Towards a better QALY model

José-María Abellán-Perpiñán, José-Luis Pinto-Prades, Ildefonso Méndez-Martínez and Xabier Badía-Llach

Summary

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 nonchronic health profiles. The chronic TTO utilities were then used to predict the responses to TTO questions with nonchronic 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 corrected using a power QALY model.

Keywords  cost-utility analysis; QALYs; power QALY model; predictive validity; time tradeoff

Introduction

It is well known that quality-adjusted life years (QALYs) can be only characterized as utilities under quite restrictive assumptions. First of all, a utility model for QALYs require that preferences over health status and longevity must be utility independent of other characteristics of the individual [1,2]. Thanks to this assumption the operationalization of QALYs requires only to estimate preferences over health. Even if we only estimate preferences over health, the traditional QALY model imposes further restrictions [3–6]. It has then been questioned if QALYs are really utility functions.

In this paper, we will keep the assumption that health and other characteristics of the individual are utility independent. We will then work with the most traditional framework. However, working within this framework, there are several ways of modelling QALYs depending on the restrictions that we impose on individual preferences. We will present some empirical evidence showing that by relaxing one of the assumptions of the traditional QALY model, namely, linearity in duration, the model is a better representation of preferences while keeping most of its simplicity. This does not imply that relaxing only this assumption (and keeping the rest) the QALY model is a perfect representation of preferences (a ‘true’ utility) but the empirical evidence that we present seems to suggest that it is an improvement that makes the model more realistic without complicating too much the task of estimating QALYs.

The basic idea of QALYs as a measure of utility consists in a multiplicative separation of quality and quantity of life [2,7]. Some authors refer to basic classes of QALY models as multiplicative utility models [8,9]. Formally, if \((Q, T)\) denotes a...
chronic health profile, where \( Q \) stands for constant health status during time period \( T \) until dead, then a general multiplicative QALY model asserts that

\[
U(Q, T) = H(Q) G(T)
\]

(1)

where \( H \) and \( G \) are utility functions over health status and duration, respectively. It is further commonly assumed that \( G(\cdot) \) 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 the assumption that the utility of life duration is linear. To obtain the linear QALY model from the multiplicative model (1), take \( G(T) = T \). Under this assumption, it follows that

\[
U(Q, T) = H(Q) T
\]

(2)

which is the standard way of computing QALYs for chronic health profiles.

As various researchers have emphasized (e.g. [10,11]) the linear QALY model greatly simplifies utility calculations, because practitioners need fewer measurements to calculate the overall utility of any nonchronic or general health profile (i.e. profiles in which health status can vary). This raises the question of how the linear QALY model for chronic health profiles can be extended to general health profiles. From now on we denote a typical general health profile as \((Q_1, T_1; Q_2, T_2; \ldots; Q_n, T_n)\) where health state \( Q_1 \) lasts for duration \( T_1 \), ... and state \( Q_n \) lasts for duration \( T_n \), followed by death. According to this notation subscripts do not necessarily denote successive and equal size units of time (e.g. subscript 1 denotes first year, subscript 2 denotes second year, etc.), but they denote durations which may vary from one health state to another health state (e.g. subscript 1 may denote a survival of 4 years in state \( Q_1 \), subscript 2 may denote a survival of 10 years in \( Q_2 \), etc.). Once the validity of the linear QALY model (2) is assumed, we only have to estimate the utility of each constituent health state described as a chronic condition (e.g. \( T_1 \) time periods in \( Q_1 \), followed by death), and then by assuming that preferences over any health profile are additive across separate (or disjoint) time durations, Equation (2) becomes

\[
U(Q_1, T_1; Q_2, T_2; \ldots; Q_n, T_n) = \sum_{i=1}^{n} H(Q_i) T_i
\]

(3)

where \( Hs \) are the utilities of health states.

Of course, health profile utilities can also be computed in a similar way by applying non-linear QALY models (i.e. models in which the utility function for duration is no longer assumed to be linear), but it requires as an additional task to fit some parametric utility function for duration to data [12–14] or, alternatively, to apply some group estimate of the parameter [15]. Under the assumption of additivity over separate time periods, the multiplicative QALY model (1) postulates that [9, p. 206, Equation (6)]:

\[
U(Q_1, T_1; Q_2, T_2; \ldots; Q_n, T_n)
= \sum_{i=1}^{n} H(Q_i) \left\{ G \left( \sum_{j=0}^{T_i} T_j \right) - G \left( \sum_{j=0}^{T_i-1} T_j \right) \right\}
\]

(4)

where \( T_0 = 0 \).

Equation (4) asserts that the utility of a general health profile is the sum of the utility increments during the separate periods. In this way, for example, the utility of the profile \((Q_1, 3\text{yr}; Q_2, 5\text{yr})\) is calculated according to Equation (4) as \( H(Q_1) \{G(3) - G(0)\} + H(Q_2) \{G(8) - G(3)\} \).

The main concern of this paper is to test the ability of the linear QALY model to predict preferences over general health profiles from utilities elicited in chronic health states. This type of empirical exercise is what we call throughout the paper a test of predictive validity. Therefore, we focus on the validity of the linear model to yield accurate QALY calculations instead of testing any of the axioms (e.g. risk neutrality) essential to the model within a specific utility framework (e.g. EU). Hence, we adopt a different approach from that followed in some empirical studies (e.g. [16–19]) devoted to test axioms directly.

One advantage of our approach is that results do not depend on a specific utility theory (e.g. EU or rank-dependent utility), because the prediction to be tested is common to any multiplicative QALY model. This is supported by recent findings reported by Doctor et al. [20]. They found robust evidence in favor of the linear QALY model for chronic health profiles within a general utility framework compatible with EU, RDU, and Prospect Theory (PT). Then they argue that the observed violations of the linear QALY model (e.g. [3,12–14,21,22]) might had been caused by violations of EU.

We build on all those previous studies that have tested some predictions of the linear QALY model. In this way, our approach resembles the methodology followed by those studies that have
compared the number of QALYs yielded by a health profile to some sort of holistic assessment (i.e. a non-separable representation of individual preference) for the same profile. These studies have generally found large discrepancies between QALYs and holistic assessments (e.g. [11,23,24]). Also, in some respect, our paper is close to the tests of predictive validity performed by Bleichrodt and Johannesson [25] and Bleichrodt et al. [26]. 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. tested the validity of the SG method with and without probability transformation. These authors found that applying probability weighting in SG measurements increased the consistency of QALYs with individual preferences. However, consistency was not significantly increased further when utility curvature was also taken into account by means of linear/exponential and log/power families of utility functions for duration.

This paper is intended as an attempt to generalize above-mentioned studies of predictive validity. In particular, we try to shed some new light on four related issues, namely:

(a) Which is the non-linear QALY model with a higher predictive validity?:

We try to go beyond simply testing whether predictions under the linear QALY model are accurate. Consequently, we also test the predictive performance of two non-linear QALY models, namely: the power QALY model and the exponential QALY model. Although the linear QALY model is by far the most widely used in practice, the power and the exponential models have also been used in health economics and medical decision making (e.g. [27–30]), and their preference foundations are well known [3,5,8,9,19]. The power model assumes constant relative risk attitude, that is, for a given health state, preference between gambles yielding different durations does not vary if all of them are multiplied by the same constant duration. In formal terms (e.g. [9]), let

\[ [(Q,T_1),p_1];\ldots;[(Q,T_n),p_n] \]

be a gamble, where \( p_i \) is the probability attached to outcome \((Q,T_i)\), and \([[(Q,T'_1),p_1];\ldots;[(Q,T'_n),p_n]] \] denotes other gamble, where \( T'_i \neq T_i \) for any \( i \). Constant relative risk attitude requires that if \([[(Q,T_1),p_1];\ldots;[(Q,T_n),p_n]] \) then \([[(Q,r T_1),p_1];\ldots;[(Q,r T_n),p_n]] \), where \( r \) is any duration. On the contrary, the exponential model assumes constant absolute risk attitude, that is, preference between gambles remains invariant if a constant duration is added to all durations. Formally, this assumption requires that if \([[(Q,T_1),p_1];\ldots;[(Q,T_n),p_n]] \) then \([[(Q,s+T_1),p_1];\ldots;[(Q,s+T_n),p_n]] \). Hence, we think that it is important to examine the impact of utility curvature on TTO measurements.

(b) The influence of utility curvature on a wide range of health states:

It seems that empirical evidence available hitherto supports that the utility function for duration is concave rather than linear [31]. However, it has not been still tested if the potential improvement of the non-linear over the linear QALY model is constant for all degrees of severity. To conduct this test we use a set of 43 EQ-5D health states. Bleichrodt et al. [26] only used two health states (full health and one intermediate health state). In other papers, researchers used three health states at best (e.g. [19,32,33]).

(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 for practical relevance. The most widely used instrument for use as a measure of health outcomes in our context is the EuroQol [34,35]. The quality weights of this instrument were estimated with the TTO on a random sample of approximately 3000 members of the general population of the United Kingdom [36,37]. However, various researchers have warned about the risk that TTO utilities are biased downwards if utility for duration is concave (e.g. [12,38,39]). Hence, we think that it is important to examine the impact of utility curvature on TTO measurements.
were seriously biased by utility curvature then the EuroQol algorithm might be also biased, leading to wrong allocations of health resources. However, the correction of this bias would not be difficult to conduct.

(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. [11,38,40]). 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 [41]. We think that it is 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 sample (nearly 1300 people) randomly drawn from general (Spanish) population.

The paper is structured as follows. In the next section, we describe the test of the linear QALY model that we 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 nonchronic health profiles which combine each health state with full health. The chronic TTO utilities are then used to predict the responses to new TTO questions involving comparisons between two nonchronic health profiles. Hence, our null hypothesis assumes that 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 utility specifications (i.e. power and exponential models) improve the prediction. This alternative hypothesis is described in the following section. In the next section, we present the study design and the statistical methods to be used. The penultimate section shows the results. Discussion closes the paper.

**Predictive validity of QALYs**

The predictive validity of the QALY model will be tested as follows:

1. Assume the validity of the linear QALY model (2).

2. Assume that the TTO is used to elicit the utility of health states $Q^I$ and $Q^W$ defined as chronic conditions [42].

3. Assume that we have two health profiles such as $(F_{H1}, T_1; Q^W, T_2; FH_3, T_3)$ and $(Q^I_1, T_1; Q^I_2, T_2; FH, T_3)$, both followed by death. Since health state $Q^I$ is the same during periods $T_1$ and $T_2$, we can represent the profiles as $(F_{H1}, T_1; Q^W, T_2; FH_3, T_3)$ and $(Q^I, T_{1+2}; FH, T_3)$ where $T_{1+2}$ stands for the sum $T_1 + T_2$.

4. Assume that the linear QALY model (3) describes correctly the preferences over non-chronic health profiles and that the individual is indifferent between the two profiles. Let $H(FH) = 1$ and $H(death) = 0$. Then applying the linear QALY model (3) to the indifference between both health profiles gives

$$T_1 + H(Q^W)T_2 = H(Q^I)T_{1+2} \quad (5)$$

In order to test the predictive validity of the linear QALY model, we may fix two of the three durations (i.e. $T_1$, $T_2$, $T_{1+2}$) and elicit the third one. The elicited duration $T_i$ for $Q^I$ should coincide with the estimate (denoted by $\hat{T}_i$) we get from Equation (5). Thus the null hypothesis of the test asserts that there is no significant difference between predicted ($\hat{T}_i$) and observed ($T_i$) durations. On the contrary, our alternative hypothesis asserts that non-linear QALY models will describe preferences better than the linear one. The best model will be that one that minimize the difference $\hat{T}_i - T_i$ for the whole set of health states.

**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. In both cases the utility function over duration is described by a single parameter $\lambda$. Derivations of TTO utilities after adjusting by power and exponential functions are provided in Appendix A.

Case 1: The exponential QALY model asserts that

$$U(Q, T) = H(Q) k(1 - e^{-\lambda T}) \quad (6)$$

where coefficient $k$ is a scaling constant equal to $T^*/(1 - e^{-\lambda T^*})$ and $T^*$ is the maximum lifetime
duration in the given domain. Within the range from 0 to $T^*$, the exponential function $k(1 - e^{-\lambda T})$ can be concave ($\lambda > 0$) as well as convex ($\lambda < 0$). Exponential model (6) reduces to the constant discounting model [43] if time is discrete.

If we assume the validity of the exponential QALY model (6), under the assumption of additivity over disjoint time periods, Equation (5) changes into

$$k(1 - e^{-\lambda T_1}) + H^*(Q_w^W)\{k(1 - e^{-\lambda T_{1+2}})
- k(1 - e^{-\lambda T_1})\} = H^*(Q_w^I)k(1 - e^{-\lambda T_{1+2}})$$

where asterisk * denotes that TTO utility $H$ has been adjusted by $\lambda$.

Case 2: The power QALY model asserts that

$$U(Q, T) = H(Q) k^T$$

where $k$ is a scaling constant equal to $T^*/(T^*)^\lambda$ and $T^*$ is the longest duration in the given domain. Within the range from 0 to $T^*$, the power function $k^T$ can be concave ($\lambda < 1$) as well as convex ($\lambda > 1$). The power specification implies that proportional time tradeoffs are constant [3]. Under EU the power parameter may reflect risk aversion, positive time preference and diminishing marginal value [44].

If we assume the validity of the power QALY model (8), under the assumption of additivity over disjoint time periods, Equation (5) changes into

$$kT_1^\lambda + H^*(Q_w^W)\{kT_{1+2}^\lambda - kT_1^\lambda\} = H^*(Q_w^I)kT_{1+2}^\lambda$$

where asterisk * denotes that TTO utility $H$ has been adjusted by $\lambda$.

**Methods**

**Subjects**

We conducted two surveys in order to perform the test. In the 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 [37] used to model the EuroQol algorithm. In the first survey, each respondent assessed a random selection of 13 health states including two very mild states, three mild states, three moderate states, three 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 (Table 1).

**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

<table>
<thead>
<tr>
<th>Table 1. Health states</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Very mild</strong></td>
</tr>
<tr>
<td>11112</td>
</tr>
<tr>
<td>11121</td>
</tr>
<tr>
<td>11211</td>
</tr>
<tr>
<td>12111</td>
</tr>
<tr>
<td>21111</td>
</tr>
<tr>
<td>21312</td>
</tr>
<tr>
<td>23112</td>
</tr>
<tr>
<td>11133</td>
</tr>
<tr>
<td>22121</td>
</tr>
<tr>
<td>12121</td>
</tr>
<tr>
<td>22112</td>
</tr>
<tr>
<td>11312</td>
</tr>
</tbody>
</table>

*Note: Health states ‘11111’ and ‘33333’ were also used.*
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. This task was only performed in order to familiarize respondents with health states.

Next, we applied the same procedure used by Dolan [37] in order to obtain direct TTO utilities of the subset of EQ-5D health states on which the EuroQol model is based on. In this way, our results will allow us to judge if EuroQol values generated from TTO valuations for 10 years of reference may be biased or not. 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’. In case of negative health states utilities, they were not restricted to move between 1 and -1, because as Patrick et al. [45] acknowledged that is an arbitrary transformation after which valuations can no longer be interpreted as true utilities. The reason is that utilities are unique up to location and scale. We can establish arbitrarily two points of the utility function but not three. If we restrict utilities to move between +1 and -1 we are arbitrarily establishing the utility of three points, namely, good health, death and the utility of the worst health state. The numbers produced in this way cannot be interpreted as utilities.

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 to compare two health profiles such as (11111, \( T_1 \); 33333, \( T_2 \); 11111, \( T_3 \)) and (\( Q^i \), \( T_1 + 2 \); 11111, \( T_3 \)) both followed by death. It was set that \( T_1 + T_2 = T_1 + 2 = 12 \) months, and \( T_3 = 9 \) years. If \( T_1 = 0 \) then \( Q^i \) is as bad as health state ‘33333’. If \( T_1 = 1 \) then \( Q^i \) is as good as state ‘11111’. Hence, all health states are covered.

We then elicited \( T_1 \) and compared it with \( \hat{T}_1 \) obtained using Equation (5), that is

\[
\hat{T}_1 = \frac{H(Q^i) - H(33333)}{1 - H(33333)} \times 12 \tag{10}
\]

An example of the method is provided in Figure 1.

Empirical evidence supports that preferences inferred from choices are more consistent than preferences inferred from matching or judgments of selling prices [46]. In this way, both chronic and non-chronic health profiles were presented to respondents as choices. For example, questions with non-chronic 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 was not indifferent, the interviewer 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 both the two-sample \( t \) test. If significant differences were found, we would then estimate the optimal value \( \hat{\lambda} \) of power and exponential coefficients of the QALY models described in ‘Power and exponential QALY models’ section.

The procedure we would employ in order to estimate the optimal coefficients would be an optimization algorithm based on the Newton–Raphson method [47]. This procedure is not a regression analysis. In the context of this study, goodness-of-fit means minimizing differences between responses constructed from sample 1 and responses obtained from sample 2.

The specific procedure is as follows. We start setting a utility function for duration with \( \hat{\lambda} = 0 \) in the exponential model, and with \( \lambda = 1 \) in the power model, i.e. we fix a linear utility function. Next, by varying \( \hat{\lambda} \) for each respondent within 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 \( \hat{T}_i \) and the mean observed responses \( \hat{T}_i \) across 41 EQ-5D health states,
where subscript \(i\) denotes the health state \((i = 1, 2, \ldots, 41)\). We note that health states '11111' and '33333' were used as reference health states in the questions involving health profiles, so these states were not included in estimations. For this reason discrepancies were only minimized for each one of the remaining 41 health states.

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

\[
\sum_{i=1}^{41} (\hat{T}_i(\lambda) - T_i)^2
\]

(11)

where \(\hat{T}_i(\lambda) = (1/J) \sum_{j=1}^{J} \hat{T}_{ij}(\lambda)\), and subscript \(j\) denote each respondent belonging to sample 1.

### Results

The two samples were very similar in terms of gender, age, and health status \(\chi^2, P = 0.97\) and 0.93, respectively), though we found a significant difference in educational status \(\chi^2, P < 0.01\).

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 were not able to make some tradeoffs. Hence, the final analysis is based on the responses of 923 and 296 subjects, respectively.

Table 2 shows that the null hypothesis cannot be rejected for eight of the 41 states displayed (they have been ranked in increasing order of severity according to chronic utilities). Thus, differences between \(\hat{T}_i\) and \(\hat{T}_i\) are statistically significant for almost all health states. The largest differences were found for severe states. The maximum difference (= 3.399, i.e. 3 months and 12 days) was observed for the health state ‘13332’.

It is noticeable that \(\hat{T}_i\) was lower than \(\hat{T}_i\) for the majority of the states. Only for three very mild

---

**Figure 1. Example of test conducted**

If \(H(Q^i) = 0.5\) and \(H(33333) = -1\) \(\Rightarrow \hat{T}_i\) should be 9 months
Table 2. Difference between mean predicted responses and mean observed responses (standard error) under linear and non-linear QALY models

<table>
<thead>
<tr>
<th>Health state</th>
<th>Linear QALY model</th>
<th>Exponential QALY model, λ = 0.48</th>
<th>Power QALY model, λ = 0.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>11121</td>
<td>0.135 (0.864)</td>
<td>-0.214 (0.382)</td>
<td>-0.108 (0.670)</td>
</tr>
<tr>
<td>11112</td>
<td>0.162* (0.065)</td>
<td>-0.296 (0.368)</td>
<td>-0.087 (0.611)</td>
</tr>
<tr>
<td>21111</td>
<td>0.136 (0.257)</td>
<td>-0.311 (0.161)</td>
<td>-0.083 (0.689)</td>
</tr>
<tr>
<td>11211</td>
<td>-0.235 (0.272)</td>
<td>-0.398 (0.392)</td>
<td>-0.095 (0.827)</td>
</tr>
<tr>
<td>12111</td>
<td>-0.425 (0.395)</td>
<td>-0.345 (0.874)</td>
<td>-0.071 (0.855)</td>
</tr>
<tr>
<td>11122</td>
<td>-0.009 (0.004)</td>
<td>-0.381* (0.158)</td>
<td>-0.096 (0.157)</td>
</tr>
<tr>
<td>12211</td>
<td>-0.446 (0.890)</td>
<td>-0.318 (0.917)</td>
<td>-0.124 (0.691)</td>
</tr>
<tr>
<td>12121</td>
<td>-0.749** (0.199)</td>
<td>-0.408* (0.166)</td>
<td>-0.132 (0.596)</td>
</tr>
<tr>
<td>12222</td>
<td>-0.570 (0.510)</td>
<td>-0.389* (0.157)</td>
<td>-0.107 (0.802)</td>
</tr>
<tr>
<td>22121</td>
<td>-0.834* (0.060)</td>
<td>-0.451* (0.186)</td>
<td>-0.118 (1.009)</td>
</tr>
<tr>
<td>11113</td>
<td>-0.746** (0.217)</td>
<td>-0.472 (0.629)</td>
<td>-0.127 (0.589)</td>
</tr>
<tr>
<td>22112</td>
<td>-0.970** (0.349)</td>
<td>-0.428 (0.999)</td>
<td>-0.148 (1.127)</td>
</tr>
<tr>
<td>21222</td>
<td>-0.637* (0.272)</td>
<td>-0.493 (0.628)</td>
<td>-0.129 (1.135)</td>
</tr>
<tr>
<td>21212</td>
<td>-1.064** (0.087)</td>
<td>-0.567 (0.635)</td>
<td>-0.183 (0.916)</td>
</tr>
<tr>
<td>22222</td>
<td>-0.793** (0.226)</td>
<td>-0.583** (0.296)</td>
<td>-0.185* (0.090)</td>
</tr>
<tr>
<td>11312</td>
<td>-0.807* (0.353)</td>
<td>-0.482* (0.203)</td>
<td>-0.196 (0.759)</td>
</tr>
<tr>
<td>11131</td>
<td>-1.559** (0.058)</td>
<td>-0.509** (0.222)</td>
<td>-0.202* (0.085)</td>
</tr>
<tr>
<td>21312</td>
<td>-0.975** (0.144)</td>
<td>-0.528** (0.215)</td>
<td>-0.248* (0.111)</td>
</tr>
<tr>
<td>12223</td>
<td>-1.794* (0.566)</td>
<td>-0.561** (0.286)</td>
<td>-0.195 (0.949)</td>
</tr>
<tr>
<td>13111</td>
<td>-1.417** (0.511)</td>
<td>-0.576** (0.268)</td>
<td>-0.241 (0.648)</td>
</tr>
<tr>
<td>21133</td>
<td>-0.889** (0.306)</td>
<td>-0.527** (0.243)</td>
<td>-0.174 (0.314)</td>
</tr>
<tr>
<td>11333</td>
<td>-1.557** (0.417)</td>
<td>-0.602** (0.246)</td>
<td>-0.128 (0.240)</td>
</tr>
<tr>
<td>21232</td>
<td>-1.858** (0.668)</td>
<td>-0.542** (0.260)</td>
<td>-0.167 (0.458)</td>
</tr>
<tr>
<td>13212</td>
<td>-2.456** (0.529)</td>
<td>-0.508** (0.254)</td>
<td>-0.111 (0.795)</td>
</tr>
<tr>
<td>21323</td>
<td>-1.385** (0.220)</td>
<td>-0.631** (0.299)</td>
<td>-0.184* (0.091)</td>
</tr>
<tr>
<td>23321</td>
<td>-2.349** (0.574)</td>
<td>-0.627** (0.250)</td>
<td>-0.198* (0.077)</td>
</tr>
<tr>
<td>32211</td>
<td>-1.158** (0.390)</td>
<td>-0.608** (0.298)</td>
<td>-0.214* (0.085)</td>
</tr>
<tr>
<td>23232</td>
<td>-2.116 (1.247)</td>
<td>-0.584** (0.233)</td>
<td>-0.149 (0.330)</td>
</tr>
<tr>
<td>22331</td>
<td>-2.142** (0.296)</td>
<td>-0.596** (0.288)</td>
<td>-0.127 (0.609)</td>
</tr>
<tr>
<td>22323</td>
<td>-2.279* (1.097)</td>
<td>-0.534** (0.211)</td>
<td>-0.159 (0.086)</td>
</tr>
<tr>
<td>22233</td>
<td>-2.241** (0.852)</td>
<td>-0.589** (0.300)</td>
<td>-0.085 (0.920)</td>
</tr>
<tr>
<td>33212</td>
<td>-1.125** (0.410)</td>
<td>-0.607** (0.280)</td>
<td>-0.099 (0.778)</td>
</tr>
<tr>
<td>23313</td>
<td>-2.565 (0.916)</td>
<td>-0.590** (0.287)</td>
<td>-0.148 (0.386)</td>
</tr>
<tr>
<td>32223</td>
<td>-1.017* (0.461)</td>
<td>-0.581** (0.241)</td>
<td>-0.214* (0.087)</td>
</tr>
<tr>
<td>32313</td>
<td>-0.783** (0.221)</td>
<td>-0.617** (0.257)</td>
<td>-0.183 (0.727)</td>
</tr>
<tr>
<td>33312</td>
<td>-3.399* (1.124)</td>
<td>-0.624** (0.303)</td>
<td>-0.194* (0.076)</td>
</tr>
<tr>
<td>33223</td>
<td>-1.347* (0.329)</td>
<td>-0.574** (0.264)</td>
<td>-0.296 (0.951)</td>
</tr>
<tr>
<td>32331</td>
<td>-1.114* (0.324)</td>
<td>-0.586** (0.281)</td>
<td>-0.219* (0.085)</td>
</tr>
<tr>
<td>33321</td>
<td>-2.263* (0.641)</td>
<td>-0.649** (0.287)</td>
<td>-0.109 (0.665)</td>
</tr>
<tr>
<td>33322</td>
<td>-2.572* (1.143)</td>
<td>-0.614** (0.266)</td>
<td>-0.158 (0.465)</td>
</tr>
<tr>
<td>33323</td>
<td>-2.629** (0.915)</td>
<td>-0.697** (0.347)</td>
<td>-0.082 (0.552)</td>
</tr>
</tbody>
</table>

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

states there is a positive difference and it is not statistically significant in two of them. Moreover, differences are larger for the more severe health states. We tested various OLS regression models in which the difference between predicted and observed responses depended on the ranking of health states as a function of severity (i.e. chronic utility). The model that best fit the data (R² = 0.54) was the linear specification β·Rank, where Rank denotes the rank-order of the health states. We found that coefficient β was negative (−0.0587) and highly significant (P < 0.0001).
suggesting that the difference \((\bar{T}_i - \bar{T}_j)\) gets more negative as the health status gets worse. Hence, it seems that discrepancies increase with severity. This implies that the linear QALY model is only a good descriptive model for mild health states.

Next, we try to remove differences between \(\bar{T}_i\) and \(\bar{T}_j\) using non-linear QALY models. We find that differences are minimized under exponential and power models for \(\lambda = 0.48\) and 0.65, respectively (see Table 2). These parameters imply a concave-shaped utility function for duration.

Table 2 shows 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.

**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, at least in the context of TTO measurements, and at least with the durations we have used, 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. [48,49]).

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. [50]) and it has been also proposed as a good instrument for medical decision analysis (e.g. [13,51]). In addition, the power model is consistent with the property of constant proportional time trade-offs [3].

4. The parameter estimated \((\hat{\lambda} = 0.65)\) indicates that the shape of the utility function for life duration is concave rather than linear. As TTO questions are framed in terms of certainty, hence 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 [52].

5. The finding of concavity is in line with most empirical studies that estimated the utility function by a power function (e.g. [12,13,19,51]). We note, however, that our parameter has a value somewhat lower than previous studies (reported mean \(\hat{\lambda}\) 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 we have estimated a summary parameter from individual data adjusting simultaneously by all health states. Lastly, experiments conducted by Bleichrodt and Pinto [19] tested the power function under unexpected 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. 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. [38] recommended. Our finding is similar to results reported by MacKeean et al. [53]. They also adjusted TTO utilities, and found a large difference between discounted TTO-based QALYs and holistic utilities.

7. Our data shows that, at least for reference duration of 10 years, TTO utilities estimated under linearity may be biased. This implies that the *EuroQol* algorithm, which is based on conventional TTO valuations, might lead to wrong allocation of resources.

Our paper has several potential 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. [26] did. Second, general health profiles that are compared in our test only differ in the first year. Hence, it would be interesting to redo our test using larger durations. Third, the time trade-off frame for the chronic profile was established arbitrarily in 10 years. This time horizon is not realistic in most cases and it may have influenced our results. However, it is not clear in the literature if there is some sort of ‘gold standard’ on this question. Fourth, there were some very extreme negative values for bad health states that may have influenced our results. It is not clear in the literature how to deal with very negative utilities. Apparently, two tentative solutions are possible. First, asymmetry between positive and negative utilities could be corrected by transforming valuations to a scale $[-1, 1]$. Second, very negative values can be treated as outliers, excluding them from the data set. The first option does not have a clear theory as we have explained above. The second option is arbitrary. It is not clear to us when scores are so negative that have to be excluded from analysis. We have then decided to accept social preferences as they are.

Most previous empirical evidence suggests that the assumption of linear utility function for duration fails to describe preferences. Our data agree with that evidence. Specifically, what our data suggest is that TTO utilities elicited by using short reference durations (10 years) seem to be biased by utility curvature. In this case, it seems that adjusting TTO utilities by means of a power function with a parameter around 0.65 is able to remove the bias. This does not mean, however, that our results can be extrapolated to those practical exercises with the TTO in which longer time horizons are fixed through the assessment. Bleichrodt et al. [54] found 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 [25]. 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. At this stage, what we can say is that TTO questions using a short reference duration may be biased by utility curvature and that this bias can be corrected by using a power QALY model.

Acknowledgements

Jose-Luis Pinto-Prades acknowledges financial support from Ministerio de Educació y Ciencia grant SEJ2004-05079 and from Fundacion BBVA (project BBVA1-04). Jose Maria Abellan acknowledges financial support from Fundacion Seneca PB/67/FS/02.

Appendix A: 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 $(Q, X)$ and $(FH, Y)$ reached by means of the TTO method, we have

$$H * (Q) = \frac{1 - e^{-\lambda Y}}{1 - e^{-\lambda X}}$$

(A1)

If $Q$ is regarded worse than death, then substituting the exponential QALY model in the indifference between $(Q, X-Y)$ and $(FH, Y)$ and dead reached by means of the TTO method, we have

$$H * (Q) = -\left(\frac{e^{-\lambda(X-Y)} - e^{-\lambda}}{1 - e^{-\lambda(X-Y)}} - 1\right)$$

(A2)

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

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

$$H * (Q) = \left(\frac{Y}{X}\right)^{\lambda}$$

(A3)

If $Q$ is regarded worse than death, then substituting the power QALY model in the indifference between $(Q, X-Y)$ and $(FH, Y)$ and dead reached by means of the TTO method, we have

$$H * (Q) = -\frac{(X)^{\lambda} - (X-Y)^{\lambda}}{(X-Y)^{\lambda}}$$

(A4)


References


Copyright © 2006 John Wiley & Sons, Ltd.

Health Econ. 15: 665–676 (2006)