Validation of the Underlying Assumptions of the Quality-Adjusted Life-Years Outcome: Results from the ECHOUTCOME European Project

Ariel Beresniak, Antonieta Medina-Lara, Jean Paul Auray, Alain De Wever, Jean-Claude Praet, Rosanna Tarricone, Aleksandra Torbica, et al.

PharmacoEconomics

ISSN 1170-7690
Volume 33
Number 1

PharmacoEconomics (2015) 33:61-69
DOI 10.1007/s40273-014-0216-0

Springer

PharmacoEconomics (2015) 33:61–69
DOI 10.1007/s40273-014-0216-0

ORIGINAL RESEARCH ARTICLE

Validation of the Underlying Assumptions of the Quality-Adjusted Life-Years Outcome: Results from the ECHOUTCOME European Project

Ariel Beresniak · Antonieta Medina-Lara · Jean Paul Auray · Alain De Wever · Jean-Claude Praet · Rosanna Tarricone · Aleksandra Torbica · Danielle Dupont · Michel Lamure · Gerard Duru

Published online: 18 September 2014

Abstract

Background Quality-adjusted life-years (QALYs) have been used since the 1980s as a standard health outcome measure for conducting cost-utility analyses, which are often inadequately labeled as 'cost-effectiveness analyses'. This synthetic outcome, which combines the quantity of life lived with its quality expressed as a preference score, is currently recommended as reference case by some health technology assessment (HTA) agencies. While critics of the QALY approach have expressed concerns about equity and ethical issues, surprisingly, very few have tested the basic methodological assumptions supporting the QALY equation so as to establish its scientific validity.

Objectives The main objective of the ECHOUTCOME European project was to test the validity of the underlying assumptions of the QALY outcome and its relevance in health decision making.

Methods An experiment has been conducted with 1,361 subjects from Belgium, France, Italy, and the UK. The subjects were asked to express their preferences regarding various hypothetical health states derived from combining different health states with time durations in order to compare observed utility values of the couples (health state, time) and calculated utility values using the QALY formula.

Results Observed and calculated utility values of the couples (health state, time) were significantly different, confirming that preferences expressed by the respondents were not consistent with the QALY theoretical assumptions.

Conclusions This European study contributes to establishing that the QALY multiplicative model is an invalid measure. This explains why costs/QALY estimates may vary greatly, leading to inconsistent recommendations relevant to providing access to innovative medicines and health technologies. HTA agencies should consider other more robust methodological approaches to guide reimbursement decisions.

Key Points

Underlying assumptions of the quality-adjusted life-year (QALY) are not validated by an experiment conducted in four European countries.

The fact that the QALY metric is an invalid measure explains why costs/QALY estimates may vary greatly.

Health technology assessment agencies should consider other current and new methodological approaches for healthcare decision making.

A. Beresniak (✉)
Data Mining International, Route de l'Aéroport, 29-31, CP 221, 1215 Geneva 15, Switzerland
e-mail: aberesniak@datamining-international.com

A. Beresniak
LIRAES, Paris-Descartes University, Paris, France

A. Medina-Lara
Exeter University, Exeter, UK

A. Medina-Lara
CERGAS, Bocconi University, Milan, Italy

J. P. Auray · G. Duru
Cyklad Group, Lyon, France

A. De Wever · J.-C. Praet
Université Libre de Bruxelle, Brussels, Belgium

R. Tarricone · A. Torbica
Department of Policy Analysis and Public Management, Bocconi University, Milan, Italy

D. Dupont · M. Lamure
Claude-Bernard University, Lyon, France

1 Introduction

In order to assist resource allocation decisions, economists in the 1980s proposed the use of the quality-adjusted life-year (QALY) as a health outcome measure for use in cost-utility analyses. The QALY outcome takes into account both the quantity and quality of life relevant to hypothetical health states, and allows comparison between healthcare interventions across different therapy areas by relating their respective cost/QALY ratios to league tables.

The principles of the QALY are derived from expected utility theory [1]; a complex theory based on the Von Neumann-Morgenstern utility theorem [2], which relates to the affine transformation property of the utility function[1]. Applied to health, this theory takes into account the effect that a healthcare intervention (defined as a medicine, a device, or a diagnostic procedure), has on both a person's quantity and quality of life [3]. Hence, the QALY is the product of life expectancy (estimated in years) and a measure of the quality of the remaining life-years (estimated in utilities or quality-of-life values):

QALY = quality of life (expressed in 'utility') × number of life-years

The calculation of a QALY is based on a simple multiplicative format, the 'multiplicative model'. Two types of models are described in the frame of the multi-attribute utility theory: the additive multi-attribute utility model and the multiplicative multi-attribute utility model (or 'multiplicative model'), which is the specification of the QALY model [1]. The multiplicative model assumes that for a given health state, the utility (preference) of one pair (time duration and health state utility) should be equal to the product of the utility for each component of the pair (time duration and health state utility). For further simplification, the QALY model assumes that the utility of the time is identical to its quantity, $u(t) = t$.

Over the years, a number of criticisms and issues about the use of QALYs have been raised. These include important ethical aspects, e.g., the rationale and moral considerations of the QALY as an outcome measure to accept or deny people access to treatments that can potentially prolong life [4], and methodological limitations of the approach, e.g., that utilities or quality-of-life indices required to compute QALYs can be measured in different ways, which can give different results [5–10].

These criticisms go some way to explaining why healthcare decisions based on cost/QALY thresholds are no longer recommended in the USA, as stated in the Patient Protection and Affordable Care Act [11]. Similarly, in Germany, the Institute for Quality and Efficiency in Health Care (IQWiG) has rejected the cost/QALY approach for ethical and methodological reasons [9, 13]. While many countries are taking other approaches, cost-utility studies expressed in cost/QALY are still recommended in the reference case of methods guidance for many health technology assessment (HTA) agencies in Commonwealth countries, e.g., the National Institute for Health and Care Excellence (NICE) in the UK, the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia, and the Canadian Agency for Drugs and Technologies in Health (CADTH) in Canada [14]. As a result, the number of published applied cost/QALY ('cost-utility' or 'cost-effectiveness') studies outweighs the number of publications evaluating the methodological limitations of the approach [6–8, 15]. Nevertheless, the fact that innovative treatments are often denied reimbursement in countries where the reference case demands a cost/QALY approach is leading a growing number of stakeholders, e.g., patient associations, industry, and healthcare professionals, to question the scientific relevance and consistency of the approach.

It is well-documented that different QALY estimates can be obtained by simply changing the utility assessment method. For example, a study conducted in patients with rheumatoid arthritis showed a statistically significantly difference in utility scores from generic preference-based measures [Health Utility Index (HUI), EuroQoL 5 Dimensions (EQ-5D) and Short-Form 6 Dimensions (SF-6D)] [16]. Other studies questioning the validity of these methods for estimating QALYs have found similar results [17]. Research has also shown that incremental cost-effectiveness ratio (ICER) threshold values can be misleading in that the cost/QALY approach has led to decisions that result in increased healthcare expenditure [18]. In addition, current approaches to eliciting time trade-off (TTO) values, and their use in economic evaluation, rest on specific assumptions about the way utility relates to time and health. While the assumptions and evidence of violations of them are discussed in the literature, the issues seem widely under-appreciated by those using and applying TTO in economic evaluation [19].

Acknowledging these methodological issues, the ECHOUTCOME (European Consortium in Healthcare Outcomes and Cost-Benefit Research) research project, funded by the European Commission under the 7th Framework Program, conducted a large European experiment with the objective of formally testing the scientific validity of the theoretical assumptions supporting the use of the QALYs as a reliable outcome for HTA. The proposed approach is based on the scientific reasoning proposed by Popper [20],

[1] If u(.) is a Neumanian-type utility function on E associated with an agent preferences, then, whatever the real numbers a and b such as $a > 0$, function $v(.) = au(.) + b$ is also a Neumanian-type utility function associated to the preferences of the same agent, and reciprocally. Then if u(.) is measured in a reference system S_1, v(.) can be measure in a reference system S_2 after a change of unit (role of coefficient a) and a change of origin (role of coefficient b).

who is known for his significant contribution regarding the robustness of scientific evidence in empirical sciences: a theory in the empirical sciences can never be proven, but it can be falsified, meaning that it can and should be scrutinized by decisive experiments. According to this author, if the outcome of an experiment contradicts the theory, one should refrain from ad hoc manoeuvres that evade the contradiction merely by making it less falsifiable. A scientific assumption is a proposition potentially refutable, which can be invalidated by one single example.

2 Methods and Data

2.1 Terminology and Experimental Study Design

Given the objective of this research, only a limited number of easily understandable health states were needed to verify the validity of the theoretical assumptions of the QALY. This is because the goal of this experiment was not to validate a new instrument or methodological approach for discriminating across health states amongst a large group of subjects, but, rather, to test the validity of the underlying assumptions supporting an existing methodological framework.

The study by Pliskin et al. [3] showed that if a set of conditions pertaining to an agent's preferences expressed by lotteries regarding life-years and quality of life are verified, then, if the lots of the lotteries are pairs (number of life-years and health state), it is possible to express the agent's preferences by an interval (also known as 'Neumannian') utility function [2]. This utility function should be equal to the product of an interval utility function on time duration 'life-years' and an interval utility function on 'health state'. Let T be considered as a set of life-year spans (excluding zero), and Z be a set of health states (excluding death). The following hypotheses form the basis of the multi-attribute utility theory:

Hypothesis H1	The preferences of an agent on T and on gamble whereby the lottery payoffs are respectively elements of T, Z, and T × Z, are 'Neumannian' and therefore defined by interval utility functions w, v, and u, respectively.
Hypothesis H2	Z and T are mutually independent utilities.
Hypothesis H3	The agent is weakly risk neutral on T according to the risk aversion measure described by Arrow-Pratt [21].
Hypothesis H4	The agent's rate γ to trade-off time is constant.

In order to assess the utilities (preferences), four health states in which only physical mobility was varied were selected for the experiment and included in set *Z*: (1) no physical disability (NPD); (2) limping (LMP); (3) walking with the assistance of a rollator (ROL); (4) confined to a wheelchair (WCH).

Three life-expectancy spans were arbitrarily chosen and deemed sufficient to validate the QALY multiplicative model, while providing a simple enough framework for the respondents. The three life spans (T1: 5 years; T2: 10 years; and T3: 15 years) thus constituted the set T: 5, 10, and 15 years, corresponding to the remaining life expectancy. The QALY multiplicative model would then be validated if the product of t with the utility of z would be equal to the utility function of the pairs (t,z).

The experiment was set to (1) measure the preferences of each health state z using lottery questions; (2) calculate the utility of the pairs (duration t, health state z) by performing the product of the utility of the health state by the time; (3) measure the preferences of the pairs (duration t, health state z) using lottery questions; and (4) compare the results as calculated (2) and measured (3).

Because the terminology 'standard gamble' has often been used to measure individuals' preferences regarding health states between death and perfect health in many health-decision analyses, the simple terms 'gamble' or 'lottery' will be used considering that death (as a potential 'lot') was excluded from the experiment in order to comply with Hypothesis H4. Nevertheless, the lottery questions are broadly similar to the standard gamble technique, which leads to a Neumannian-type measure under certain assumptions. This approach allows the expression of preferences in a manner consistent with the theoretical foundation of the QALYs.

2.2 Population Sample

This objective of this analysis, conducted in four European countries (Belgium, France, Italy, UK), was to test the validity of hypotheses H1, H2, H3, and H4. The study sample comprised individuals from academia to ensure that the participants would be educated enough to understand the questions and to express their preferences for different pairs (t,z) from combining four different selected health states with three different time spans.

A similar study conducted in France [8] allowed estimating the average difference (0.15) and standard deviation (SD) (0.18), leading to a sample size of 765 participants in order to discriminate a difference greater than 0.03 (alpha risk 5 % and beta risk 80 %). As a precautionary measure, for the current study, a minimum sample size of 300 participants per country (minimum total 1,200) was set so as to allow potential subgroup analyses, and would allow the rejection of a difference of 0.03 for an observed SD of 0.55.

2.3 Preference Assessment

A total of 18 gambles were included starting with four demonstration questions where participants were instructed to weigh risk and to express their preferences for various pairs (obtained from combining the health states and time spans selected for this experiment), in order to derive utility scores for each pair. For example, in one gamble participants are asked to assume that they have a severe disease with a life expectancy of 15 years. A treatment was proposed which gave a 50 % chance of increasing life expectancy to 24 years, and a 50 % chance of decreasing life expectancy to 6 years. The participants were then asked to express their preferences for three scenarios: (1) 'not taking the treatment'; (2) 'taking the treatment and accepting the risks'; and (3) 'indifferent between taking or not taking the treatment'. In a similar way, the 14 lotteries of the study provided key situations on which to evaluate the consistency of the multiplicative approach of the QALY.

An interviewer presented the notions of risks and lotteries to the participants using two lotteries as a warming up exercise. The choice of lottery questions was essential considering that the QALY model assumes that the utilities used are based on intervals (Neumannian) [3]. The risk lotteries facilitated the utility assessment for the following pairs comparing different time spans and health states:

- (15 years, WCH): 15 years confined to a wheelchair
- (10 years, ROL): 10 years walking assisted with a rollator
- (15 years, LMP): 15 years limping
- (10 years, NPD): 10 years without any physical disability.

When using a metric, it is necessary to define a system of reference. One may choose a range between 0 and 1. Very frequently, a utility of 0 represents the health state 'death', and a utility of 1 represents the health state of 'perfect health'. As in any formula based on a multiplicative model, the use of '0' as an origin raises many issues; e.g., "Would you prefer to be dead 10 years or 5 years?" In order to make the questions meaningful, the experiment excluded 'death' as the 0 origin but considered two reference systems S_1 and S_2. The reference system S_1 was defined by one origin: (10, WCH), namely $u_1(10, WCH) = 0$; and one unit: (15, NPD), namely $u_1(15, NPD) = 1$. The reference system S_2 was defined by one origin: (5, WCH), namely $u_2(5, WCH) = 0$, and one unit: (15, NPD), namely $u_2(15, NPD) = 1$.

For each individual, i, we also measured the utility, $v_i(.)$, of the health states LMP and ROL within the reference system having for its origin the health state WCH and for unit the health state NPD; by this, we get $v_i(WCH) = 0$ and $v_i(NPD) = 1$.

The QALY multiplicative model would be validated if the observed utilities of the couples (health state, time) would be equal to the calculated utilities 'time (health states)' whatever any affine transformations. Let H_i be a utility function associated with the same preference $\succcurlyeq_i$ of the individual i as u_{ki} within the reference system S_k, $k = 1, 2$, then $a_{ki} > 0$ and b_{ki} exist so that there is an affine relation $u_{ki}(t, z) = a_{ki} H(t,z) + b_{ki}$; we thus have to verify that for each individual i, $i = 1, \ldots n$, two real numbers existed a_{1i} and b_{1i}, $(a_{1i} > 0)$ and b_{1i} such that:

(1) $u_{1i}(15, WCH) = 15\ v_i(WCH)a_{1i} - b_{1i}$
(2) $u_{1i}(10, ROL) = 10\ v_i(ROL)a_{1i} - b_{1i}$
(3) $u_{1i}(15, LMP) = 15\ v_i(LMP)a_{1i} - b_{1i}$
(4) $u_{1i}(10, NPD) = 10\ v_i(NPD)a_{1i} - b_{1i}$
(5) $u_{1i}(5, NPD) = 5\ v_i(NPD)a_{1i} - b_{1i}$

or

(1) $[u_{1i}(15, WCH) - 15\ v_i(WCH)a_{1i} - b_{1i}] = 0$
(2) $[u_{1i}(10, ROL) - 10\ v_i(ROL)a_{1i} - b_{1i}] = 0$
(3) $[u_{1i}(15, LMP) - 15\ v_i(LMP)a_{1i} - b_{1i}] = 0$
(4) $[u_{1i}(10, NPD) - 10\ v_i(NPD)a_{1i} - b_{1i}] = 0$
(5) $[u_{1i}(5, NPD) - 5\ v_i(NPD)a_{1i} - b_{1i}] = 0$

where $u_{1i}(.,.)$ designates the utility function of the individual i for the pairs (t,z), measured in the frame of the reference system S_1. The five pairs tested in the frame of the two reference systems are represented in Fig. 1. Coefficient b is calculated using Eq. 1 ($b_{1i} = u_{1i}$ (15, WCH)). Eqs. 4 and 5 above allow the calculation of a coefficient a by difference eliminating coefficient b. Eqs. 2 and 3 can then be tested in the frame of the two reference systems S_1 and S_2 using known coefficients a and b estimated using Eqs. 1, 4, and 5 (see Fig. 1). Statistical analyses were performed independently using SPSS® version 19.0 (IBM Corp., Armonk, NY, USA) and the free prosoftware R. Comparisons between distributions of utilities' means were performed with *t* tests or ANOVA. All statistical tests were to be performed at the 5 % significance level.

In order for subject i from the target population to express a Neumannian-type preference on $T \times Z$, it is necessary that, if u_{1i} and u_{2i} are two Neumannian utilities representing this preference, coordinate points $(u_{1i}(t, z), u_{2i}(t,z))$ are linked with an affine relationships and can be tested by linear correlation. For each of the subjects, this hypothesis was tested by considering the following six points (t,z): (NPD, 15 years) (NPD, 10 years); (NPD, 5 years) (LMP, 15 years); (ROL, 10 years,), (WCH, 15 years). As a conservative approach, it was considered that when a linear correlation was strictly inferior to 0.8, utilities could not be considered as 'Neumannian' type.

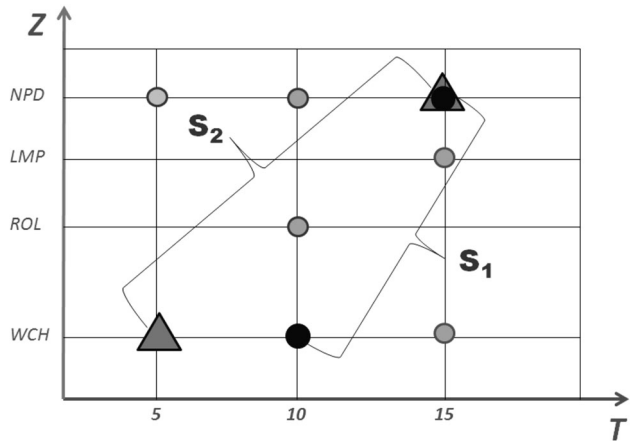

Fig. 1 Graphical representation of the five pairs (time, health state) represented by the *gray dots* and tested in the frame of the two reference systems S_1 (*black dots*) and S_2 (*gray triangles*), respectively, for the origin couples (10, WCH), (5, WCH) and for the unit couple (15, NPD). According to the multi-attribute utility theory, if preferences of the subject on health state and time (T × Z) are represented by Neumaniann-type utility functions noted as u and v, respectively, in reference systems S_1 and S_2, then: u(10, WCH) = 0, u(15, NPD) = 1 and v(5, WCH) = 0, v(15, NPD) = 1; there is an affine transformation with coefficient a (>0) and b such that u = av + b. *LMP* health state 'limping' *NPD* health state 'no physical disability', *ROL* health state 'walking with the assistance of a rollator', S_1 reference system S_1, S_2 reference system S_2, *T* set of time duration, *WCH* health state 'confined to a wheelchair', *Z* set of health states

3 Results and Analysis

Of 1,361 participants, a total of 1,250 participants responded to all 14 questions and provided probabilities different from 0. The remaining 111 were not considered eligible for inclusion in the analysis as they did not provide Neumannian-type preferences. Population characteristics are summarized in Table 1, and mean values and *t* tests are presented in Table 2.

In the frame of the reference system: v(NPD) = 1 and v(WCH) = 0, the mean utility of the health state 'LMP' was 0.777 (SD 0.168) and the mean utility of ROL was 0.688 (SD 0.185).

The analyses of the preferences of the subjects about the pairs (time span, health states) are presented in Table 3. In every situation, the utility of the pair u(t, z) was compared to the utility of each component of pair w(t) and v(z): u(z,t) = v(z) × w(t), then two coefficients a and b exist such as a > 0 and b as u(z,t)−[a × v(z) × w(t) + b] = 0.

t Tests rejected the hypothesis of the nullity of each equation related to the situations (15 years, LMP) and (10 years, ROL) comparing the observed and calculated utility values ($p < 0.0001$).

Only 70.9 % (886) of subjects expressed consistent preferences for the health states, namely: they prefer to live 15 years in a wheelchair rather than 10 years in a wheelchair; and, they also prefer to live 15 years in a wheelchair rather than 5 years in a wheelchair. In this sub-group, *t* tests rejected the hypothesis of the nullity of the equation comparing the observed and predicted utility values ($p < 0.0001$), confirming that the QALY multiplicative model is not valid in this sub-group. In order for subject i from the target population to express a Neumannian-type preference on T × Z, it is necessary that, if u_{1i} and u_{2i} are two Neumannian utilities representing this preference, the coordinate points ($u_{1i}(t, z)$, $u_{2i}(t,z)$) are linked with affine relationships and can be tested by linear correlation. From the initial population of 1,250 subjects, 410 (33 %) subjects were considered not having utilities as 'Neumannian' type, after testing potential linear correlation between preferences points. Again, in this second scenario these results suggest that the observed and calculated values are not equal ($p < 0.0001$), thus questioning the validity of the QALY multiplicative model.

Table 1 Study population gender and academic background per country

	Belgium	France	Italy	UK
Gender (%)				
Male	62.7	65.6	52.8	57.1
Female	37.3	34.4	47.2	42.9
Academic background (%)				
Sciences	2.1	59.8	4.7	7.6
Humanities	43.9	10.5	1.9	33.6
Health	15.6	24.1	8.1	5.6
Management	11.4	1.4	63.2	4.3
Economics	14.1	14.1	17.1	31.9
Technology	0.2	0.2	1.2	1.3
Other	12.7	12.7	3.9	15.6

4 Discussion

Welfare theory concepts are usually very difficult to validate in real life. Given that the QALY approach is known to be inconsistent and unreliable [15], it is particularly disconcerting that the main reasons for recommending the QALY are its apparent pragmatic approach and its extensive usage. Numerous cost/QALY studies are incorrectly labelled and referred to in the international literature as 'cost-effectiveness analyses'. However, the methodological controversy should not focus solely on terminology issues as however they are labelled their underlying assumptions and limitations remain the same.

In terms of limitations of the current analysis, criticisms could be expressed regarding the lottery technique used in this experiment in order to elicit individuals' preferences.

Table 2 Values and *t* tests testing the difference from 0 of Eqs. 2 and 3 in the frame of the two reference systems S_1 and S_2

t test difference from 0		*n* = 1,250			*n* = 886			*n* = 410		
		Mean	*t*	*p*	Mean	*t*	*p*	Mean	*t*	*p*
S_1	$[u_{1i}(10, ROL) - 10\ v_i(ROL)a_{1i} - b_{1i}] = 0$	−0.223	−19.30	0.000	−0.275	−20.03	0.000	−0.288	−13.58	0.000
	$[u_{1i}(15, LMP) - 15\ v_i(LMP)a_{1i} - b_{1i}] = 0$	−0.272	−16.47	0.000	−0.349	−17.09	0.000	−0.363	−11.36	0.000
S_2	$[u_{1i}(10, ROL) - 10\ v_i(ROL)a_{1i} - b_{1i}] = 0$	−0.185	−18.90	0.000	−0.236	−20.57	0.000	−0.249	−14.25	0.000
	$[u_{1i}(15, LMP) - 15\ v_i(LMP)a_{1i} - b_{1i}] = 0$	−0.243	−16.53	0.000	−0.318	−18.05	0.000	−0.329	−12.71	0.000

LMP limp, *ROL* walk with the assistance of a rollator

Table 3 Mean utilities (standard deviations in brackets) of the studied pairs (time, health states)

Utilities	Reference system S_1[a]	Reference system S_2[b]
5 years, NPD	0.676 (0.233)	0.727 (0.213)
10 years, NPD	0.877 (0.163)	0.898 (0.142)
15 years, NPD	1	1
15 years, LMP	0.744 (0.198)	0.846 (0.170)
10 years, ROL	0.570 (0.239)	0.730 (0.203)
5 years, WCH		0
10 years, WCH	0	
15 years, WCH	0.536 (0.243)	0.689 (0.239)

LMP limp, *NPD* no physical disability, *ROL* walk with the assistance of a rollator, *WCH* wheelchair

[a] Reference system S_1: origin (0) = 10 years in wheelchair; unit (1) = 15 years without physical disability

[b] Reference system S_2: origin (0) = 5 years in wheelchair; unit (1) = 15 years without physical disability

Lottery-type questions were selected in order to be consistent with the Von Neumann–Morgenstern utility theorem [1, 2]. This technique is based on the hypothesis that an individual is neutral to risk probabilities. Intuitively, this hypothesis seems very fragile, as most subjects are either highly risk adverse or risk seeking. The use of other preference elicitation methods such as TTO, however, would have created other issues[2]. In addition, the use of multi-attribute utility instruments (MAUI) would not have been appropriate for the objective of this research, as it has never been established that utilities derived from instruments such as EQ-5D are Neumannian in type.

Many publications [16, 17, 22–26] have confirmed that results obtained from one utility preference-elicitation method cannot be reproduced using another instrument and thus cannot be compared. However, comparison between utility values are done and benchmarked in the frame of 'league tables'. Richardson et al. stated that the general population would be strongly opposed to the use of QALY league tables for maximizing health benefits, including a formula which would 'abandon' any given patient because of the presumed cost ineffectiveness of a treatment [27].

In addition to the issue of different instruments generating different results, another challenge consists of deciding on the target population for the derivation of utility scores. For example, it is well-known that citizens, patients, caregivers, or healthcare providers all attribute different preference scores to the same health states and life expectancies. For these reasons, many authors have urged that considerable caution be exercised when interpreting cost/QALY for decision making because of the lack of comparability between methods, the use of inappropriate comparators, and the fact that the results cannot be generalized [28, 29].

As for any experimental study dealing with individuals' preferences, potential bias could have occurred, such as selection or cognitive bias. The consistency of the answers were tested against the QALY multiplicative theory and it was observed that the answers did not correspond with the theory assumptions, suggesting that the theory is not valid in this population, as well as potentially in many other groups. One could argue about the possibility of statistical error, which could occur in any study, and which should not be considered as a disproof of one theory. First, this experiment was designed to test the consistency of the answers provided according to the theory arguments, and not to study the distribution of the answers. However, even if there were some variability in the responses, statistical tests yielded *p* values lower than 0.0001, which disregards the possibility that the results could be considered erroneous. Concerning the possibility of cognitive bias, e.g., framing effect, the same technique was used to assess the utility of the pairs and the utility of each attribute, which would cancel such an effect when making the difference. Nevertheless, these biases could occur in all preference studies for obtaining utility scores for calculating QALYs.

The experimental results suggest that the theoretical underpinning of the QALY approach does not correspond to stated preferences of members of this population. It is

[2] In order to measure the utility of a health state z, so that $z_1 \succcurlyeq z$, where the utility of z_1 is equal to 1, the duration $t < T$ for which the pair (T,z) is indifferent to the pair (t, z_1) is assessed. It is then postulated that the utility of z, $v(z)$, is equal to t/T. The assumption that enables this result, and which is not often stated, is that the utility function on the pairs (t,z) is of the type $tv(z)$. It is precisely this specification that is at the origin of our interrogations.

thus fair to question the representativeness of the study population. The majority of respondents in this study were young adults in higher education, who were expected to attach different values to impaired quality of life, and relative death risks, from those of older adults. On the other hand, these 1,361 participants are more likely to understand lottery scenarios, although they did not come from a representative sample of the general population (or from any patient population), which could have led to potential selection bias. This could have created some 'selection effects' if the study objective had been to obtain utility values for estimating QALYs in this population and extrapolate these values to the general population; but the objective of the study was to 'test the consistency of the responses from the participants' regarding their own preferences (whatever the preferences may be). According to scientific reasoning, a theory is valid until one single counter-example invalidates this theory, otherwise it is considered a metaphysical doctrine [20]. Consequently, if the QALY approach would still be used after demonstrating the existence of one single counter-example (e.g., the results of this study), the QALY approach should then be considered a metaphysical doctrine (such as a common belief), and not as a scientific theory in the sense of Popper [20].

The fact that the responses in this group were more likely to provide valid answers to a preference-elicitation exercise contradicts the value judgments implicit in the decision-making framework of NICE, and provides evidence to refute the validity of such framework. Unless we adopt a point of view where individuals are compelled to conform to the model (and not the other way around), these findings suggest that the use of the QALY multiplicative model cannot be justified in healthcare decision making.

Hence, given that the many methodological inconsistencies of the QALY may lead to divergent results and to dramatically different health decisions, the convenience of the QALY approach in cost-utility analyses should not be seen as the main advantage because its inconsistencies may restrict access to innovative treatments in countries where cost/QALY is recommended in the reference case of HTA agencies.

Importantly, the limitations of the QALY assumptions have already been recognized by many authors, which in the 1990s led to the development of alternative 'synthetic' outcomes such as the healthy years equivalent (HYE), believed to be more robust [30–33]. Surprisingly, the literature questioning the validity of the QALY outcome appears to have been ignored by important HTA organizations or networks [such as NICE or the European Network for Health Technology Assessment (EUNETHA)]. There are several reasons that could explain why the QALY approach is still promoted despite its well-known and well-documented methodological limitations and flaws. The first reason is the large number of QALY empirical studies that have been published without questioning the underlying assumptions. A second reason is the intense economic activity and financial interests behind the QALY application studies, which are for the most part funded by pharmaceutical companies and often carried out by consulting companies who assist them in ensuring that the requirements of national HTA agencies are met in order to achieve reimbursement. Thirdly, as the QALY method can generate highly divergent results by slightly altering only one underlying assumption, this is useful to HTA agencies for challenging the results of cost-utility studies funded by the pharmaceutical industry. Moreover, the QALY method uses arbitrarily set thresholds which provide a useful means for the justification of cost-containment measures, and against which the level of reimbursement of innovative medicines is recommended or negotiated. In addition, the educational content of many university programs is developed by QALY supporters who have taught and continue to teach the QALY method to many students. Lastly, the influence of QALY advocates within the international communication and consultancy services of some important HTA agencies (e.g., NICE) contributes to the retention of the QALY approach in the reference case.

While many users and authors acknowledge that the QALY outcome is "not perfect" [34], they insist that it is the "best method available" in order to compare health interventions for resource-allocation decisions. In light of the evidence questioning the validity and reliability of the QALY outcome, and its inherent risks which may lead to erroneous health decisions, maintaining such a defensive attitude could denote a lack of rigorous attention towards patient populations, and ignores the development of evidence-based and more robust assessment methods, such as multi-criteria analyses, simulation models, Bayesian or Neuronal networks (provided that these techniques are not used to replicate the limitations of the multiplicative model in order to derive QALYs). Of course, each new technique should be evaluated for use in healthcare decision making and any limitations should be documented [35–38]. Because of the scientific complexity of the situations raised by HTA, there is currently no single alternative paradigm to propose at this time, but a spectrum of additional analytical techniques which could handle various outcomes including costs and health consequences, and which are not based on a simple multiplicative formula or associated methodological issues. It was not the purpose of this experiment to test the robustness of alternative techniques, which should be investigated using established methods. If evidence has been generated that a specific method is invalid, then it is not appropriate to continue using it. In

other industries, for example, the aviation industry, when faults are identified the model is usually withdrawn from use until it can be either fixed or replaced. With this analogy in mind, this questions the scientific and ethical acceptability of continuing to use the QALY on the basis that "…the QALY method is not perfect, but we do not have anything else to use" [39].

HTA agencies, stakeholders and researchers are urged to consider the implications of the ECHOUTCOME results, and to develop alternative methods to assess and compare health interventions and technologies. The rapidly changing healthcare environment, the interests for targeted therapies and personalized medicines, and the increasing economic pressures on healthcare systems underlines the importance of conducting robust HTA to assist resource-allocation decisions. The ECHOUTCOME consortium hopes that these findings will contribute to developing and implementing more robust methodologies for HTA that will allow further methodological development in this field, and that will contribute to establishing best practices for optimal and timely allocation of limited resources, for the benefits of patients, health systems, and societies in Europe.

Acknowledgments The authors gratefully acknowledge the suggestions of Louise Crathorne who kindly reviewed the manuscript.

Disclaimer This project has received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under agreement n° 242203. The sole responsibility for the content of this article lies with the authors and does not necessarily reflect the opinion of the European Union. The European Commission is not responsible for any use that may be made of the information contained therein.

Conflicts of interests No conflict of interests have been declared by the authors.

Authors' contributions AB, JPA, GD, and ML have designed the study. AB, AM-L, JPA, GD, and ML have carried out the methodological framework. ADW, J-CP, RT, AT, AM-L, and ML have organized the data collection. AB, DD, GD, and JPA have contributed to the writing of the manuscript. All authors have contributed to the interpretation of the results and to the revision of the manuscript. AB is acting as the overall guarantor.

References

1. Keeney RL, Raïffa H. Decisions with multiple objectives, preferences and value tradeoffs. Cambridge: Cambridge University Press; 1993.
2. Von Neumann J, Morgenstern O. Theory of games and economic behavior. 3rd ed. Princeton: Princeton University Press; 1953.
3. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. Oper Res. 1980;28:206–23.
4. Holmes D. Report triggers quibbles over QALYs, a staple of health metrics. Nat Med. 2013;19(13):248.
5. Lipscomb J, Drummond M, Fryback D, Gold M, Revicki D. Retaining, and enhancing, the QALY. Value Health. 2009;12(Suppl 1):S18–26.
6. Auray Beresniak JP. About the robustness of theoretical foundations of QALY. Estud de Econ Aplicada. 2006;24–3:685–96.
7. Beresniak A, Russell AS, Haraoui B, et al. Advantages and limitations of utility assessment methods in rheumatoid arthritis. J Rheumatol. 2007;34(11):2193–200.
8. Duru G, Auray JP, Béresniak A, et al. Limitations of the methods used for calculating quality-adjusted life-year values. Pharmacoeconomics. 2002;20(7):463–73.
9. Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.
10. Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. Annu Rev Public Health. 2000;21:587–611.
11. The Patient Protection and Affordable Care Act (PPACA). 2010. PL111-148. p. 3–23.
12. IQWiG. General methods for the assessment of the relation of benefits to cost. Cologne: German Institute for Quality and Efficiency in Health Care (IQWiG); 2009. p. 19.
13. Neumann PJ. What next for QALYs? JAMA. 2011;305(17):1806–7.
14. Guidelines for the economic evaluation of health technologies: Canada. 3rd ed. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2006.
15. McGregor M, Caro JJ. QALYs: are they helpful to decision makers? Pharmacoeconomics. 2006;24(10):947–52.
16. Marra CA, Woolcott JC, Kopec JA, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60(7):1571–82.
17. Conner-Spady B, Suarez-Almazor ME. Variation in the estimation of quality adjusted life-years by different preference-based instrument. Med Care. 2003;41:791–801.
18. Gafni A, Birch S. Incremental cost-effectiveness ratios (ICERs): the silence of the lambda. Soc Sci Med. 2006;62(9):2091–100.
19. Buckingham KJ, Devlin NJ. A note on the nature of utility in time and health and implications for cost utility analysis. Soc Sci Med. 2009;68(2):362–7.
20. Popper K. Conjectures and refutations: the growth of scientific knowledge. New York: Methodology Basic Books; 1962. p. 428.
21. Cohen MD. Risk aversion concept in expected and non-expected-utility models. Geneva Pap Risk Insur Theory. 1995;20:73–91.
22. Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making. 2010;30(1):58–67.
23. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002;11(5):447–56.
24. Nord E, Menzel P, Richardson J. Multi-method approach to valuing health states: problems with meaning. Health Econ. 2006;15(2):215–8.
25. Ariza-Ariza R, Hernández-Cruz B, Carmona L, et al. Assessing utility values in rheumatoid arthritis: a comparison between time trade-off and the EuroQol. Arthritis Rheum. 2006;55(5):751–6.
26. Jorstad IC, Kirstiansen IS, Uhlig T, et al. Performance of four utility measures in 1041 patients with rheumatoid arthritis (RA): well correlated but differing widely in valuing health states. Arthritis Rheum. 2005;52(9):S660.
27. Richardson J, Sinha K, Iezzi A, et al. Maximising health versus sharing: measuring preferences for the allocation of the health budget. Soc Sci Med. 2012;75(8):1351–61.
28. Mauskopf J, Rutten F, Schonfeld W. Cost-effectiveness league tables: valuable guidance for decision makers? Pharmacoeconomics. 2003;21(14):991–1000.

29. Gerard K, Mooney G. QALY league tables: handle with care. Health Econ. 1993;2(1):59–64.
30. Hall J, Gerard K, Salkeld G, et al. A cost utility analysis of mammography screening in Australia. Soc Sci Med. 1992;34(9):993–1004.
31. Gafni A, Birch S, Mehrez A. Economics, health and health economics: HYEs (healthy-years equivalent) versus QALYs (quality-adjusted live-year). J Health Econ. 1993;12(3):325–39.
32. Johannesson M. The ranking properties of healthy-years equivalents and quality-adjusted life-years under certainty and uncertainty. Int J Technol Assess Health Care. 1995;11(1):40–8.
33. Birch S, Gafni A, Markham B, et al. Health years equivalents as a measurement of preferences for dental interventions. Community Dent Health. 1998;15(4):233–42.
34. Dreaper J. Researchers claim NHS drug decisions 'are flawed'. BBC News, 2013 Jan 24. http://www.bbc.com/news/health-21170445. Accessed 4 Sep 2014.
35. Thokala P, Duenas A. Multiple criteria decision analysis for health technology assessment. Value Health. 2012;15(8):1172–81.
36. Beresniak A, Dupont DM, Becker JC, et al. Interest of modelling in rheumatoid arthritis. Clin Exp Rheumatol. 2012;30(4 Suppl 73):S96–101.
37. van der Wilt GJ, Groenewoud H, van Riel P. Bridging the gap between aggregate data and individual patient management: a Bayesian approach. Int J Technol Assess Health Care. 2011;27(2):133–8.
38. Rosas MA, Bezerra AF, Duarte-Neto PJ. Use of artificial neural networks in applying methodology for allocating health resources. Rev Saude Publica. 2013;47(1):128–36.
39. Williams A. QALYS and ethics: a health economist's perspective. Soc Sci Med. 1996;43(12):1795–804.