VALUE IN HEALTH 14 (2011) 282–290

available at www.sciencedirect.com

SCIENCE DIRECT®

journal homepage: www.elsevier.com/locate/jval

Outcomes Assessment

Imputing QALYs from Single Time Point Health State Descriptions on the EQ-5D and the SF-6D: A Comparison of Methods for Hepatitis A Patients

Jeroen Luyten, MSc[a,], Christiaan Marais, MSc[a], Niel Hens, MSc, PhD[a,b], Koen De Schrijver, MD, PhD[c,d], Philippe Beutels, MSc, PhD[a]*

[a] *Centre for Health Economics Research & Modeling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium*

[b] *Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium*

[c] *Department Control of Infectious Diseases, Flemish Public Health Authorities, Antwerp, Belgium*

[d] *Epidemiology and Social Medicine, University of Antwerp, Antwerp, Belgium*

Keywords:
Health benefits
Health valuation
Quality of life
Recall period
Scoring algorithms

ABSTRACT

Objectives: To explore the impact of applying different non-standardized analytical choices for quality of life measurement to obtain quality-adjusted life years (QALYs). In addition to more widely discussed issues such as the choice of instrument (e.g. EQ-5D or SF-6D?) researchers must also choose between different recall periods, scoring algorithms and interpolations between points of measurement.

Methods: A prospective survey was made among 114 Belgian patients with acute hepatitis A illness. Using non-parametric tests and generalized linear models (GLM's), we compared four different methods to estimate QALY losses, two based on the EQ-5D (administered during the period of illness without recall period) and two based on the SF-6D (administered after illness with 4 weeks recall period).

Results: We found statistically significant differences between all methods, with the non-parametric SF-6D-based method yielding the highest median QALY impact (0.032 QALYs). This is more than five times as high as the EQ-5D-based method with linear health improvement, which yields the lowest median QALY impact (0.006 QALYs).

Conclusions: Economic evaluations of health care technologies predominantly use QALYs to quantify health benefits. Non-standardised analytical choices can have a decision-changing impact on cost-effectiveness results, particularly if morbidity takes up a substantial part of the total QALY loss. Yet these choices are rarely subjected to sensitivity analysis. Researchers and decision makers should be aware of the influence of these somewhat arbitrary choices on their results.

Funding: Financial support for this study came from "SIMID," a strategic basic research project funded by the Institute for the Support of Innovation through Science and Technology in Flanders (IWT), project number 060081; and from the Belgian Health Care Knowledge Centre. The authors have no other financial relationships to disclose.

* *Address correspondence to:* Jeroen Luyten, MSc, Centre for Health Economics Research and Modeling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Campus Drie Eiken, Universiteitsplein 1, 2610, Antwerpen, Belgium.

E-mail: Jeroen.luyten@ua.ac.be.

1098-3015/$36.00 – see front matter Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2010.10.004

Introduction

To compare health benefits across different medical treatments the quality-adjusted life year (QALY) has become a widely used summary health outcome comprising both morbidity and mortality effects. Nevertheless QALYs and cost-utility analyses continue to be seen by some as a controversial instrument for policy analysis [1–4]. One reason for criticisms is that there is no single objective standardised tool to obtain the morbidity impact of disease and different methods appear to generate different results [5–7]. Amongst the many instruments available to estimate the quality of life (QoL) associated with living in different health states, the EQ-5D [8] and the SF-6D [9–11] are both popular multi-attribute classification systems. These surveys typically generate a single time point utility value for each relevant health state, which is then used to value an average patient's health path. There are several important steps in this process, all with their specific pitfalls and problems. The health state of interest should be defined and described and it should be valued. This is done by asking questions across several dimensions that are assumed to contribute to enjoying a healthy life. Respondents summarise their position on the ordinal scale of each dimension. In order to be a valid measure of health, these dimensions should exhaustively cover all relevant dimensions of health-related QoL. The EQ-5D covers the dimensions of mobility, self care, usual activities, pain/discomfort and anxiety/depression. The SF-6D distinguishes between physical functioning, role limitations, social functioning, pain, mental health and vitality. The choice of respondents (most often patients, health practitioners or the general population) can be of substantial influence on such valuations [12]. In this article we consider the patient to be the most appropriate judge of his/her health experience. Ideally a survey, which asks for the patient's current health state, should be taken repeatedly during the period of illness, in order to estimate the loss in health related QoL. However, a day-to-day evaluation of a sufficiently large patient group is usually impractical. A common approach is therefore to have a survey filled in once (or multiple times) and consequently to assume a linear change between the QoL at the measurement point(s) and perfect health [13]. An alternative method consists of measuring patients' health status during a specified recall period. Both methods aim to estimate the same health path and it is therefore important to establish to which extent this is the case, and whether the choice for one of these methods introduces an additional source of uncertainty in economic evaluation.

We aimed to explore the effects of different methods for estimating the impact of an acute illness on the QoL valued by patients. Figure 1 illustrates five different methods to calculate the QoL impact for a patient experiencing 4 weeks of illness. Three of these methods are based on a current state health survey such as the EQ-5D and two are based on a survey with a specified recall period such as the SF-6D. At *T1* patients were surveyed regarding their health state. Their health path is consequently estimated using either of the following assumptions: the patient has a constant QoL for the duration of the illness, there is a linear improvement in QoL, or the marginal improvement in QoL increases every day (an exponential health improvement). The alternative for these interpolations is to ask the patient after recovering, at *T2*, to recall his/her health state over the entire disease episode. In Figure 1 these different methods yield the following contingent health burden: the rectangle ABDE for a constant health state, the triangle ABE for a linear health improvement and the area between the points ABEH for an exponential change. Assuming that respondents value their health state between time *T1* and *T2*, at "F", their retrospective health loss is illustrated as the rectangle ABCF in Figure 1. This rectangle can be obtained with two different scoring algorithms.

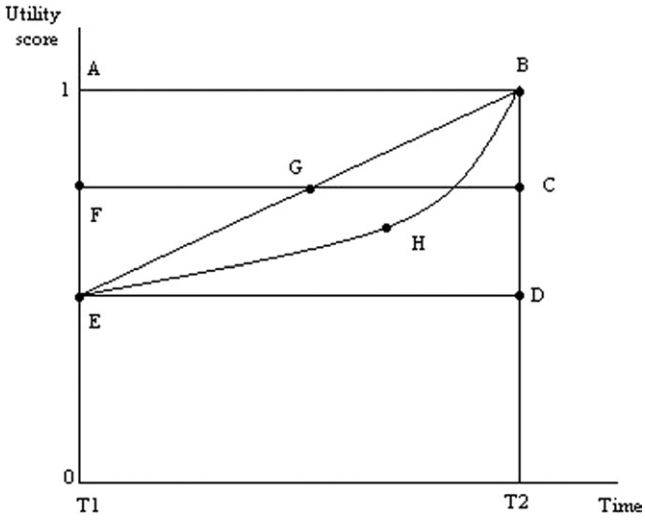

Fig. 1 – Graphical representation of the different health measurement methods for a patient experiencing 4 weeks of illness. The X-axis represents the duration of illness and the Y-axis the health utility score. The health burden for the EQ-5D is taken at *T1* and estimated by the area between ABDE for a constant health state, the triangle ABE for a linear health improvement and ABEH for an exponential change. The SF-6D is taken at *T2* and the estimated health burden is the rectangle ABCF (for both scoring algorithms).

In this study we empirically investigate whether these conceptually different methods lead to significantly different QALY loss estimates. To test this we focused on hepatitis A, which—in contrast to hepatitis B or hepatitis C—can only give rise to acute disease. Hepatitis A affects the liver, and may cause mild to severe illness (exceptionally leading to fulminant hepatitis and death) for on average 2 to 4 weeks [14–16]. We compared the QALY loss estimates of patients resulting from four different but conceptually defensible measurement methods. Two methods are based on the EQ-5D (taken during their illness, without a recall period): one with a linear health improvement and one with a constant health state during the period of illness. We will not discuss the exponential method since the results would always fall between the other two methods. The other two methods are based on the SF-6D (taken after illness, with a recall period of 4 weeks): one using the commonly used parametric scoring algorithm [9] and one using the more recent non-parametric scoring algorithm [17,18].

Methods

Data

Hepatitis A is a notifiable disease in Belgium, implying that physicians and laboratories should communicate information about new hepatitis A patients to the health inspection services.

Three interlinked surveys were administered prospectively to all hepatitis A patients about whom notification was provided to the Flemish public health services from February 1, 2008 to January 31, 2009. First, the EQ-5D, including a visual analogue scale (VAS), was sent to the patient immediately after notification of illness (approximately 1 week after the first symptoms appeared) and requested the patient to value his/her health state at that time. The SF-12 was sent to the same patients approximately 3 weeks later, and patients were asked to describe their health state during the preceding 4 weeks. A third survey was administered immediately after the SF-12, in order to collect information on the number of days of symptomatic illness the patient had experienced from the hepatitis A episode as well as on the nature of the symptoms experienced, and health care consumption associated with the period of illness [19]. All three surveys were linked to an individual patient by a unique anonymous code.

QALY calculation

The SF-12 cannot be directly translated into an utility score, but Brazier et al. [9] reduced the number of questions, creating a shorter survey (the SF-6D) that can be used for calculating QALYs. Utility scores from the EQ-5D and the SF-6D were calculated using respectively the UK York time trade-off (TTO) algorithm [20] and the UK Sheffield standard gamble (SG) algorithm (we use both the parametric [9] and the non-parametric [17] versions of the latter).

To transform these incomparable utilities into comparable QALY loss estimates, a time dimension needs to be introduced. Assuming that the utility weight associated with non-affected health states in these patients corresponds to a value of 1, the parametric SF-6D utility score, representing the patient's health during the entire month of illness, can be transformed into QALY loss estimates as follows:

$$\text{QALYloss SF-6D} = 1 - \frac{(\text{SF-6D score} + 11)}{12}$$

The QALY loss from the same SF-6D-survey but with use of the more recent non-parametric method is calculated in a similar manner but uses the algorithm described by Kharroubi et al. [21].

$$\text{QALYloss SF-6DNP} = 1 - \frac{(\text{SF-6DNP score} + 11)}{12}$$

The terms SF-6D and SF-6DNP will be used to refer to the QALY loss resulting from using the parametric and the non-parametric scoring algorithm for the SF-6D-survey, respectively.

To calculate QALYs from the EQ-5D we need to make an assumption about the evolution of the patient's health experience. In this study we only consider a constant and a linear progression (referred to as EQ-5DConstant and EQ-5DLinear respectively hereafter), and potential alternative improvements in adverse health experience (e.g., exponential) are not explored.

$$\text{QALYloss EQ-5DConstant} = \frac{\text{days of illness} \times (1 - \text{EQ5DScore})}{365}$$

$$\text{QALYloss EQ-5DLinear} = \frac{1}{2}\frac{\text{days of illness} \times (1 - \text{EQ5DScore})}{365}$$

Figure 1 shows a graphical representation of the intuitive meaning of these equations, with $1 - \text{EQ5DScore}$ being represented by the line AE and $\frac{\text{Days of illness}}{365}$ being represented by $T2 - T1$.

Analyses

The responses from the survey were analysed using SPSS 15.0 (Chicago, Illinois), R 2.8.1, @Risk (Ithaca, New York), and MS Excel 2003 (Redmond, WA) software. Missing data were excluded list wise. We tested the potential bias for excluding data by calculating the correlation between the VAS score and the number of questions answered and found this to be insignificant ($P = 0.362$). Summary statistics of the variables obtained in the survey were produced. The QALY loss predicted by the EQ-5D and the SF-6D was compared using a paired Wilcoxon signed rank test. The QALY loss estimated by the SF-6D was regressed onto that of the EQ-5D to enable analysis of the relationship between these two measures. The natural logarithm of the ratio of the QALY loss predicted by the SF-6D and the EQ-5D was taken to further investigate the difference in these measures. A linear regression model was used to determine and compare the significant explanatory variables that affect the QALY loss predicted by the EQ-5D and the SF-6D. The sensitivity of the QALY loss measures with respect to the explanatory variables was compared using a generalized linear model (GLM) with interaction term [22]. This approach is similar to that followed by Kontodimopoulos et al.[5] Relationships between the five dimensions of the EQ-5D and the six dimensions of the SF-6D were analysed with Kendall's tau rank correlation coefficient. A 95% level of confidence was used.

Results

The total sample size was 161. Of the entire sample 111 patients completed the number of days of illness and all questions on the EQ-5D, both necessary to obtain a valid QALY loss estimate from the EQ-5D. The SF-6D survey was filled in by 114 respondents. For 96 patients we were able to obtain a valid QALY-estimate for both surveys. Table I shows the sample characteristics of the discrete variables.

Table 1 – Sample characteristics.

	Percent (n)
Gender	
Male	52.2% (72)
Female	47.8% (66)
Education	
Primary education	18.6% (24)
Secondary education	41.1% (53)
Tertiary education	27.9% (36)
Other education	12.4% (16)
Working status	
Student or child	38.5% (50)
Full-time & part-time employed	50.8% (66)
Unemployed	6.2% (8)
Other employment	4.6% (6)

Comparison of QALY loss as predicted by EQ-5D and SF-6D

Table II provides summary statistics of the continuous variables in the sample, which includes the QALY loss estimates from the four measurement methods. Due to the asymmetric nature of the EQ-5D and SF-6D measures, the median QALY loss values offer the best basis for comparison between these two instruments. The SF-6D based method resulted in median QALY loss values that were more than twice that of the EQ-5DConstant and nearly five times that of the EQ-5DLinear. The SF-6DNP yielded median values more than five times those of the EQ-5DLinear and more than 2.5 times those of the EQ-5DConstant. The standard deviations indicate that the range of health outcomes is smaller for the SF-6D-based method than for the EQ-5D-based methods. The difference in spread and location can be seen from the descriptive statistics in Table II. As suggested by Kharoubi et al. [17], the differences between outcomes from the SF-6D and SF-6DNP were small. The comparisons between the EQ-5D and the SF-6D survey are consequently similar for both algorithms. Therefore we focus mainly on the parametric SF-6D score and expand our results with the non-parametric method where necessary.

Non-parametric tests indicate a statistical difference between the median QALY loss predicted by the EQ-5DLinear and the SF-6D ($P < 0.0001$). Similarly, there is a significant difference between the median QALY loss predicted by the EQ-5DConstant and the SF-6D ($P = 0.001$). Also, the EQ-5D-Constant, the EQ-5DLinear and the SF-6D outcomes are all significantly different from the SF-6DNP ($P < 0.0001$ for all comparisons). A scatter plot of the coupled health surveys, shown in Figure 2, compares the outcomes of the QALY loss measures for the EQ-5DConstant and the SF-6D. Theoretically, all observations on the scatter plot should be on the 45° reference line if the two QALY measures produce the same QALY loss estimates. This is rarely the case. In line with other studies, the EQ-5D is more responsive for patients with apparently more severe illness, but relatively insensitive for patients in milder disease states, whereas the SF-6D is also sensitive for apparently milder illness [10]. Patients, who indicated greater disease severity on the EQ-5D, tend to have a lower score on the SF-6D. Patients with lower QALY loss on the EQ-5D tend to generate higher health burdens on the SF-6D. A number of patients who indicated no health impact on the EQ-5D during the first week of illness indicated an adverse health impact on the SF-6D. The EQ-5DLinear by definition results in half the QALY loss compared to the EQ-5DConstant and consequently only results in a higher QALY loss than the SF-6D in patients who indicate severe levels of adverse health impact on the EQ-5D. The QALY loss estimated by the SF-6DNP is not shown on Figure 2 since its relationship is similar to that of the EQ-5D and the SF-6D. We found a linear relationship between SF-6D and SF-6DNP with greater values of QALY loss for the SF-6DNP as shown in Table II.

We fitted the QALY loss resulting from the SF-6D as a function of the EQ-5D. The best fit was a third order polynomial (adjusted $R^2 = 0.343$) and is also shown in Figure 2. The estimated regression equation is as follows:

$$\widehat{\text{SF-6D}} = 0.02 + 0.685\ \text{EQ5DConstant} - 7.057(\text{EQ-5DConstant})^2 + 17.413(\text{EQ-5DConstant})^3 \quad (1)$$

The regression line between the EQ-5DLinear and the SF-6D has a similar shape to that of the EQ-5DConstant and the SF-6D. The turning point for the EQ-5DConstant is 0.063 and for the EQ-5DLinear it is 0.032. Therefore it is estimated that the SF-6D-based method is an increasing function of the EQ-5DConstant if EQ-5DConstant < 0.063 and decreasing otherwise (we ignore the second turning point since it is

Table 2 – Descriptive statistics.

	Number of valid responses	Mean	Median	Min	Max	Standard deviation
Age	138	27.71	25.50	0	74	17.192
VAS score	127	66.31	67.00	20	100	19.765
No. of symptoms	161	5.60	6.00	0	11	3.12
Days of illness	133	17.82	14.00	0	65	12.32
EQ-5DConstant QALY loss	111	2.55×10^{-2}	1.22×10^{-2}	0	1.73×10^{-1}	3.33×10^{-2}
EQ-5DLinear QALY loss	111	1.28×10^{-2}	6.10×10^{-3}	0	8.65×10^{-2}	1.67×10^{-2}
SF-6D QALY loss	114	2.80×10^{-2}	2.98×10^{-2}	0	4.68×10^{-2}	1.08×10^{-2}
SF-6DNP QALY loss	114	3.00×10^{-2}	3.27×10^{-2}	0	4.66×10^{-2}	1.02×10^{-2}

Note: These statistics were obtained for all respondents who indicated a response to the specific question. The minimum number of days of illness and symptoms can be attributed to an asymptomatic patient. They are also notified to the health inspection services through laboratory tests of family members of an infected patient. This is also part of the clinical image of hepatitis A virus and, therefore, should be included.
QALY, quality-adjusted life year; VAS, visual analogue scale.

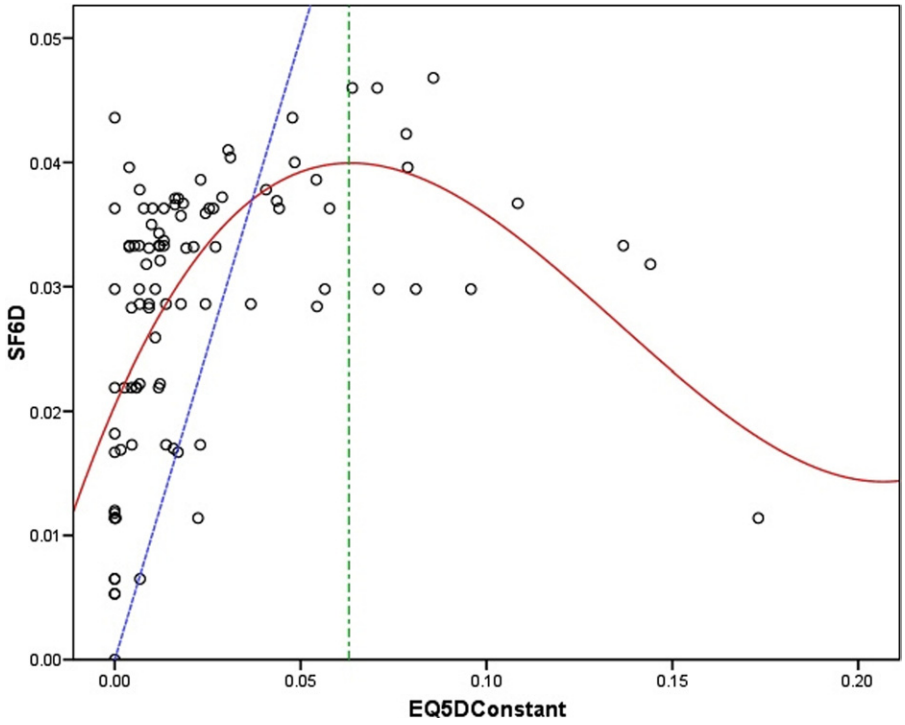

Fig. 2 – Fitted third order polynomial estimating the QALY loss resulting from the SF-6D as a function of the EQ-5D. Dots are observations, red line is third order polynomial fitted to the data, blue line is the 45° reference line, and green line is reference line at the turning point of the polynomial.

outside the range of the EQ-5DConstant). We found similar results for the SF-6DNP estimate. The quadratic equation was the best fit with an adjusted R^2 of 0.369 and the estimated curve had the same turning point.

The differences between the EQ-5D and SF-6D QALY loss estimates were further investigated by calculating the ratio of the natural logarithm of the EQ-5D and SF-6D QALY loss estimates. A cluster emerged for respondents that indicated perfect health on the EQ-5D and imperfect health on the SF-6D. This clustering was confirmed by a two-step cluster analysis. The same cluster emerged for the SF-6DNP.

Figure 3 summarizes the distribution of the responses on the different pairs of similar health-dimensions on both instruments [23]. The EQ-5D dimensions reflect the health experience as it is measured during the illness (usually at the start). The SF-6D dimensions reflect how the illness was experienced in retrospect, 4 weeks after onset of illness. Patients indicate more intermediate levels of health problems on the SF-6D survey, especially on the dimensions 'role limitations' and 'anxiety/depression' (>80% vs. ≤20%). The 'vitality' dimension is lacking on the EQ-5D but is shown to be an important aspect of hepatitis A illness on the SF-6D survey. The percentage of respondents indicating the most severe levels of problems also differs between the surveys. Fifty-one percent indicated most severe problems with 'role limitations' on the SF-6D vs. 5% in the EQ-5D. 'Usual activities' was a stronger

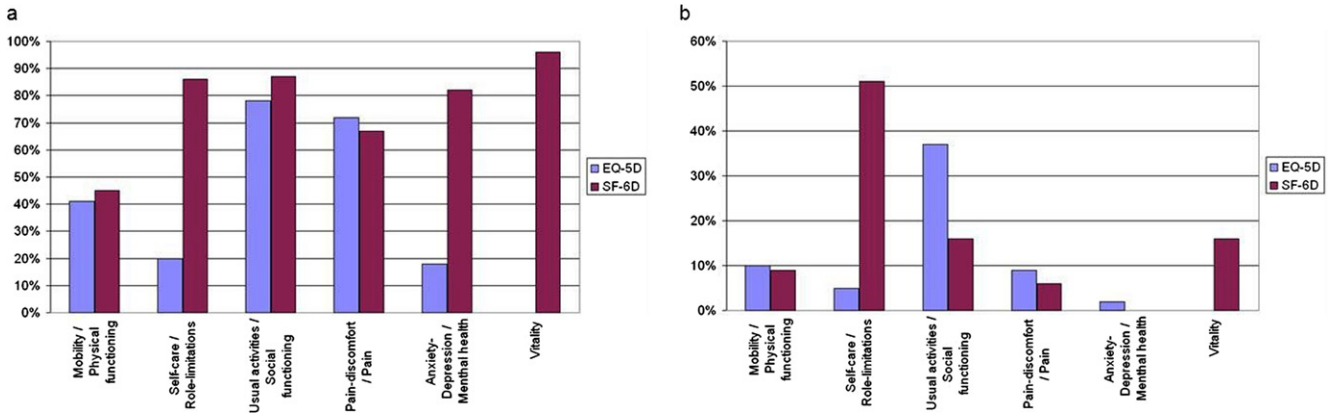

Fig. 3 – (A) Comparison of dimensions for patients indicating any problem. (B) Comparison of dimensions for patients indicating severest level of problems.

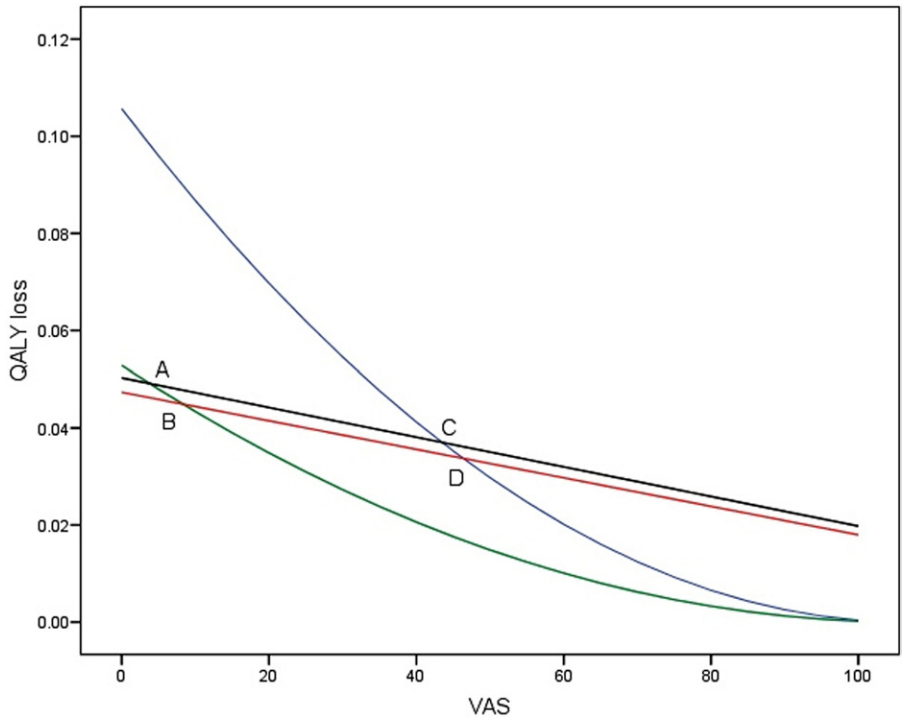

Fig. 4 – Regression lines of the EQ-5DConstant, EQ-5DLinear and SF-6D onto visual analogue scale (VAS). Blue line, EQ-5DConstant; green line, EQ-5DLinear; red line, SF-6D; black line, SF-6DNP. Intersection points: A, EQ-5DLinear and SF-6DNP (VAS score = 4.95); B, ED-5DLinear and SF-6D (VAS score = 7.17); C, EQ-5DConstant and SF-6DNP (VAS score 43.35); D, EQ-5DConstant and SF-6D (VAS score = 46.11).

contributor to the patients health state on the EQ-5D than 'social functioning' on the SF-6D.

The relationship between the dimensions of the EQ-5D and the SF-6D survey were analysed using Kendall's tau rank correlation coefficient. This indicated that all of the dimensions are correlated to each other with the exception of 'pain/discomfort' and 'self-care' ($P = 0.197$), 'vitality' and 'anxiety/depression' ($P = 0.074$); and 'social functioning' and 'self-care' ($P = 0.057$) being uncorrelated.

Significant predictors of QALY loss

A multiple regression model indicated that the VAS score and number of symptoms were significant predictors of the QALY loss estimated by the EQ-5D and the SF-6D. Multicollinearity in these linear models was not significant despite a degree of correlation (Pearson correlation coefficient = −0.2) between VAS and the number of symptoms. The cost of treatment, number of days of illness (only tested for the SF-6D), gender, smoking status, previous experience with serious illness, and age of the patient were found to be non-significant.

As can be expected, there is a negative relationship between the VAS score and the QALY loss for both the EQ-5D and the SF-6D. Both QALY loss measures indicated a positive relationship with the number of symptoms. The same results were obtained for the SF-6DNP as for the SF-6D.

The sensitivity of the QALY loss measures to the VAS score and the number of symptoms was analysed by fitting two GLMs. The GLMs had the QALY loss estimates as dependent variable and an intercept and three dummy variables for the four possible QALY loss estimators. Furthermore, the first model had three interaction terms for the QALY measures and the VAS score, and the second model had three interaction terms for the QALY measures and the number of symptoms. A significant coefficient for an interaction term indicates a threshold for the VAS/number of symptoms such that below this threshold, the QALY loss estimated by one of the methods will be greater than that of the other methods and an opposite relationship above the threshold.

The interaction between the VAS score and the QALY loss measure was significant ($P < 0.001$ for EQ-5DLinear*VAS; $P < 0.001$ for SF-6D*VAS; $P < 0.001$ for SF-6DNP*VAS), but not between the VAS score and the number of symptoms ($P = 0.123$ for EQ-5DLinear*Number of symptoms; $P = 0.140$ for SF-6D*Number of symptoms; $P = 0.098$ for SF-6DNP*Number of symptoms). This suggests the existence of a threshold for VAS. Below this threshold the QALY loss estimated by the EQ-5D is greater compared to the one based on the SF-6D, and vice versa above the threshold. Four linear regression equations were fitted to explore the relationship between the QALY loss estimates as predicted by the three measures and the VAS score. These regression lines are shown in Figure 4, which indicates that for some VAS scores the QALY loss estimates as predicted by the EQ-5DConstant are larger than those of the SF-6D and SF-6DNP, and similarly for the EQ-5DLinear. The threshold between EQ-5DLinear and SF-6D is estimated at 7.17 and at 46.11 between EQ-5DConstant and SF-6D. Therefore, it is estimated that for patients with a VAS score less than 7.17, a greater QALY loss will be predicted by the EQ-

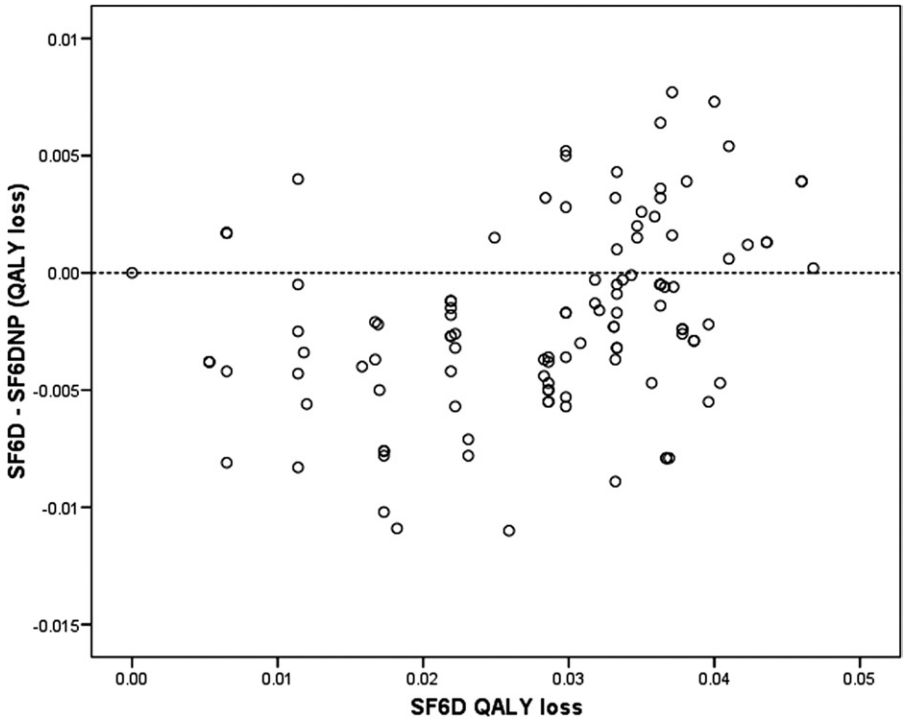

Fig. 5 – Difference of the resulting QALY loss with the parametric (SF-6D) and the nonparametric (SF-6DNP) scoring algorithm for the responses to the SF-6D survey.

5DLinear than by the SF-6D, but for VAS scores greater than 7.17 the QALY loss estimates from the SF-6D will be greater than those of the EQ-5DLinear. A similar conclusion can be made for the threshold between the EQ5DConstant and the SF-6DNP. The thresholds between the EQ5DLinear and the SF-6DNP and between the EQ5DConstant and the SF-6DNP are 4.95 and 43.35, respectively. It is estimated that such thresholds do not exist for the number of symptoms due to the insignificant interaction between the QALY loss measure and the number of symptoms.

Comparison of QALY loss as predicted by the SF-6D with parametric and non-parametric scoring algorithm

Figure 5 shows the difference in QALY loss obtained with the two scoring algorithms for the SF-6D in hepatitis A patients. In our sample the utility values ranged [0.439; 1] in the SF-6D and [0.4406; 1] in the SF-6DNP. This does not allow us to compare the poorest health states in the entire spectrum of possible health states on the SF-6D survey since the lowest possible utility is 0.257 and 0.203, respectively. On the opposite side of the scale, with the best health states, we found higher QALY loss in the non-parametric method. For average health states the results seem to be more equally distributed.

Discussion

We compared the QALY loss estimates produced by measurement methods based on the EQ-5D and the SF-6D by means of an interlinked survey from hepatitis A patients in Belgium. The four different health valuation methods produced significantly ($P < 0.05$) different results. Expressed as a median from the population sample, the SF-6D generated QALY losses that were about two to five times those obtained with the EQ-5D, assuming a constant QALY loss or a linear improvement in QALY loss over the illness period, respectively. All methods were able to identify groups of patients with a different health state according to criteria with a known relationship to health-related QoL such as the VAS and the number of symptoms. Both surveys yielded higher values of QALY loss for patients with an apparently more severe form of illness, thus supporting their validity. However, the magnitude of the difference between severely ill and mildly ill patients differed between the instruments. The EQ-5D resulted in higher values of QALY loss than the SF-6D for patients with more severe illness (as indicated by a low VAS score and a relatively high QALY loss on both surveys). The SF-6D generated greater QALY losses for the majority of patients with only minor or moderately severe illness. Most patients (93.33%) who indicated no health burden on the EQ-5D indicated a positive adverse impact on the SF-6D. On average the SF-6DNP had a similar relation to the EQ-5D as the SF-6D but yielded higher QALY loss. The differences in our sample between the SF-6D with parametric and with non-parametric scoring algorithm are consistent with the finding of Kharroubi et al. [17] who state that the parametric method overestimates the utility of patients in superior health states, thereby resulting in lower QALY loss estimates than the non-parametric method.

Our study highlights the variety of analytical choices that underlie QALY-calculations in practice, and indicates the magnitude of differences in outcome they may generate. Be-

cause of multiple contributory factors, which are associated with the choice and use of a particular standard instrument (i.e. the recall period, the algorithm, and the requirement to infer many time points from a single one or a small number of time points), our study design could not quantify the effect of a single of these associated factors. Future research may estimate the effect of a single factor by creating more subgroups in which more combinations of these associated factors can be kept constant.

Nonetheless several explanations can be proposed for the observed differences between the EQ-5D and the SF-6D survey. First, there is a difference in recall period. For the SF-6D survey, patients recall their health state over a 4-week period, while the EQ-5D survey captures a patient's current state. For hepatitis A, with an expected duration of clinical symptoms during 3 to 4 weeks (and in our sample up to 65 days [median 14 days, mean 17.8 days; proportion of patients ill for >1 week, 79%), the so-called "acute" version of the SF (with only 1 week recall period) would not enable us to capture the full disease-episode for the majority of patients. However, when considering the course of illness in retrospect, the memory may average out the experienced extremes to a more constant level [24]. Since the patients we surveyed were not ill for exactly 4 weeks they had to reconcile in their valuation their period of ill-health with their period of good health over the total recall period.

Second, the SF-6D and the EQ-5D surveys also differ in the coverage of health dimensions and in the number of levels of severity in each dimension. The SF-6D survey can generate 18,000 health states while the EQ-5D survey distinguishes between 243 states. Some studies found the SF-6D survey to be suffering from a 'floor effect'. Responses on the dimensions of physical functioning and role limitations have a disproportionate number of responses at the lowest level for patient groups with more severe illness [10,23,25]. In our study, the difference was only pronounced for 'role limitations'. The EQ-5D survey however is said to be suffering from a 'ceiling effect'. Other studies reported that patients indicated perfect health (state 11111) on the EQ-5D survey and imperfect health on the SF-6D survey [23,25]. In our sample 15% of the coupled responses indicated perfect health on the EQ-5D and lower levels on the SF-6D. This refers to the cluster analysis mentioned above.

The SF-6D survey has a dimension that is not present in the EQ-5D, vitality, which can be of special significance for an illness such as hepatitis A. In our sample 95% indicated at least some problems with vitality on the SF-6D. Patients with a milder form of illness may not suffer from pain, mobility or anxiety but still score low on the vitality level. More than 50% of the respondents indicated severe role limitations, while 5% had problems with self-care on the EQ-5D. Role limitations are however more severe than self-care and are perhaps more appropriate to capture the limiting effects of having an infectious disease on the patient's behaviour. Infected people may be capable of performing several duties but be limited in the fulfilment thereof out of precaution for infection to others. This effect, if it is considered to be part of health-related QoL, is possibly better measured by the SF-6D than in the EQ-5D and may explain the higher QALY loss for patients with only mild symptoms. Apart from the differing dimensions, the descriptive system of the SF-6D allows for more response levels on each dimension (four to six vs. three). Patients with only slight problems may be more able to indicate these on the SF-6D scale than on the EQ-5D.

A third possible explanation is the scoring algorithm used to value the described health states. A multidimensional health description needs to be reduced to one dimension in order to be useable for economic evaluation. The EQ-5D and the SF-6D both use different methods to do this; respectively, the TTO and the SG method. A number of publications have indicated that SG generates higher values for more severe health states than TTO. For milder states the TTO seems to generate higher values [10,26,27]. These differences in scoring algorithm lead to an EQ-5D scale with twice the range of the SF-6D scale: [−0.59 to 1.00] vs. [0.25 to 1.00] for the parametric one and [0.20 to 1.00] for the non-parametric one. This is reflected in the range of the different QALY loss measures as shown in Table II. The differences between both SF-6D based scoring algorithms remains small compared to the difference with the EQ-5D.

There may be several reasons why cost-utility analysis is considered a controversial instrument for policy advice. One of them is the lack of an unambiguous method for valuing morbidity. Different methods for valuing the health burden of an acute illness for which only a single time point measurement is available, can lead to different results. The inclusion of a recall period, or the assumption of a linearly improving or a constant health state, both are analytical choices with a great impact if the health burden of the illness is substantially attributable to morbidity. The most appropriate method for health description is a matter of discussion since there is no 'gold-standard' to determine which measure approximates best the 'true health state'. Much depends on the concept of health that policymakers want to use to prioritize. For example, is it the better digested memory of a specific health state (captured in a survey with recall period) that counts, or is it the immediate experience (as captured in a current state survey)? Clearly, relying on memory suffers from all kinds of biases [28,29]. The fact that the SF-6D produces more consistent outcomes across the sample than the EQ-5D may explain more on the functioning of the memory than on the actual disease experience.

Conversely, a current state description requires the analyst to make assumptions on the further evolution of the illness between the different points of measurement, which can have a significant impact on the estimated QALY, if there is no opportunity to administer surveys regularly. Opting for a linear health improvement instead of a constant health state between two points in time yields, per definition, half the QALY loss. This leads to an incremental cost-effectiveness ratio of twice the magnitude (if only morbidity QALYs are considered). Considering these differences and the advantages and disadvantages of each method, researchers should therefore consider performing economic evaluations using different estimates of health benefits, and incorporate the range of the measures by means of sensitivity analysis rather than assuming a single utility score. We have provided a conversion equation between the QALY loss estimated by the EQ-5D and SF-6D

measures in the case of hepatitis A, which may simplify this process if one of the measures is available.

Conclusion

This study made clear that four conceptually defensible methods for measuring the health burden of an acute illness, based on the EQ-5D and the SF-6D surveys, produce statistically different results. The hypotheses that the methods based on the EQ-5D lead to better health scores in healthy patients and worse scores in severely ill patients than methods based on the SF-6D were confirmed. However, the most appropriate method remains a matter for further debate. Researchers should not only choose between surveys but also consider the effect of the time at which the survey is taken, the length of the recall period, the assumed disease progression and the scoring algorithm underlying the resulting outcome. Cost-utility analyses should make explicit the underlying techniques and assumptions behind a summary health utility score and if possible expand the results with other measurement methods in a sensitivity analysis.

Acknowledgments

We would like to thank the members of the Department of Control of Infectious Diseases (Flemish public health authorities), for their efforts in contacting hepatitis A patients and collecting the surveys, under supervision of Dr. Petra Claes, Dr. Anmarie Forier, Dr. Ruud Mak, Dr. Emmanuel Robesyn and Dr. Geert Top. We also want to thank Dr. S.A. Kharroubi and Prof. J.E. Brazier for providing us the 'ScHARR SF-36 to SF-6D converter for Bayesian and parametric utility scores'.

REFERENCES

[1] Nord E, Daniels N, Kamlet M. QALYs: some challenges. Value Health 2009;12(Suppl. 1):S10–5.
[2] Lipscomb J, Drummond M, Fryback D, et al. Retaining, and enhancing, the QALY. Value Health 2009;12(Suppl. 1):S18–26.
[3] Hausman DM. Valuing health properly. Health Econ Policy Law 2008;3(Pt 1):79–83.
[4] Hausman DM. Valuing health. Philos Public Aff 2006;34:246–74.
[5] Kontodimopoulos N, Pappa E, Papadopoulos A, et al. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Qual Life Res 2009;18:87–97.
[6] Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ 2009;10:15–23.
[7] Barton GR, Sach T, Avery A, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. Health Econ 2008;17:815–32.
[8] The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
[9] Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851–9.
[10] Brazier JE, Ratcliffe J, Salomon J, et al. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford: Oxford University Press, 2007.
[11] Ware JE Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
[12] Earle CC, Chapman R, Baker C, et al. Systematic overview of cost-utility assessments in oncology. J Clin Oncol 2000;18:3302–17.
[13] Richardson G, Manca A. Calculation of quality adjusted life years in the published literature: a review of methodology and transparency. Health Econ 2004;13:1203–10.
[14] Craig AS, .Schaffner W. Prevention of hepatitis A with the hepatitis A vaccine. N Engl J Med 2004;350:476–81.
[15] Bauch CT, Rao A, Pham B, et al. A dynamic model for assessing universal Hepatitis A vaccination in Canada. Vaccine 2007;25:1719–26.
[16] Cuthbert JA. Hepatitis A: old and new. Clin Microbiol Rev 2001;14:38–58.
[17] Kharroubi SA, Brazier JE, Roberts J, et al. Modelling SF-6D health state preference data using a nonparametric Bayesian method. J Health Econ 2007;26:597–612.
[18] Kharroubi SA, O'Hagan A, Brazier JE, Estimating utilities from individual health preference data: a nonparametric Bayesian method. J R Stat Soc Ser C Appl Stat 2005;54:879–95.
[19] Beutels P, Luyten J, Lejeune O, et al. Evaluation of universal and targeted hepatitis A vaccination options in Belgium, Health Technology Assessment (HTA), Brussel: Federaal Kenniscentrum voor de Gezondheidszorg (KCE), 2008. KCE reports 98A.
[20] Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
[21] Kharroubi SA, O'Hagan A, Brazier J. Estimating utilities from individual health preference data: a nonparametric Bayesian method. J Royal Stat Soc: Series C (Applied Statistics) 2005;54: 879–95.
[22] McCullagh P, Nelder JA. Generalized Linear Models (2nd ed). London: Chapman and Hall, 1989, p. 515.
[23] Brazier JE, Roberts J, Tsuchiya A, et al. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873–84.
[24] Bansback N, Sun H, Guh D, et al. Impact of the recall period on measuring health utilities for acute events. Health Econ 2008;17:1413–9.
[25] Longworth L, Bryan S. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. Health Econ 2003;12:1061–7.
[26] Dolan P, Gudex C, Kind P, et al. Valuing health states: a comparison of methods. J Health Econ 1996;15:209–31.
[27] Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. J Health Econ 2006;25:334–46.
[28] Litwin MS, McGuigan KA. Accuracy of recall in health-related quality-of-life assessment among men treated for prostate cancer. J Clin Oncol 1999;17:2882–8.
[29] Clarke PM, Fiebig DG, Gerdtham UG. Optimal recall length in survey design. J Health Econ 2008;27:1275–84.