 16 questions for significance.Therefore, it was justified to use a t-test for testing the difference of means.[1]

[1] Throughout this section, significance at .05 level is indicated by *, 0.01 level by **.

Table 7: Normality tests, GM tasks

Variable	p(S-K test)		p(S-W test)
GM1A	0.005	**	0.048
GM1B	0.927		0.995
GM2A	0.771		0.751
GM2B	0.328		0.128
GM3A	0.670		0.223
GM3B	0.792		0.204
Consistency check	0.438		0.149

Table 8: Normality tests, UI tasks

Variable	p(S-K test)	p(S-W test)
UI1A	0.806	0.392
UI1B	0.865	0.881
UI2A	0.791	0.742
UI2B	0.836	0.742
UI3A	0.991	0.574
UI4B	0.932	0.932
UI4A	0.798	0.502
UI4B	0.436	0.220
Consistency check	0.347	0.196

GM condition was satisfied at the aggregate level. Using a standard t-test, we were not able to reject at .05 confidence level the null hypothesis that the mean probability is equal between the pairs of tasks. Using non-parametric Wilcoxon signed-rank test yield similar results - we did not reject the condition at the .05 level.

Table 9: Tests of GM at the aggregate level

	Parametric		Non-parametric	
GM1	0.604		0.484	
GM2	0.766		0.429	
GM3	0.629		0.614	
Consistency check	0.064		0.038	*

Testing – less restrictive – UI generally satisfied the condition, with a single exception. In the second UI task, the condition was strongly rejected parametrically and non-parametrically.

Table 10: Tests of UI at the aggregate level

	Parametric		Non-parametric	
UI1	0.237		0.380	
UI2	0.008	**	0.016	*
UI3	0.214		0.369	
UI4	0.239		0.116	
Consistency check	0.005	**	0.011	*

Nonetheless, the general pattern of not rejecting the two conditions may have been caused by limited statistical power of the tests rather than by the fact that the model represents subjects' preferences. Sample size may have been too small, or variation of elicited probabilities too high to reject the hypothesis (confidence intervals were quite wide, Figure 2 and 3).

Figure 2: Median indifference probabilities and confidence intervals

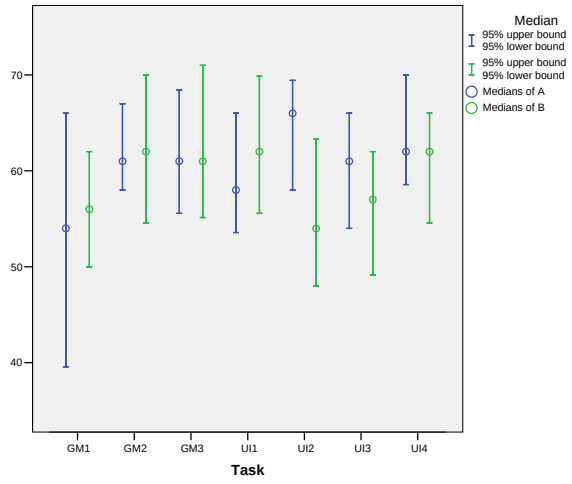
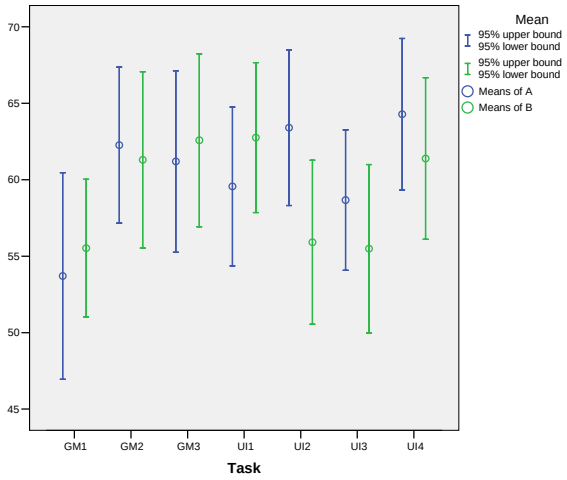


Figure 3: Mean indifference probabilities and confidence intervals



On the other hand, our sample size with a standard deviation of 15 would have enough power to detect a difference in aggregate values of 6 percentage points or higher ($\alpha = .05$, $1 - \beta = .80$). Therefore, not finding many violations of either of the conditions is probably not due to the limited power of our tests.

Furthermore, subjects may have tended to adjust their choices towards the 50% probability, especially in cognitively demanding tasks as those used in our experiment. We performed two tests to verify whether these conjectures could be the case. It is unlikely that targeting the middle of the probability interval was prevalent among the subjects. Mean elicited probabilities for the vast majority of the tasks were significantly different from 50% (Table 11 and 12).

Table 11: Mean difference to 50%, GM

	p-value	
GM1A	0.277	
GM1B	0.017	*
GM2A	0.000	**
GM2B	0.000	**
GM3A	0.000	**
GM3B	0.000	**

Table 12: Mean difference to 50%, UI

	p-value	
UI1A	0.001	**
UI1B	0.000	**
UI2A	0.000	**
UI2B	0.031	*
UI3A	0.000	**
UI3B	0.051	
UI4A	0.000	**
UI4B	0.000	**

This finding is further strengthened by the fact that there was not a single subject whose elicited indifferences always fell between 40% and 60% in either GM or UI tasks.

Table 13: Mean difference between non-corresponding tasks, GM

	GM2A		GM2B		GM3A		GM3B	
GM1A	0.003	**	0.048	*	0.046	*	0.015	*
GM1B	0.007	**	0.037	*	0.088		0.038	*
GM2A					0.754		0.914	
GM2B					0.975		0.735	

Table 14: Mean difference between non-corresponding tasks, UI

	UI2A		UI2B		UI3A		UI3B		UI4A		UI4B	
UI1A	0.174		0.281		0.710		0.205		0.059		0.535	
UI1B	0.783		0.025	*	0.111		0.027	*	0.488		0.626	
UI2A					0.062		0.003	**	0.707		0.492	
UI2B					0.301		0.873		0.005	**	0.059	
UI3A									0.027	*	0.275	
UI3B									0.001	**	0.042	*

In addition, differences in mean elicited probabilities across decision tasks (i.e., between the non-corresponding tasks) were statistically significant in almost half of the cases (Table 13 and 14). Therefore, empirically meaningful differences in elicited probabilities can be detected from our sample.

Individual-level analysis

To evaluate the robustness of the data, as well as to assess the suitability of the QALY model to represent individual preferences, we also conducted an individual-level analysis. Robustness of the model was evaluated by looking at two metrics: magnitude of individual differences in indifference probabilities in the corresponding pairs of the tasks and number of reversals of the first choice, as described above.

Mean individual difference between the pairs was, especially in the UI tasks, not much higher than the sensitivity of the experiment (8 percentage points, Table 15 and 16). This was even more so for median differences, suggesting that individual differences are driven by outliers.

Table 15: Descriptives for individual differences, GM

	Mean	Median	IQR
GM1	20.04	16	6-30
GM2	16.68	13	5-20
GM3	17.63	15	9-24

Table 16: Descriptives for individual differences, UI

	Mean	Median	IQR
UI1	13.65	7	4-18
UI2	14.40	12	4-20
UI3	13.53	8	5-16
UI4	13.39	10	3-18

A question of what constitutes an inconsistency arises. As noted above, we could expect errors up to 8 percentage points purely due to the indifference elicitation procedure. In addition, more errors may have been caused by a "stochastic", "constructed" or

"discovered" nature of people's preferences for unfamiliar choices (Braga and Starmer, 2005).

The frequency of "inconsistencies" depends crucially upon the chosen threshold value above which we call a difference in elicited probabilities an inconsistency. For example, only 7% of the subjects always "satisfied" the GM condition if the threshold was set to 8%, but the proportion rose to 33% if it was set to 18%. It is interesting that although a similar proportion of the subjects always satisfied the condition in both UI and GM tasks, the number of inconsistencies was generally lower in the UI tasks (Table 17 and 18).

Significant proportion of inconsistencies may be interpreted as violating the QALY model. However, it can also reflect difficulty of the question and the fact that preferences over health profiles may be less readily available than those over e.g. monetary gambles with low stakes. Nonetheless, the proportion of individual inconsistencies is comparable to the findings of the studies using simpler, for example monetary stimuli.

Table 17: Frequency and cumulative proportion of individual differences in elicited probabilities under the threshold, GM

	Frequency			Cum. %		
	8<	13<	18<	8<	13<	18<
0	4	12	19	7%	21%	33%
1	12	13	18	28%	44%	65%
2	22	21	15	67%	81%	91%
3	19	11	5	100%	100%	100%

Table 18: Frequency and cumulative proportion of individual differences in elicited probabilities under the threshold, UI

	Frequency			Cum. %		
	8<	13<	18<	8<	13<	18<
0	4	14	19	7%	25%	33%
1	16	21	18	35%	61%	65%
2	23	15	15	75%	88%	91%
3	12	7	5	96%	100%	100%
4	2	0	0	100%	100%	100%

We did not find any systematic differences in the number of inconsistencies among the elicited probabilities within the tasks testing one condition. Nonetheless, individual reversals were far less common among UI than GM tasks (Table 19 and 20). This is not surprising, because the former is a less restrictive condition.

Table 19: Proportion of individual differences in elicited probabilities over the threshold by task, GM

	Frequency			%		
	8<	13<	18<	8<	13<	18<

GM1	39	32	24	68%	56%	42%
GM2	34	26	16	60%	46%	28%
GM3	40	30	23	70%	53%	40%

Table 20: Proportion of individual differences in elicited probabilities over the threshold by task, UI

	Frequency			%		
	8<	13<	18<	8<	13<	18<
UI1	20	17	12	35%	30%	21%
UI2	30	19	15	53%	33%	26%
UI3	26	16	13	46%	28%	23%
UI4	30	20	14	53%	35%	25%

The magnitude of the inconsistencies is rarely correlated across tasks (Table 21). Thus, the results do not suggest individual differences in the ability to consistently answer the type of questions used in the experiment.

Table 21: Pairwise correlation of differences in elicited probabilities across tasks

	GM1		GM2		GM3		UI1	UI2	UI4
GM2	0.48								
p-value	*0.000*	**							
GM3	0.21		0.08						
p-value	*0.115*		*0.568*						
UI1	0.02		0.27		0.21				
p-value	*0.888*		*0.041*	*	*0.123*				
UI2	0.04		0.00		-0.03		0.02		
p-value	*0.758*		*0.987*		*0.840*		*0.906*		
UI3	0.15		-0.04		0.00		-0.08	-0.08	
p-value	*0.268*		*0.762*		*0.996*		*0.578*	*0.535*	
UI4	0.17		0.00		0.38		0.17	0.21	0.02
p-value	*0.203*		*0.989*		*0.003*	**	*0.206*	*0.124*	*0.890*

6. Discussion

Our results support both UI and GM as empirically valid preference conditions of the QALY model. However, a note of caution is important here. Decision tasks used in the experiment are cognitively very demanding. Subjects have to trade off at least three dimensions (quality of life, duration and probability), with potential sequence effects constituting – possible and tested – fourth consideration. Thus, some subjects reported confusion in the practice question, and may have tended to use some simple heuristics later on. Two of these heuristics are especially plausible in our context.

A cognitively difficult decision task under uncertainty may lead to targeting 50%-50% indifference. Indifference elicitation procedure utilizing a midpoint technique ("ping-

ponging" the subject from extreme to more moderate probabilities) might have facilitated this heuristic. However, we were able to rule out this possibility by testing the indifference values against a 50% probability and, in most of the cases, finding significant differences.

More importantly, subjects may have limited the complexity of the situation by eliminating some aspects of the decision tasks. For example, they could have eliminated the common outcome in the UI tasks in order to simplify it. Thus, support for the conditions may have resulted from the heuristic, rather than from genuine preferences.

Although von Neuman-Morgenstern expected utility does not have to hold for our test to be valid, we do not correct for all observed phenomena in human decision making. The validity of the test depends crucially upon the validity of generalized rank-dependent utility theory. If it is not a reasonably accurate description of preferences, results may be confounded in a similar way than when assuming expected utility. Nonetheless, even if that was the case, due to the fact that many of the biases have been corrected for, the extent of confounding is probably much lower than in similar experiments which assumed expected utility.

In addition, a preference reversal specific to health domain known as maximal endurable time (MET, Sutherland et al, 1982; Stalmeier and Bezembinder, 1996; Dolan and Stalmeier, 2003; Spencer, 2003b) could pose a problem. Although its very existence is subject to discussion, we tried to limit its extent by choosing moderate health states, for which it is unlikely to be present. During the experiment, and while debriefing the subjects, none of them expressed any views suggesting that this anomaly may have influenced their choices. Furthermore, the existence of MET would constitute a significant violation of the QALY model, and as such would have been detected by our tests.

Although the conditions tested hold at the aggregate level, one might question the robustness of the model by pointing out the number of reversals at the individual level. However, QALY is usually not aimed at describing individual preferences, but rather at making policy decisions. Thus, satisfying the conditions using aggregate values may be sufficient for these purposes.

7. Conclusion

We have performed the most robust test of the QALY model available in the literature today. Our tests do not require additional confounding assumptions about for example discounting and take account of violations of expected utility. At the aggregate level we observed support for the QALY model as we could not reject generalized marginality, the central condition of the QALY model.

Nonetheless, care should be taken when using the model for decision making at the individual level. Although the subjects exhibited reasonable levels of consistency when

probabilities differed significantly from indifference, most of them exhibited less robustness in their elicited indifference values.

We also tested for utility independence, a less restrictive preference condition than generalized marginality, which still implies a tractable model. Utility independence was supported at the aggregate level. At the individual level we find more support for utility independence than for generalized marginality. For a substantial proportion of our subjects the observed deviations from utility independence can reasonably be attributed to the elicitation procedure and imprecision of preference. Our aggregate findings on utility independence are consistent with the findings of Spencer and Robinson, forthcoming in spite of the differences in experimental design between their and our study. Spencer and Robinson do not report individual-level results.

Our results provide support for the QALY model at the aggregate level. It should be pointed out though that this conclusion is based on three tests only. It should also be kept in mind that we only used mild to moderate health states to avoid considerations like maximal endurable time. Our conclusions may no longer hold when more severe health states are involved. Before QALYs can be safely applied in cost effectiveness analysis, more evidence is needed and we invite other researchers to try and replicate our findings using other experimental designs.

At the individual level, the support for QALYs is much weaker. It appears that QALYs cannot be applied in medical decision making without some additional tests of the decision maker's preference structure. The tests developed in this paper may be helpful in doing so. Even when QALYs are found not to hold, not all is lost. Our results, suggest that there is more support for utility independence at the individual level. Utility independence still implies a tractable model that can be applied in practice. Hence, in contrast with a common belief that QALYs are not consistent with people's preferences for health, the overall message of this paper seems to be supportive of the use of QALY-type models in health economics.

References

Ariely, D., & Loewenstein, G. (2000). The importance of duration in ratings of, and choices between, sequences of outcomes. *Journal of Experimental Psychology: General*, 129 (4), 508-523.

Bleichrodt, H, & Johanesson, M. (1997). The validity of QALYs: an experimental test of constant proportional tradeoff and utility independence. *Medical Decision Making,* 17 (1), 21-32.

Bleichrodt, H, & Quiggin, J. (1997). Characterizing QALYs under a general rank dependent utility model. *Journal of Risk and Uncertainty*, 15, 151-165.

Braga1, J., & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics*, 32 (1), 55-89.

Chapman, G. B. (2000). Preferences for improving and declining sequences of health outcomes. *Journal of Behavioral Decision Making*, 13 (2), 203-218.

Doctor, J. N., Bleichrodt, H., Miyamoto, J., Temkin, N. R., Dikmena, S. (2004). A new and more robust test of QALYs. *Journal of Health Economics,* 23, 353-367.

Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical care*, 35 (11), 1095-1108.

Dolan, P. & Stalmeier P. F. M. (2003). The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics,* 22 (3), 445-458.

Gafni, A. (1995). Time in health: can we measure individuals' pure time preference. *Medical Decision Making,* 15, 31-37.

Temkin, N. R., Dikmena, S., Krabbe, PF, Bonsel GJ (1998). Sequence effects, health profiles, and the QALY model: in search of realistic modeling. *Medical Decision Making,* 18 (2),178-86.

Kahneman, D., Fredrickson, B.L., Schreibner, C.A., Redelmeier, D.A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science,* 4, 401-405.

Kahneman, D., Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.

Kahneman, D., Tversky, A. (2000). Choices, Values, and Frames. Cambridge: Cambridge University Press.

Keeney, R. & Raiffa, H. (1976). Decisions with Multiple Objectives: Preferences and Value Tradeoffs. New York: Wiley.

Krabbe PF, & Bonsel GJ (1998). Sequence effects, health profiles, and the QALY model: in search of realistic modeling. *Medical Decision Making,* 18 (2), 178-186.

Kuppermann, M., Shiboski, S., Feeny, D., Elkin, E. P.,Washington, A.E. (1997). Can preference scores for discrete states be used to derive preference scores for an entire

path of events? An application to prenatal diagnosis. *Medical Decision Making*, 17 (1), 42-55.

Miyamoto, J.M., & Eraker, S.A. (1988). Amultiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology: General*, 117 (1), 3-20.

Miyamoto, J. M., Wakker, P. P., Bleichrodt, H., Peters, H. (1996). The zero condition: a simplifying assumption in QALY measurement. *Management Science,* 44 (6), 839-849.

MacKeigan, L.D., O'Brien, B.J., Oh, P.I. (1999). Holistic versus composite preferences for lifetime treatment sequences for type 2 diabetes. *Medical Decision Making*, 19 (2), 113–121.

McNeil, B.J., Weichselbaum R., Pauker S.G. (1981). Speech and survival: trade offs between quality and quantity of life in laryngeal cancer. *New England Journal of Medicine*, 305, 982–987.

Bala, M. V., Wood, L. L., Zarkin, G. A., Norton, E. C., Gafni, A., O'Brien, B. J. (1999). Are health states "timeless"? the case of the standard gamble method. *Journal of Clinical Epidemiology,* 52 (11), 1047-1053.

Richardson, J., Hall, J., Salkeld, G. (1996). The measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care,* 12(1), 151-162.

Spencer, A. & Robinson, A. (2007). Tests of utility independence when health varies over time. *Journal of Health Economics,* forthcoming.

Ross Jr., W. T., & Simonson, I. (1991). Evaluations of pairs of experiences: a preference for happy endings. *Journal of Behavioral Decision Making,* 4, 155–161.

Spencer, A. (2000). Testing the additive independence in the QALY model. Working Paper No. 427.

Spencer, A. (2003a). A test of the QALY model when health varies over time. *Social Science & Medicine*, 57, 1697-1706.

Spencer, A. (2003b). The TTO method and procedural invariance. *Health Economics*, 8 (9 Suppl. II), II-138-II-50.

Stalmeier, P. F. M., Bezembinder, T. G. G., Unic, I. J. (1996). Proportional heuristics in time tradeoff and conjoint measurement. *Medical Decision Making,* 16(1), 36-44.

Stalmeier, P., Wakker, P. P., Bezembinder, T. (1997). Preference reversals: violations of unidimensional procedure invariance. *Journal of Experimental Psychology: Human Perception and Performance,* 23 (94), 1196-1205.

Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38 (2), 332-382.

Sutherland, H. J., Llewellyn-Thomas H., Boyd, N. F., Till, J. E. (1982). Attitude toward quality of survival: the concept of maximal endurable time. *Medical Decision Making,* 2, 299-309.

Treadwell, J. R. (1998). Tests of preferential independence in the QALY Model. *Medical Decision Making,* 18 (4), 418-428.

Treadwell, J. R., Kearney, D., Davila, M. (2000). Health profile preferences of hepatitis C patients. *Digestive Diseases and Sciences*, 45 (2), 345-50.

Tsuchiya, A., & Dolan, P. (2005): The QALY model and individual preferences for health states and health profiles over time: a systematic review of the literature. *Medical Decision Making,* 25 (4), 460-467.

Tversky, A., Kahneman, D., (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty,* 5, 297-323.

Annexes

Annex 1: Description of health states used in the experiment

K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K

Health state K

In a health state K, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**No pain** or **discomfort**
Anxiety/Depression	**Not anxious** or **depressed**

K K K K K K K K K K K K K K K K K K K K K K K K K K K K K K

L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L

Health state L

In a health state L, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Not anxious** or **depressed**

L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L L

M M M M M M M M M M M M M M M M M M M M M M M M M M

Health state M

In a health state M, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Moderately anxious** or **depressed**

M M M M M M M M M M M M M M M M M M M M M M M M M M

N N N N N N N N N N N N N N N N N N N N N N N N N N N N

Health state N

In a health state N, your health is characterized by:

Mobility	**Some problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**Some problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Moderately anxious** or **depressed**

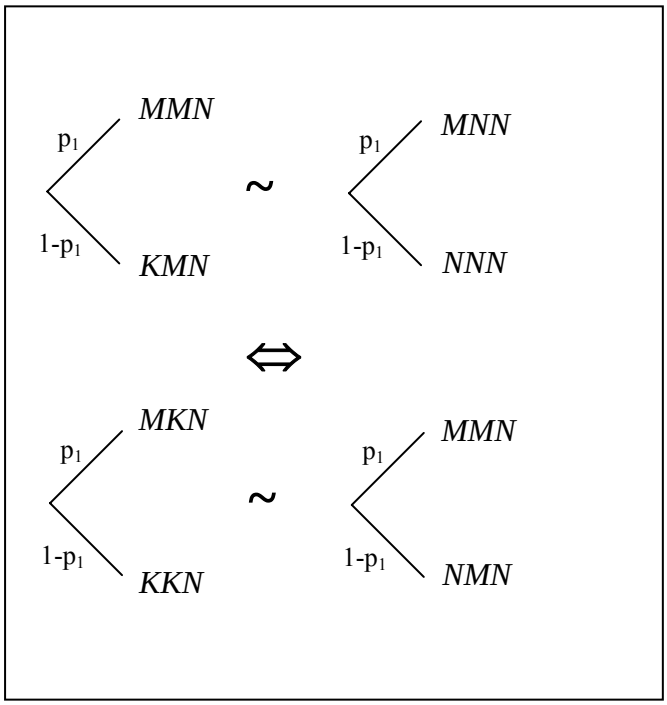
N N N N N N N N N N N N N N N N N N N N N N N N N N N N

Annex 2: Initial ordering of decision tasks

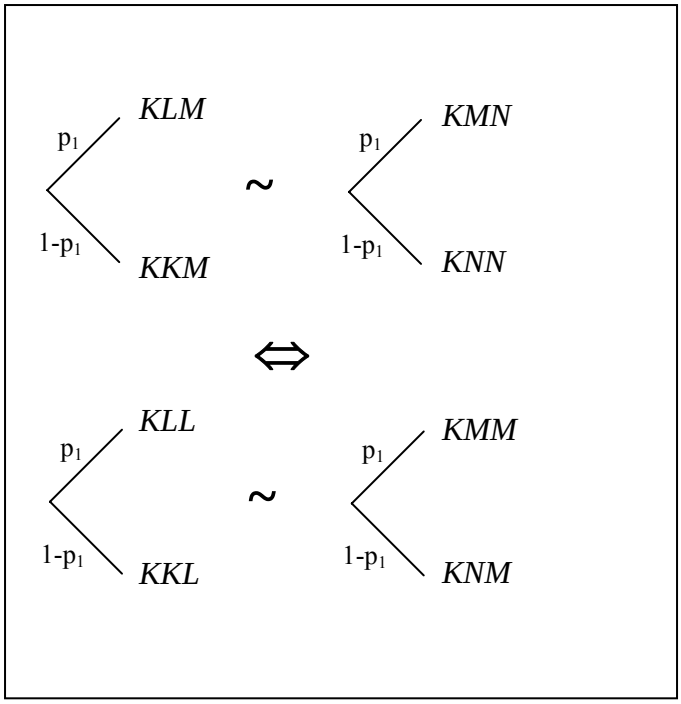
1	GM1A
2	UI5B
3	GM3B
4	UI4A
5	GM2A
6	UI1A
7	GM1A_CC
8	UI4B
9	UI2B
10	GM3A
11	UI5A
12	GM1B
13	UI1B
14	UI4B_CC
15	GM2B
16	UI2A

Annex 3: Decision trees for GM tasks

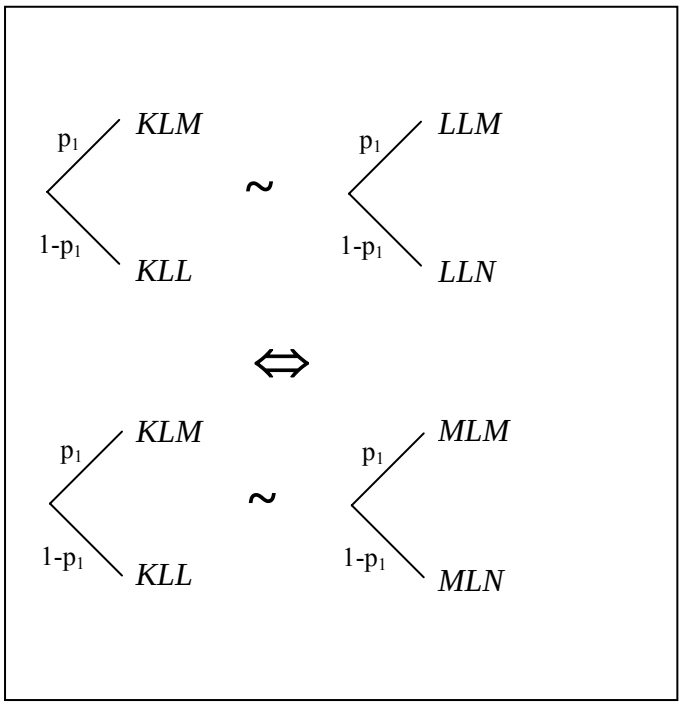
GM1



GM2

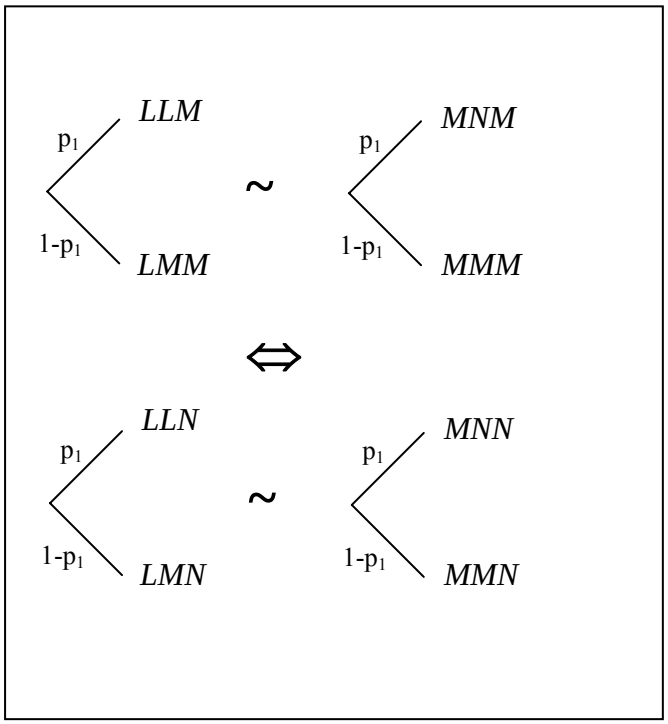


GM3

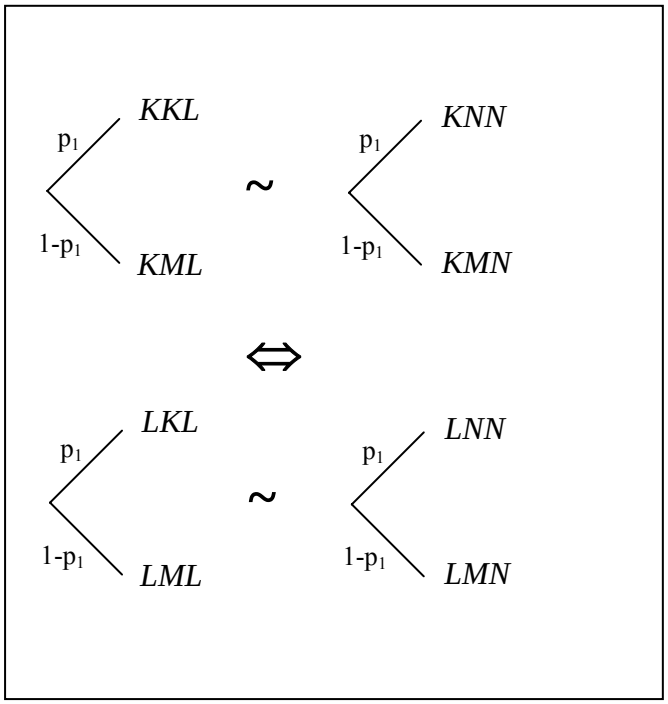


Annex 4: Decision trees for UI tasks

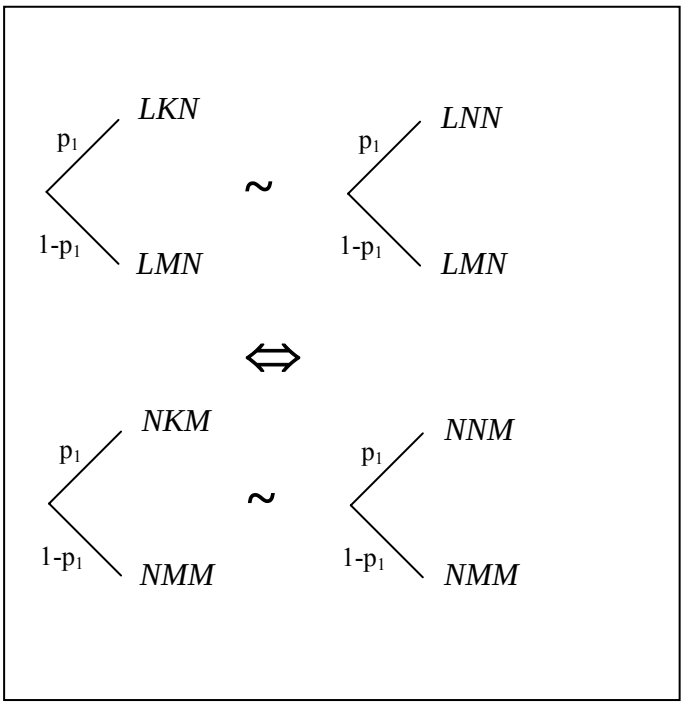
UI1



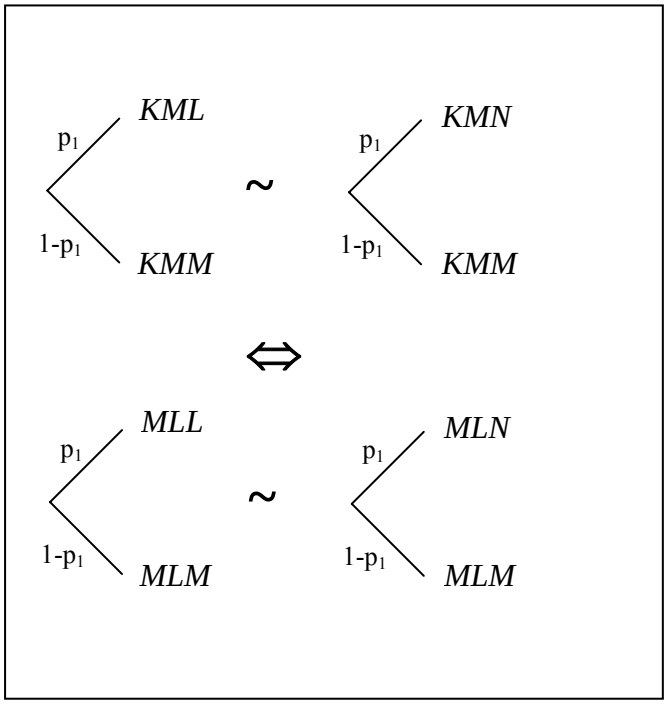
UI2



UI3


LKN
p1
1-p1
LMN
~
LNN
p1
1-p1
LMN
⇔
NKM
p1
1-p1
NMM
~
NNM
p1
1-p1
NMM

UI4


KML
p1
1-p1
KMM
~
KMN
p1
1-p1
KMM
⇔
MLL
p1
1-p1
MLM
~
MLN
p1
1-p1
MLM