A Test of the Predictive Validity of Non-linear QALY Models Using Time Trade-off Utilities

José Mª Abellán Perpiñán, Department of Applied Economics, University of Murcia, Spain.
José Luis Pinto Prades*, Centre for Health Economics and Department of Economics, Pompeu Fabra University, Spain.
Ildefonso Méndez Martínez, Department of Applied Economics, University of Murcia, Spain
Xabier Badía Llach, Health Outcomes Research Europe, Barcelona, Spain

Abstract

This paper presents a test of the predictive validity of various classes of QALY models (*i.e.*, linear, power and exponential models). We first estimated TTO utilities for 43 EQ-5D chronic health states and next these states were embedded in health profiles. The chronic TTO utilities were then used to predict the responses to TTO questions with health profiles. We find that the *power QALY model* clearly outperforms *linear* and *exponential QALY models*. Optimal power coefficient is 0.65. Our results suggest that TTO-based QALY calculations may be biased. This bias can be avoided using a *power QALY model.*

JEL classification: I10, D99.

Key words: Cost-utility analysis, QALYs, Power QALY Model, predictive validity, Time Tradeoff.

* *Corresponding author:* José Luis Pinto Prades, Department of Economics, Pompeu Fabra University, Ramón Trias Fargas 25-27, 08005 Barcelona. Spain. Tel.: 34935422638; Fax: 392521746. E-mail: jose.pinto@econ.upf.es.

Introduction

It is current practice in economic evaluation of health care programs to use *quality-adjusted life years* (*QALYs*) as a measure of the utility of health outcomes. A QALY is basically a weighting scheme in which each period of life duration (e.g., years or a fraction of a year) is adjusted by the health state in which it is spent. According to this definition what is essential to QALYs is that life duration and health status are separable. Due to this reason some authors refer to basic classes of QALY models as *multiplicative* utility models (e.g., Miyamoto and Eraker, 1988; Miyamoto, 1999). That is, QALY models in which duration and health status are combined in a multiplicative way. Formally, if (Q, T) denotes a chronic health state, where Q stands for constant health status throughout T periods of time until dead, then the multiplicative QALY model asserts that

$$U(Q, T) = H(Q) \cdot G(T), \tag{1}$$

where H and G are utility functions over health status and duration, respectively. It is further assumed that G(·) is increasing in duration and G(0) = 0.

However, whenever publications refer to the QALY model, authors usually mean the *linear QALY model*, that is, the specific multiplicative utility model characterized by assuming that the utility of the life duration is linear. Despite of QALYs are often discounted for practical proposes the linear QALY model remains as the simplest and most widely used QALY model. If G(T) is equal to T in Equation 1, then we obtain the linear QALY model

$$U(Q, T) = H(Q) \cdot T, \tag{2}$$

As various researchers have remarked (e.g., Wakker, 1993; Miyamoto, 2000) the linear QALY model greatly simplifies utility calculations, because fewer measurements

are needed to calculate the utilities of sequences of health states (henceforth: health profiles). Once the validity of the linear QALY model is assumed, practitioners only need to estimate utilities for chronic health states, and then by assuming that the utility of the health profile is additive over separate (or disjoint) time periods (*i.e.*, each period during which health status does not vary) the number of QALYs follows. Of course, the utility of health profiles can be also computed in a similar way by using non-linear QALY models (e.g., Miyamoto, 1999), but it requires as an additional task to fit some parametric utility function for duration to data (e.g., Miyamoto and Eraker, 1985; Sttigelbout et al., 1994) or, alternatively, to apply a group estimate of the parameter that is regarded as a good approximation (e.g., Miyamoto, 2000).

Let us to explain how Equations (1) and (2) can be extended to health profiles. Let $(Q_1, T_1; Q_2, T_2)$ stand for a health profile where state Q_1 lasts for T_1 periods, and state Q_2 lasts for T_2 periods, followed by death. Next, for simplicity, assume that both Q_1 and Q_2 are better-than-death health states (*i.e.*, states in which longer duration is preferred to shorter duration) and also assume that G(.) is concave. Figure 1 (bottom) shows concave utility functions for life duration in chronic health states Q_1 and Q_2. To apply Equation 1 to $(Q_1, T_1; Q_2, T_2)$ we have to assume that $U(Q_1, T_1; Q_2, T_2)$ is additive over separate periods T_1 and T_2. Under this assumption, the upper part of Figure 1 shows how the segments of the chronic health state utility functions are combined by adding the utility increments $H(Q_1) \cdot \{G(T_1) - G(0)\}$ and $H(Q_2) \cdot \{G(T_1 + T_2) - G(T_1)\}$ during the separate periods T_1 and T_2. The height of the point *a* denotes the utility of $H(Q_1) \cdot G(T_1)$, since $G(0) = 0$. The height of the point *b* denotes the overall utility of the health profile computed as the sum of the abovementioned utility increments. Hence, under the additivity assumption, the multiplicative QALY model (2) implies that

$$U(Q_1, T_1; Q_2, T_2) = H(Q_1) \cdot G(T_1) + H(Q_2) \cdot \{G(T_1 + T_2) - G(T_1)\} \qquad (3)$$

If we extent this example to the more general case of *n-tuples* (Q_1, T_1; Q_2, T_2; ...; Q_n, T_n) where duration ranges from *1* to *n* time periods, the multiplicative QALY model will be defined by (Miyamoto, 1999: p. 206, Eq. 6):

$$U(Q_1, T_1; Q_2, T_2; \ldots; Q_n, T_n) = \sum_{t=1}^{n} H(Q_t) \cdot \left\{ G\left(\sum_{i=0}^{t} T_i \right) - G\left(\sum_{i=0}^{t-1} T_i \right) \right\} \qquad (4)$$

We note that although Equation (4) may look complicated, it simply describes algebraically the normal process of construction of the utility of a health profile when the utility function for duration is not linear. For example, suppose that we want computing the utility of health profile (Q_1, 3 yrs; Q_2, 5 yrs) and we discount each year by applying a non-linear utility function. We can then elicit $H(Q_1)$ and $H(Q_2)$ and then computing the utility of the entire profile as $H(Q_1) \cdot G(1) + H(Q_1) \cdot G(2) + H(Q_1) \cdot G(3) + H(Q_2) \cdot G(4) + \ldots + H(Q_2) \cdot G(8)$, where 1 stands for "first year", 2 stands for "second year" and so on. Alternatively, we can compute the utility of the profile according to Equation (4), that is, $H(A) \cdot G(3) + H(B) \cdot \{G(8) - G(3)\}$, where 3 and 8 denote three and eight years respectively. As it is easy of checking, both ways of calculating QALYs are equivalent.

It is straightforward that if we apply the linear QALY model (1) to any arbitrary health profile then Equation (4) is reduced to the standard formula

$$U(Q_1, T_1; Q_2, T_2; \ldots; Q_n, T_n) = \sum_{t=1}^{n} H(Q_t) \cdot T_t \qquad (5)$$

In sum, Equations (1) and (5) characterize the linear QALY model, whereas Equations (2) and (4) characterize the multiplicative QALY model.

Empirical evidence on the validity of the linear QALY model is mixed. Most empirical tests of the assumptions underlying the linear QALY model for chronic health states have generally yielded negative results (*e.g.*, McNeil et al. 1978; Pliskin et al. 1980; Miyamoto and Eraker 1985; Stiggelbout et al. 1994; Verhoef et al. 1994; Wakker and Deneffe 1996; Stalmeier and Bezembinder 1999). In particular, evidence available hitherto supports that the utility function for duration is concave rather than linear (Dolan, 2000).

However, the majority of the previous tests were performed within the realm of Expected Utility (EU). The exceptions are the tests performed by Bleichrodt and Pinto (2001) and Doctor et al. (2003). Whereas the former rejected the linear QALY model within a framework consistent with Rank-Dependent Utility (RDU), the latter found considerable support for the linear QALY model within a more general framework compatible even with Prospect Theory (PT). Hence, as Doctor et al. argue, it is possible that the observed violations of the linear QALY model might have been caused by violations of EU.

Empirical tests of the linear QALY model for health profiles are not conclusive either. Treadwell (1998) and Spencer (2001) found empirical support for the assumption of additivity over time periods. However, evidence on sequencing and discounting effects (*e.g.*, Krabbe and Bonsel 1998; MacKeigan et al. 2002) challenges the validity of the linear QALY model.

This study aims to test the validity of the linear QALY model in a different way from that typically followed in most previous empirical studies. Instead of testing any critical assumption to the linear model we perform a test of its predictive validity. With predictive validity we mean to test whether preferences over health profiles can be accurately predicted from utilities for chronic health states. This approach has the

advantage that it does not depend on a specific utility framework (*e.g.*, EU or PT), because we are testing a prediction common to any multiplicative QALY model.

Although there are some other studies that have also tested predictions of the linear QALY model they have followed a different approach. Most predictive tests have compared the number of QALYs yielded by a health profile to some sort of holistic assessment for the same profile. These studies have generally found large discrepancies between QALYs and holistic assessments (*e.g.*, Richardson et al., 1989; Lipscomb, 1989; Kupperman et al., 1997). However, discrepancies can only interpreted as failures of the linear QALY model as long as holistic assessments are accepted as a norm or gold standard.

In this respect, our approach is closer to that followed by Bleichrodt and Johannesson (1997) and Bleichrodt et al. (1999). Bleichrodt and Johannesson compared the predictive validity of three elicitation methods (Standard Gamble, Time Trade-Off, and Visual Analogue Scale) under EU both for linear and discounted QALY models. Predictive validity was assessed by testing the degree of association between direct and predicted rankings of health profiles. Bleichrodt and Johannesson found that the combination of TTO utilities and a null discount rate had the highest predictive validity. In a similar way, Bleichrodt et al. (1999) tested the validity of the SG method with and without rank-dependent weighting. They also examined linear/exponential and log/power families of utility functions for duration. They found that utility curvature improved the predictions of the multiplicative QALY model.

This study also try to go beyond simply testing the performance of the linear QALY model. We try to shed some light on each of four related questions, namely:

a. *Which is the non-linear QALY model with a higher predictive validity.*

We test the predictive validity of two non-linear QALY models, namely: the *power QALY model* and the *exponential QALY model*. We have chose these two models because they have been frequently used in axiomatic work on QALYs both under EU (e.g., Pliskin et al., 1980; Miyamoto et al., 1998; Miyamoto, 1999) and non-expected utility (e.g., Miyamoto, 1988; Miyamoto and Eraker, 1988; Bleichrodt and Pinto, 2001). Also, power and exponential specifications have been widely employed in medical decision analysis (*e.g.*, Pauker, 1976; Mass and Wakker, 1994; Cher et al., 1997; Enemark et al., 1998). We will choose as the better QALY model that one with higher predictive validity. To that end we will estimate the parameter values that most improves the predictive validity of the two non-linear QALY models selected.

Bleichrodt et al. (1999) also tested the predictive validity of the power and exponential models, but they did not estimate the optimal parameters of these models. They gave different values to power and exponential parameters and selected the value that better predicted preferences.

b. *The influence of utility curvature on a wide range of health states*.

It has not been tested if the potential improvement of the non-linear over the linear QALY model is constant for all degrees of severity. To test this we use a set of 43 EQ-5D health states. Bleichrodt et al. (1999) only used two health states (*i.e.*, full health and an intermediate health state). In other papers researchers used three health states at best (*e.g.*, Richardson et al., 1996; McKeigan et al., 1999; Bleichrodt and Pinto, 2001).

c. *The influence of utility curvature on Time Trade-Off utilities*.

The elicitation method selected to perform our test is the TTO. We choose this method because there is some evidence that the TTO is more consistent with individual preferences than the SG (*e.g.*, Dolan et al., 1996; Dolan, 2000). Indeed, as noted above, Bleichrodt and Johannesson (1997) found that the TTO without discounting yielded the highest predictive validity. However, it is also true that various researchers have warned about the risk that TTO utilities are biased downwards if utility for duration is concave (*e.g.*, Miyamoto and Eraker, 1985; Johannesson et al., 1994; Dolan and Jones-Lee, 1997). In consequence, we think that it is important to examine the influence of utility curvature over the consistency of TTO utilities. For example, if TTO utilities were biased by utility curvature the widely employed *EuroQol* algorithm (TTO-based) might be also biased, leading to wrong allocations of health resources.

d. *The structure of preferences on a wide sample of the general population*.

Most of previous empirical papers were based on small convenience samples (*e.g.*, Dolan and Gudex, 1995; Kupperman et al., 1997; MacKeigan et al., 1999). However, it is a common view in Cost-Effectiveness Analysis (CEA) that health state utilities should be collected from a representative sample of general population (Gold et al. 1996). We think that it is very important for policy decisions to check if the potential superiority of the non-linear over the linear QALY model also holds when a representative sample of the general population is used. In consequence, we elicit preferences from a relatively large sample (nearly 1,300 people) of the general (Spanish) population.

The paper is structured as follows. In Section 2 we describe the test of the linear QALY model that we are going to perform. In a first survey, we estimate TTO utilities for 43 health states described to the respondents as chronic conditions. Next, in a second survey, the same collection of health states are embedded in health profiles which combine the specified health state with full health. The TTO utilities estimated in a first survey are then used to predict the responses to TTO questions used in the second survey. We emphasize that respondents were randomly drawn from a single population. Hence, our null hypothesis assumes that, up to sampling error, the linear QALY model leads to the right prediction. This implies that no significant difference is found between predicted and observed responses. Otherwise, if the null hypothesis is rejected we would next examine whether two alternative single-parametric utility specifications (*i.e.*, power and exponential models) improve the prediction. This alternative hypothesis is described in Section 3. In Section 4 we present the study design and the statistical methods to be used. Section 5 shows the results. Conclusion closes the paper.

2. Predictive validity of QALYs

First, assume that the set of chronic health states may contain both *better-than-death* and *worse-than-death* states. Next, assume that the linear QALY model (2) represents the individual's true preferences over the overall set of chronic health states. We note that the existence of better-than-death and worse-than-death health states is not in conflict with the linear QALY model (nor with the multiplicative QALY model). In fact, as various authors have emphasized (*e.g.*, Miyamoto and Eraker, 1988; Miyamoto et al., 1998; Miyamoto, 1999) the existence of better-than-death and worse-than-death states is a diagnostic of a multiplicative relationship.

Assume now that the TTO is used to elicit the utility of health states Q^I and Q^W defined as chronic conditions (Torrance, 1986). Assume that if any of the health states was regarded as worse than death its TTO utility would be set equal to – Y/(T – Y), where T denotes the duration in the state Q^I and Y stands for the duration in Full Health (FH). Hence, we do not rescale negative utilities so that they range from -1 to 1. This is an arbitrary transformation after which valuations cannot longer interpreted as true utilities (Patrick et al., 1994).

Next, assume that we have two health profiles such as (FH_1, T_1; Q_2^W, T_2; FH_3, T_3) and (Q_1^I, T_1 ; Q_2^I, T_2; FH, T_3) followed by death, where T_i denotes duration. Since health state Q^I is the same during periods T_1 and T_2, we can represent the profiles as (FH_1, T_1; Q^W, T_2; FH_3, T_3) and (Q^I, T_{1+2}; FH_3, T_3) where T_{1+2} stands for T_1 +T_2. Finally, assume that the linear QALY model (5) describes correctly the preferences over health profiles and that the individual is indifferent between the two profiles. Let H(FH) =1 and H(death) =0. Then Equation (5) yields

$$T_1 + H(Q^W) \cdot T_2 = H(Q^I) \cdot T_{1+2} \qquad (6)$$

One way of testing the predictive validity of the linear QALY model is fixing two of the three durations (*i.e.*, T_1, T_2, T_{1+2}) and eliciting the third one. The elicited T_i should coincide with the estimate (denoted by $\hat{T}_i$) we get from Equation (6). Thus the null hypothesis of the test asserts that there is no significant difference between predicted and observed durations. On the contrary, our alternative hypothesis asserts that non-linear QALY models will describe preferences better than the linear one. If non-linear QALY models outperformed the linear one we would expect that taking into account utility curvature could partly remove the discrepancy between predictions and

actual data. Indeed, as noted in introduction, various authors have proposed previously to adjust TTO measurements for utility curvature. Hence, if our test rejects the linear QALY model, we will then test the degree of improvement in predictive validity of the power and exponential QALY models. The better they describe preferences the lower the difference between $\hat{T}_i$ and T_i.

3. Power and exponential QALY models

We will compare two classes of non-linear QALY models, namely: the *exponential QALY model* and the *power QALY model*[1]. In both cases the utility function over duration is described by a single parameter λ. The power specification implies that *proportional time tradeoffs* are *constant* (Pliskin et al., 1980) and, under EU, the power parameter may encapsulate attitude towards risk, time preference, and diminishing marginal value (Gafni and Torrance, 1984).

The exponential function satisfies *constant absolute* (rather than proportional) *tradeoffs* (Happich, 2001) and it reduces to the constant discounting model if time is discrete. Derivations of TTO utilities after adjusting by utility curvature are provided in Appendix 1.

CASE 1. The *exponential QALY model* asserts that:

$$U(Q, T) = \mathrm{H(Q)} \cdot \mathrm{k}\left(1 - \mathrm{e}^{-\mathrm{u}}\right) \qquad (7)$$

[1] See Miyamoto (1999) for a characterization of these models under EU and under RDU assumptions. Bleichrodt and Miyamoto (2002) present axiomatizations for these models under Cumulative Prospect Theory.

where coefficient k is a scaling constant equal to $\frac{T*}{(1-e^{-\lambda T*})}$ and T* is the maximum lifetime duration in the given domain. Within the range from 0 to T*, the exponential function $k \cdot (1\text{-}e^{-\lambda T})$ can be concave ($\lambda > 0$) as well as convex ($\lambda < 0$).

If we assume the validity of the *exponential QALY model* (7), under the assumption of additivity over disjoint time periods, Equation (6) changes into

$$k\left(1-e^{-\lambda T_1}\right)+H*\left(Q^W\right)\cdot\left\{T_{1+2}-k\left(1-e^{-\lambda T_1}\right)\right\}=H*\left(Q^I\right)\cdot T_{1+2} \quad (8)$$

where asterisk * denotes that TTO utility H has been adjusted by λ.

CASE 2. The *power QALY model* asserts that:

$$U(Q, T) = H(Q)\cdot k\,T^{\lambda} \quad (9)$$

where k is a scaling constant equal to $\frac{T*}{(T*)^{\lambda}}$ and T* is the longest duration in the given domain. Within the range from 0 to T*, the power function kT^{λ} can be concave ($\lambda < 1$) as well as convex ($\lambda > 1$).

If we assume the validity of the *power QALY model* (9), under the assumption of additivity over disjoint time periods, Equation (6)changes into

$$k\,T_1^{\lambda}+H*\left(Q^W\right)\cdot\left\{T_{1+2}-k\,T_1^{\lambda}\right\}=H*\left(Q^I\right)\cdot T_{1+2} \quad (10)$$

where asterisk * denotes that TTO utility H has been adjusted by λ.

4. Methods

- *Subjects*

We conducted two surveys in order to perform the test. In a first survey, 43 EQ-5D health states were valued as chronic conditions by 977 respondents (sample 1). In a second survey, 300 respondents (sample 2) the same health states were embedded in

different health profiles.

The surveys were carried out by 11 interviewers over a 6-month period, following a 2-day training period. The two groups of respondents were randomly selected from Spanish general population. Age and gender quotas were used to ensure representativeness on these parameters according to the 1991 Spanish census. Potential respondents were contacted initially by letter, and then by follow up telephone calls.

Respondents who were unable to read or write, or who were cognitively impaired (according to the Pfeiffer test), or who had a severe illness or mental disorder, were replaced by others in the same sex-age quota. Background data, health expectations, and opinions regarding the interview were collected at the end of the interview.

- *Health states*

The subset of EQ-5D health states selected is the same that Dolan (1997) used to model the EuroQol algorithm. In the first survey, each respondent assessed a random selection of 13 health states including 2 very mild states, 3 mild states, 3 moderate states, 3 severe states, and the states '11111' (*i.e.*, full health) and '33333'. In the second survey, respondents did not value states '11111' and '33333' because of they were used as reference health states.

[Insert Table 1 about here]

- *Elicitation procedure*

Respondents belonging to sample 1 first described and rated their own health state using the EQ-5D descriptive system and the VAS method with endpoints 0 –100 of worst and best imaginable health state respectively. They then ranked their selection

of 13 health states (plus unconsciousness, but excluding 'death') in order of preference. Respondents were asked to imagine that each health state would last for 10 years without change, followed by death. After they ranked the health states, they were asked to rank 'death' among those states, and were given the option at this point of reordering their previous ranking.

Utilities were elicited by means of the TTO method for chronic health states (Torrance, 1986). For health states regarded as better than death, we asked for the duration Y that yields indifference between surviving 10 years in the target health state and surviving Y years in health state '11111'. For states regarded as worse than death, we asked for the duration Y that leads to the indifference between dead and surviving (10- Y) years in the target state followed by Y years in '11111'. The order in which each respondent ranked and valued his/her selection was randomized to avoid anchoring and adjustment biases.

Respondents belonging to sample 2 first ranked their selection of 13 health states plus death and unconsciousness. Then they were asked two compare two health profiles such as (FH, T_1; 33333, T_2; FH, T_3) and (Q^I, T_{1+2}; FH, T_3) both followed by death. It was set that $T_1 + T_2 = T_{1+2} = 12$ months, and $T_3 = 9$ years. We then elicited T_I and compared it with $\hat{T}_I$ obtained using Equation (6), that is

$$\hat{T}_I = \frac{H(Q^I) - H(33333)}{1 - H(33333)} x\ 12 \qquad (11)$$

Empirical evidence supports that preferences inferred from choices are more consistent than preferences inferred from matching or judgments of selling prices (*e.g.*, Bostic et al., 1990). In this way, all chronic health states and health profiles were

presented to respondents as choices. Indifference points in both samples were stated at a level of proportions of months in order to ensure that precision in responses were similar across samples. For example, questions with health profiles were first presented as a choice between a profile with 6 months in health state '11111' and 6 months in state '33333' followed by 9 years in '11111' and a profile with 12 months in the target health state followed by 9 years in '11111' as well. After the choice was described respondents were asked whether they preferred the first or the second profile, or whether they were indifferent between both. In case the respondent preferred either the first or the second health profile, the interviewer next varied duration in health states '11111' and '33333' until indifference was reached. Throughout this choice-bracketing exercise respondents were allowed to revise earlier answers and, in order to avoid response errors, they were asked to confirm the elicited indifference values.

- *Estimation methods*

Differences between predicted and observed durations obtained under the linear QALY model are tested by the *two-sample t* test. If significant differences were found, we would then estimate the optimal value (*i.e.*, $\hat{\lambda}$) of power and exponential coefficients of the QALY models described in Section 3.

The procedure we would employ in order to estimate the optimal coefficients would be an optimization algorithm based on the *Newton-Raphson* method (Greene, 1999). We remark that this procedure is not a regression analysis. In the context of this study, goodness of fit means minimizing the differences between responses constructed from sample 1 and observed responses from sample 2.

The specific procedure is as follows. We start setting a utility function for duration with λ=0 in the exponential model, and with λ=1 in the power model. That is,

we fix a linear utility function. Next, by varying λ for each respondent in sample 1, we obtain a value of the parameter such that is minimized the sum of squares over the differences between the mean predicted responses $\overline{\hat{T}_i}$ and the mean observed responses $\overline{T}_i$ across 41 EQ-5D health states[2]. Therefore, the objective function to be used in estimations (see below, Equation 12) would be a summary obtained from individual data. As Miyamoto (2000) has argued, the aggregation across responses in order to yield the fit may be a more accurate representation of preference than are the individual data because the individual points almost always exhibit greater random variation than a summary constructed from the data. From this perspective, since the responses to any elicitation method (and the TTO is not an exception) are not free from random variation, the overall minimization of differences across responses and health states may be one way to obtain a more accurate approximation to true preferences.

In sum, we would find the optimal estimate $\hat{\lambda}$ of the parameter λ that minimizes

$$\left[\sum_{i=1}^{41}\left(\overline{\hat{T}}_i(\lambda)-\overline{T}_i\right)^2\right], \qquad (12)$$

with $\overline{\hat{T}}_i(\lambda)=\frac{1}{J}\sum_{j=1}^{J}\overline{\hat{T}}_{ij}(\lambda)$, where subscript *i* denotes the health state and subscript *j* denotes the respondent belonging to sample 1 valuing state *i*.

5. Results

Table 2 compares the sociodemographic characteristics of both samples. The two samples were very similar in terms of gender and age (Chi-square, P = 0.97 and

[2] Health states '11111' and '33333' were used as reference states in the questions with health profiles, so they were not included in estimations.

0.93 respectively), although there existed significant differences in educational and employment status (Chi-square, P < 0.01 in both comparisons).

[Insert Table 2 about here]

We excluded 54 subjects from the data analysis of the TTO utilities because they did not assign the lowest value to health state `33333´. Four participants in the second survey were also excluded from the data analysis because they did not want to make some tradeoffs. Hence, the final analysis is based on the responses of 923 and 296 subjects, respectively.

[Insert Table 3 about here]

Table 3 shows that the null hypothesis cannot be rejected for 7 of the 41 states displayed (they have been ranked in increasing severity order according to chronic utilities). Thus, differences between $\overline{\hat{T}}_i$ and $\bar{T}_i$ are statistically significant for almost all health states. The largest differences were found for severe states. The maximum difference (= 3.399, *i.e.*, three months and twelve days) was observed for the *worse-than-death* state '13332'.

It is noticeable that $\overline{\hat{T}}_i$ was lower than $\bar{T}_i$ for the majority of the states. Only for three very mild states there is a positive difference and it was not statistically significant in two of them. Moreover, differences increase with severity, although there are variations in the final part of the domain of the health states.

The discrepancy between $\overline{\hat{T}_i}$ and $\bar{T}_i$ was minimized under exponential and power models for $\hat{\lambda} = 0.48$ and 0.65 respectively. These estimates imply a concave-shaped utility function for duration.

[Insert Figure 2 about here]

Figure 2 displays clearly that the power QALY model outperforms the rest. Under this model differences were not significant in 32 out of the 41 health states used in the comparison. In those health states where differences were greater than zero they were heavily reduced.

6. Discussion

The main findings and implications derived from the present study are summarized as follows:

1. We find significant differences for almost all the health states under the *linear QALY model*. This result suggests that the linear model is not a good descriptive model.
2. It seems that the linear QALY model is most likely to hold for mild health states. It is for severe health states when the deviation seems larger. This result is consistent with other findings reported elsewhere in the literature (*e.g.*, Sutherland et al., 1982; Kirsch and McGuire, 2000).
3. The *power QALY model* had the highest predictive validity. It removed differences between observed and predicted responses for the majority of health states (around 80%). This functional form has been frequently used in the decision-theory literature (*e.g.*, Tversky and Kahneman, 1992) and it has been also proposed as a good instrument for medical decision analysis (*e.g.*,

Stiggelbout et al., 1994; Bleichrodt and Pinto, 2000). In addition, and opposite to exponential models, the power model satisfies that TTO measurements remain constant with independence of the reference duration fixed throughout the assessment task (Pliskin et al., 1980).

4. The parameter estimated ($\hat{\lambda}$=0.65) indicates that the shape of the utility function for life duration is concave rather than linear. Under EU concavity of the utility function for duration can reflect risk aversion, positive time preference and diminishing marginal value (Gafni and Torrance, 1984). Under other non-EU theories, like rank-dependent utility, concavity would just display time preference and diminishing marginal utility because risk attitude is reflected in the probability weighting function (Wakker, 1994). Since TTO measurements are riskless, in our study concavity cannot be interpreted under EU as reflecting risk aversion. Hence that concavity just displays 'time' and 'quantity' effects. Obviously, as the magnitude of the lifetime duration is confounded with the timing of the health outcomes, separating time preference from quantity effects is problematic (*e.g.*, Chapman, 1997).
5. The finding of concavity is in line with those empirical studies that estimated the utility function by a power function (*e.g.*, Miyamoto and Eraker, 1985; Stiggelbout et al., 1994; Bleichrodt and Pinto, 2000; Bleichrodt and Pinto ,2001). We note, however, that our parameter has a value somewhat lower than previous studies (reported mean λ ranges from 0.74 to 1.03). Nevertheless, we emphasize that those studies are different from ours. First, previous studies used one or, at best, three health states while we have used 43 different health states. Second, we have also used a larger *sample-size* and

estimated a summary parameter from individual data adjusting simultaneously all the health states. Lastly, experiments conducted by Bleichrodt and Pinto tested the power function under nonexpected utility, in which case it was expected a lower curvature (*i.e.*, a higher value).

6. Discounting is a standard practice in economic evaluation of health care. However, our results cast doubt on the conventional approach of applying an exponential factor in order to discount TTO-based QALYs. We find that the exponential model works only slightly better than the linear model. Indeed, the discrepancy between predictions and observed data remains quite large (it only vanishes for ten health states) even after TTO utilities are adjusted by utility curvature as Johannesson et al. (1994) recommended. Our finding is similar to results reported by MacKeigan et al. (2002). They also adjusted TTO utilities, and found a large difference between discounted TTO-based QALYs and holistic utilities.
7. TTO utilities estimated under linearity may be biased. This implies that the *EuroQol* algorithm, which is based on conventional TTO valuations, may lead to wrong allocation of resources. We have obtained empirical support for adjusting TTO utilities by concavity in order to correct the utility curvature bias. However, Bleichrodt et al. (2002) find some indication that the downward bias in the TTO utility caused by utility curvature approximately offsets other upwards biases (*e.g.*, loss aversion), so the TTO is consistent for longer reference durations than that we have used here (*i.e.*, 10 years). This may explain that the TTO was the best method in the study conducted by Bleichrodt and Johannesson (1997). In that study, the reference duration used in TTO measurements was set equal to 30 years. Thus, we

cannot reject the possibility that the different biases in the TTO cancel out for long temporal horizons. Further investigation is needed to identify the specific domains in which conventional TTO measurements may be acceptable.

Our paper has, at least, three limitations. First, we have used and *inter-rater* test. In some respect, this avoids anchoring effects and guarantees that the results at the aggregate level are consistent. However, we cannot estimate the degree of validity of the non-linear QALY model at the individual level as Bleichrodt et al. (1999) did. Further research is needed in order to study this question. Second, despite the use of a very similar methodology in the two surveys, the two samples varied in terms of educational and employment status. Both factors may potentially affect health state valuations (Badía et al., 1995). Lastly, the health profiles that are compared in our test only differ in the first year. Hence, it would be interesting to reply our test using larger durations. This question should be address in future investigations.

Our results seem to support that TTO utilities are estimated according to a non-linear (*power*) QALY model. In principle, it seems that the best estimate is around 0.65. This value is not very different from previous findings where values range from 0.74 to 0.8. In some respect, given the size and representativeness of the sample, and the number of health states evaluated, it seems that our estimate is a good candidate in practical decision making. We believe that for cost-effectiveness studies TTO utilities should be adjusted using a power function for duration.

Appendix 1: Elaboration of TTO utilities adjusted for utility curvature

CASE 1. *Elaboration of the adjusted TTO utility under the exponential QALY model*:

If health state Q is regarded as better than death, then substituting the exponential QALY model in the indifference between (Q_i, T_1) and (FH, T_2) reached by means of the TTO method, we have

$$H^*(Q_i) = \frac{1-e^{-\lambda T_2}}{1-e^{-\lambda T_1}} \qquad [A1]$$

If Q_i is regarded worse than death, then substituting the exponential QALY model in the indifference between (Q_i, T_1–T_2; FH, T_2) and dead reached by means of the TTO method, we have

$$H^*(Q_i) = -\left(\frac{e^{-\lambda(T_1-T_2)}-e^{-\lambda}}{1-e^{-\lambda(T_1-T_2)}}-1\right) \qquad [A2]$$

CASE 2. *Elaboration of the adjusted TTO utility under the power QALY model*:

If health state Q_i is regarded as better than death, then substituting the power QALY model in the indifference between (Q_i, T_1) and (FH, T_2) reached by means of the TTO method, we have

$$H^*(Q_i) = \left(\frac{T_2}{T_1}\right)^{\lambda} \qquad [A3]$$

If Q_i is regarded worse than death, then substituting the power QALY model in the indifference between (Q_i, T_1–T_2; FH, T_2) and dead reached by means of the TTO method, we have

$$H^*(Q_i) = - \frac{(T_1)^{\lambda} - (T_1 - T_2)^{\lambda}}{(T_1 - T_2)^{\lambda}} \qquad [A4]$$

References

Abellán, J. M. and J. L. Pinto, 1999, Quality-Adjusted Life-Years as Expected Utilities. Spanish Economics Review 2, 49-63.

Badía, X., E. Fernández and A. Segura, 1995, Influence of socio-demographic and health status variables on evaluation of health states in a Spanish population. European Journal of Public Health 5, 87-93.

Bleichrodt, H., 2002, A New Explanation for the Difference Between SG and TTO Utilities. Health Economics 11, 447-456.

Bleichrodt, H. and M. Johannesson, 1997, Standard Gamble, Time Trade-Off and Rating Scale: Experimental Results on the Ranking Properties of QALYs. Journal of Health Economics 16, 155-175.

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

Bleichrodt, H. and J.L. Pinto, 2000, A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. Management Science 46, 1485-96.

Bleichrodt, H. and A. Gafni, 1996, Time Preference, the Discounted Utility Model and Health. Journal of Health Economics 15, 49-67.

Bleichrodt, H. and J. L. Pinto, 2001, The Descriptive Validity of Quality-Adjusted Life-Years under Non-Expected Utility. Working Paper Erasmus University.

Bleichrodt, H. and J. Quiggin, 1997, Characterizing QALYs under a General Rank-Dependent Utility Model. Journal of Risk and Uncertainty 15, 151-165.

Bleichrodt, H., J. L. Pinto and J. M. Abellán, 2002, A Consistency Test of the Time Trade-Off. Working Paper Erasmus University.

Bostic, R., R. J. Hernstein and R. D. Luce, 1990, The effect on the preference reversal phenomenon of using choice indifferences. Journal of Economic Behavior and Organization 13, 193-212.

Chapman, G.B., 1997, Risk attitude and time preferences (Editorial). Medical Decision Making 17, 355-356.

Cher, D. J., J. Miyamoto and L. A. Lenert, 1997, Incorporating Risk Attitude into Markov-process Decision Models. Medical Decision Making 17: 340-350.

Dolan, P. A note on QALYs versus HYEs, 2000, Health states versus health profiles. International Journal of Technology Assessment in Health Care 16, 1220-4.

Dolan, P. and C. Gudex, 1995, Time preference, duration and health state valuations. Health Economics 4, 289-99.

Dolan, P. and M. Jones-Lee, 1997, The time trade-off: a note on the effect of lifetime reallocation of consumption and discounting. Journal of Health Economics 16, 731-739.

Dolan, P., 1997, Modeling valuations for EuroQol health states. Medical Care 35, 1095-108.

Dolan, P., 2000, The measurement of health-related quality of life for use in resource allocation decisions in health care, in: A. J. Culyer and J. P. Newhouse, eds., Handbook of Health Economics, vol. 1B (Elsevier Science, Amsterdam) 1723-1760.

Enemark, U., C.H. Lyttkens, T. Troeng, H. Weibull and J. Ranstam, 1998, Implicit discount rates of vascular surgeons in the management of abdominal aortic aneurysms. Medical Decision Making 18, 168-177.

Fishburn, P., 1965, Independence in Utility Theory with Whole Product Sets. Operations Research 13, 28-45.

Gafni, A. and G. W. Torrance, 1984, Risk attitude and time preference in health. Management Science 30, 440-451.

Gold, M. R., J. E. Siegel, L. B. Russell and M. C. Weinstein, 1996, Cost-effectiveness in health and medicine (Oxford University Press, New York).

Greene, W., 1999, Econometric analysis, third edition (MacMillan, New York).

Hall, J., K. Gerard, G. Salked y J. Richardson, 1992, A cost-utility analysis of mammography screening in Australia. Social Science and Medicine 34: 993-1004.

Happich, M., 2001, Utility functions for life years and health status – an additional remark. Disskusionpapiere der Technischen Universität Berlin.

Holmes, A. M., 1998, Measurement of short term health effects in economic evaluations. Pharmaeconomics 13, 171-174.

Jansen, S., A. Stiggelbout, P. Wakker, T. Vliet, J-W. Leer, M. Nooy and J. Kievit, 1998, Patient expected utilities for cancer treatments: a study on the feasibility of a chained procedure for the standard gamble and the time trade-off 18, 391-399.

Johannesson, M. , J. S. Pliskin and M. C. Weinstein, 1994, A Note on QALYs, Time Tradeoff and Discounting. Medical Decision Making 14: 188-193.

Kirsch, J. and A. McGuire, 2000, Establishing health state valuations for disease specific states: An example from heart desiase. Health Economics 9, 149-158.

Krabbe, P. F. M. and G. J. Bonsel, 1998, Sequence effects, health profile, and the QALY model. Medical Decision Making 18, 178-186.

Kuppermann, M., S. Shiboski, D. Feeney, E. P. Elkin and A. E. Washington, 1997, Can preference scores for discrete states be used to derive preference scores for an entire path of events? Medical Decision Making 17, 42-55.

Lipscomb, J., 1989, The preference for health in cost-effectiveness analysis. Medical Care 27, S233-53.

Loomes, G. and L. McKenzie, 1989, The use of QALYs in health care decision making. Social Science and Medicine 28, 299-308.

MacKeigan, L. D., A. Gafni, and B. J. O'Brien, 2003, Double discounting of QALYs. Health Economics 12, 165-169.

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

Maas, A. and P. Wakker, 1994, Additive conjoint measurement of multiattribute utility. Journal of Mathematical Psychology 38, 86-101.

McNeil, B. J., R. Weichselbaum and S. G. Pauker, 1978, Fallacy of the Five-Year Survival in Lung Cancer. New England Journal of Medicine 299, 1397-1401.

McNeil, B. J., R. Weichselbaum and S. G. Pauker. Tradeoffs between quality and quantity of life in laryngeal cancer. New England Journal of Medicine 1981; 305, 982-987.

Miyamoto, J. M. and S. A Eraker . A multiplicative model of the utility of survival duration and health quality. Journal of Experimental Psychology: General 1988; 117,3-20.

Miyamoto, J. M. and S. A. Eraker. Parameter Estimates for a QALY Utility Model. Medical Decision Making 1985; 5, 191-213.

Miyamoto, J. M. and S. A. Eraker. Parametric models of the utility of survival duration: Tests of axioms in a generic utility framework. Organizational Behavior and Human Decision Processes 1989; 44, 166-202.

Miyamoto, J. M. Quality-adjusted life years (QALY) utility models under expected utility and rank-dependent utility assumptions. Journal of Mathematical Psychology 1999; 43, 201-237.

Miyamoto, J. M. Utility assessment under expected utility and rank dependent utility assumptions. In G. B. Chapman and F. Sonnenberg. Decision making in health care: Theory, psychology, and applications. New York: Cambridge University Press 2000, 65-109.

Miyamoto, J., P. P. Wakker, H. Bleichrodt and H. J. M. Peters. The Zero-condition: A Simplifying Assumption in QALY Measurement and Multiattribute Utility. Management Science 1998; 44, 839-849.

Pauker, S. G., 1976, coronary artery surgery: The use of decision analysis. Annals of Internal Medicine, 85, 8-18.

Pliskin, J. S., D. S. Shepard and M C. Weinstein. Utility functions for life years and health status. Operations Research 1980; 28, 206-24.

Richardson, J., J. Hall and G. Salkeld, 1989, The compatibility of measurement techniques and the measurement of utility through time, in: C. S. Smith, ed., Economics and Health: proceedings of the eleventh conference of health Economics (Melborne: Public Sector management institute, Monash University).

Richardson, J., J. Hall and G. Salked, 1996, The measurement of utility in multiphase health states. International Journal of Technology Assessment of Health Care 12, 151-162.

Ried, W., 1998, QALYs versus HYEs – what's right and what's wrong. A review of the controversy. Journal of Health Economics 17, 607-626.

Stalmeier, P. F. M. and T. G. Bezembinder, 1999, The Discrepancy between Risky and Riskless Utilities: A Matter of Framing?. Medical Decision Making 19: 435-447.

Stiggelbout, A. M., G. M. Kiebert, J. Kievit, J. W. H. Leer, G. Stoter and J. C. J. M. de Haes, 1994, Utility Assessment in Cancer Patients: Adjustment of Time Tradeoff

Scores for the Utility of Life Years and Comparison with Standard Gamble Scores. Medical Decision Making 14: 82-90.

Sutherland, H. J., H. Llewellyn-Thomas, N. F. Boyd and J. E. Till, 1982, Attitudes towards quality of survival: The concept of maximum endurable time. Medical Decision Making 2, 299-309.

Torrance, G. W., 1986, Measurement of health state utilities for economic appraisal: a review. Journal of Health Economics 1,1-30.

Treadwell, J. R., 1998, Tests of Peferential Independence in the QALY model. Medical Decision Making 18, 418-428.

Verhoef, L. C. G., A. F. J. de Haan and W. A. J. van Daal, 1994, Rsik attitude in gambles with years of life: empirical support for prospect theory. Medical Decision Making 14, 194-200.

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

Wakker, P., 1994, Separating marginal utility and probabilistic risk aversion. Theory and Decision 36, 1-44.

Wakker, P. and A. Sttigelbout, 1995, Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Medical Decision Making 15, 180-186.

Wakker, P. and D. Deneffe, 1996, Eliciting von Newman-Morgenstern Utilities When Probabilities Are Distorted or Unknown. Management Science 8, 1131-1150

Figures and tables

Figure 1. Assessment of health profiles under the non-linear QALY model

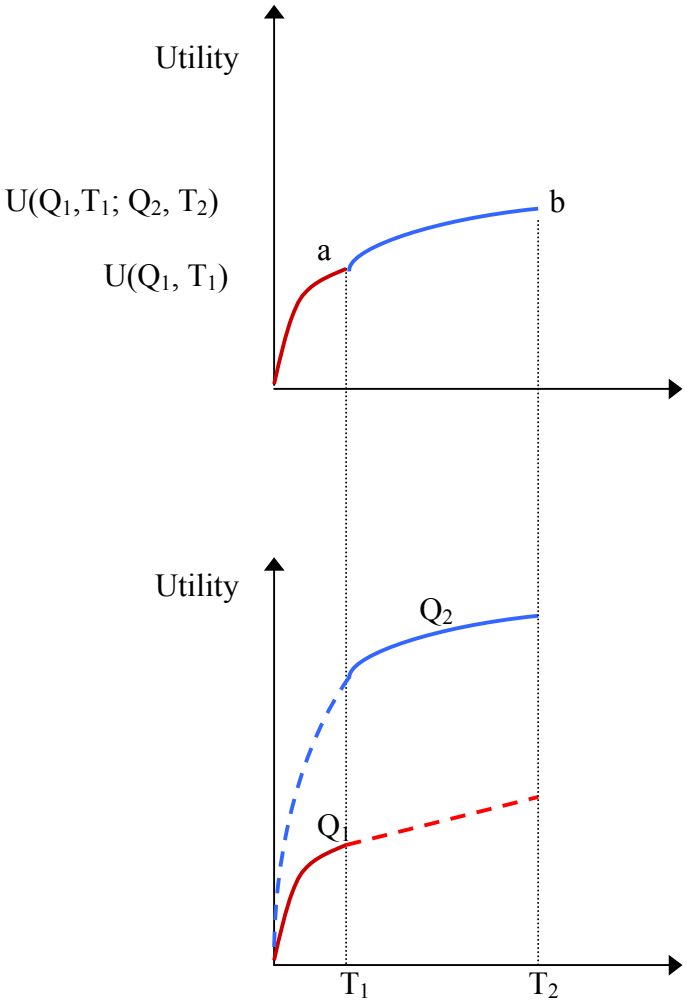

Figure 2. Differences under linear and non-linear QALY models

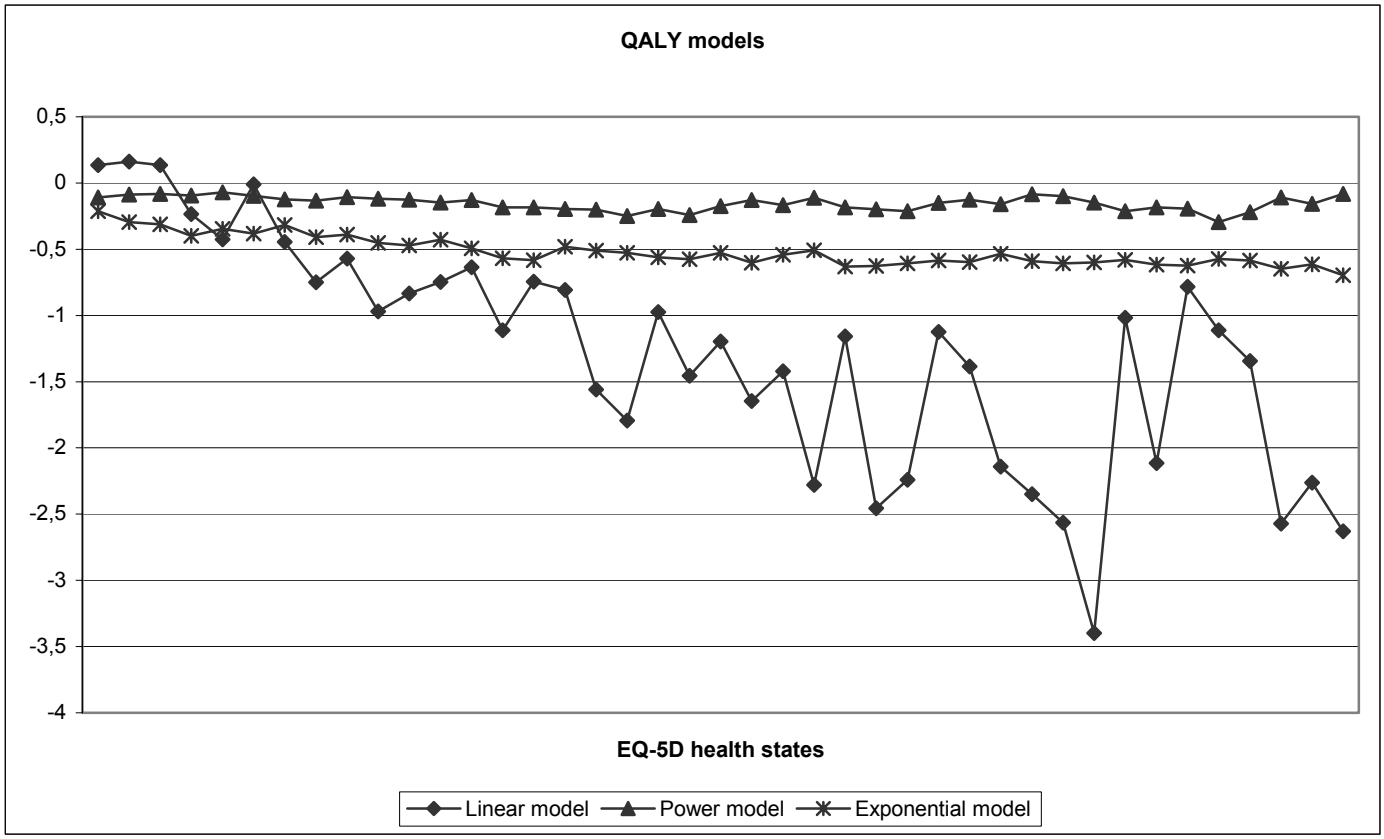

Table1. Health states

Very mild	Mild	Moderate	Severe
11112	11122	13212	33232
11121	11131	32331	23232
11211	11113	13311	23321
12111	21133	22122	13332
21111	21222	12222	22233
	21312	21323	22323
	12211	32211	32223
	11133	12223	32232
	22121	22331	33321
	12121	21232	33323
	22112	32313	23313
	11312	22222	33212

Note: health states '11111' and '33333' were also used.

Table 2. Sociodemographic characteristics

	Sample 1 (N_1=974)	Sample 2 (N_2=300)
Gender		
Female	535 (54.8%)	164 (54.7%)
Male	442 (45.2%)	136 (45.3%)
Age (years)		
Mean (SD)	45.5 (17.9)	45.6 (18.0)
Educational level		
No formal studies	149 (15.2%)	65 (21.7%)
Elementary	570 (58.3%)	184 (61.2%)
Secondary	140 (14.3%)	38 (12.8%)
Universitary	119 (12.2%)	13 (4.3%)
Employment status		
Employed	406 (41.6%)	121 (40.3%)
Unemployed	103 (10.5%)	14 (4.6%)
Housewife	193 (19.8%)	95 (31.7%)
Retired	143 (14.6%)	32 (10.7%)
Student	77 (7.9%)	28 (9.3%)
Others	55 (5.6%)	10 (3.4%)

Table 3. Difference between $\overline{\hat{T}_i}$ and $\bar{T}_i$ under linear and non-linear QALY models

Health state	**Linear QALY model**	**Exponential QALY model** $\lambda = 0.48$	**Power QALY model** $\lambda = 0.65$
11121	0.135	-0.214	-0.108
11112	0.162*	-0.296	-0.087
21111	0.136	-0.311	-0.083
11211	-0.235	-0.398	-0.095
12111	-0.425	-0.345	-0.071
11122	-0.009*	-0.381*	-0.096
12211	-0.446	-0.318	-0.124
12121	-0.749**	-0.408*	-0.132
12222	-0.570	-0.389*	-0.107
22121	-0.834**	-0.451*	-0.118
11113	-0.746**	-0.472	-0.127
22112	-0.970**	-0.428	-0.148
21222	-0.637*	-0.493	-0.129
22122	-1.064**	-0.567	-0.183
22222	-0.795**	-0.583**	-0.185*
11312	-0.807*	-0.482**	-0.196
11131	-1.559**	-0.509**	-0.202*
21312	-0.975**	-0.528**	-0.248*
12223	-1.794**	-0.561**	-0.195
13311	-1.417**	-0.576**	-0.241
21133	-0.889**	-0.527**	-0.174
11133	-1.557**	-0.602**	-0.128
21232	-1.856**	-0.542**	-0.167
13212	-2.456**	-0.508**	-0.111
21323	-1.385**	-0.631**	-0.184*
23321	-2.349**	-0.627**	-0.198*
32211	-1.158**	-0.608**	-0.214*
23232	-2.116	-0.584**	-0.149
22331	-2.142**	-0.596**	-0.127
22323	-2.279*	-0.534**	-0.159
22233	-2.241**	-0.589**	-0.085
33212	-1.125**	-0.607**	-0.099
23313	-2.565**	-0.599**	-0.148
32223	-1.017*	-0.581**	-0.214*
32313	-0.783**	-0.617**	-0.183
13332	-3.399**	-0.624**	-0.194*
32232	-1.345**	-0.574**	-0.296
32331	-1.111**	-0.586**	-0.219*
33321	-2.263**	-0.649**	-0.109
33232	-2.572*	-0.614**	-0.158
33323	-2.629**	-0.697**	-0.082

* Differences significant at the 5% level. ** Differences significant at the 1% level