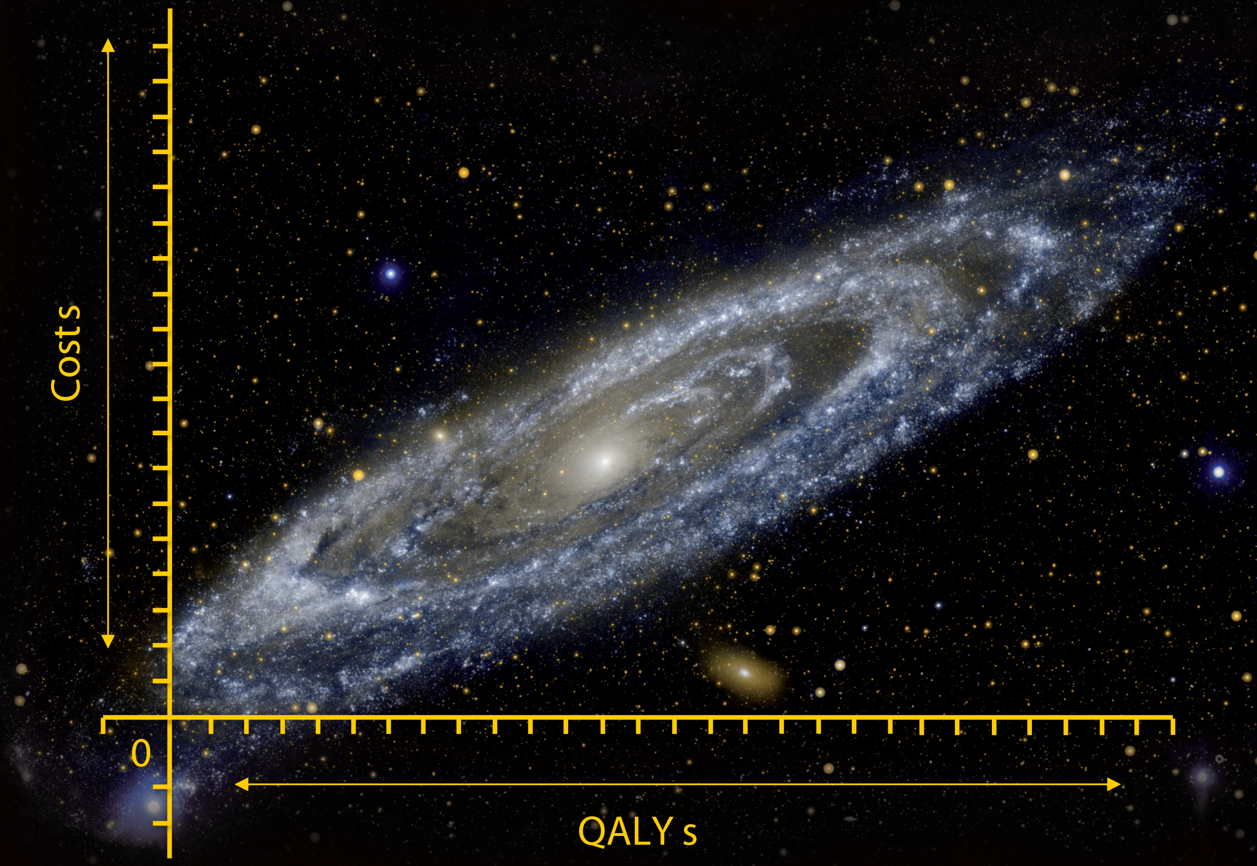
Mathematical Approaches in Economic Evaluations

Applying techniques from different disciplines



Mark Oppe

Mathematical Approaches in Economic Evaluations

Applying techniques from different disciplines

Mark Oppe

ISBN 978-94-6169-396-9

Layout and printing: Optima Grafische Communicatie, Rotterdam, The Netherlands

Cover picture: M81 Andromeda Galaxy, by courtesy of NASA-JPL-Caltech

Mathematical Approaches in Economic Evaluations
Applying techniques from different disciplines

Wiskundige methodes in economische evaluaties
Het gebruik van technieken uit verschillende disciplines

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.G. Schmidt
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 30 mei 2013 om 11.30 uur

door
Mark Oppe
geboren te Leiderdorp

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie

Promotor:
Prof.dr. C.A. Uyl-de Groot

Overige leden:
Prof.dr. M.P.M.H. Rutten-van Mölken
Prof.dr. J.J. van Busschbach
Prof.dr. N.J. Devlin

Copromotor:
Dr. F.Th. de Charro

Table of contents

Chapter 1

General introduction

Health economics (HE) is a multi-disciplinary field with links to economics, psychology and medicine. This is especially apparent in economic evaluations (EE) which have become an integral part in the management of health care systems in many western countries. In economic evaluations, information on a disease, on the cost of a treatment and on the effectiveness of the treatment is combined into a single mathematical model. This model is then used to assess the cost effectiveness of a treatment for the disease. The mathematical techniques employed to obtain and describe the information originate from three distinct mathematical disciplines associated with economics, psychology and medicine: econometrics, psychometrics and (bio)statistics. Even though there is a large amount of overlap, they all originated as separate disciplines and were developed with different perspectives in mind. This means that researchers in HE have a wide variety of different statistical and mathematical techniques at their disposal.

This dissertation shows how ideas and approaches from different disciplines can be applied in solving health economic problems. Basic statistical techniques common to all fields, such as linear regression, are common. They are applied in most of the studies presented in this thesis. In addition to this, the studies described in this thesis show how more specialised techniques and approaches can be used outside the field where they were originally developed. In particular they are used in economic evaluations and the measurement and valuation of health related quality of life.

The first of these is Monte Carlo simulation. Monte Carlo (MC) simulation is a numerical technique that was invented by Stan Ulam and John von Neumann while they were working on the Manhattan project [1,2]. With the increase in computing power over the years, MC simulation has become a very popular technique in many different fields including the social and health sciences. Different applications can also have different mathematical objectives and therefore different perspectives of regarding MC. For example, in mathematical finance MC is used as a *numerical integration technique* for the calculation of option pricing [3,4]. In contrast, in health economics MC is commonly viewed as a *numerical sampling technique* in simulation studies. In this thesis Monte Carlo simulation was used in two different applications. Firstly to determine the uncertainty around the outcome parameters of the Markov model for an economic evaluation of a new chemotherapy for colorectal cancer. Secondly to compute the uncertainty around the utility values for EQ-5D health states. Furthermore MC was used to determine the relationship between the number of health states and the number of respondents included in EQ-5D valuation studies and the uncertainty surrounding the utility values.

The second technique used outside its main field of application is factor analysis (FA), a statistical technique with strong ties to psychometrics. In psychology FA is applied in the development of questionnaires. FA can be used to explore the underlying dimensional structure of questionnaire data. The basic idea behind it is to investigate whether a number of items generate information about a more general underlying construct [5]. FA determines common factors and the way questionnaire items are associated to these factors by analysing the pattern of correlation. Items with relatively high inter-correlation are assumed to reflect the same construct, and items with low inter-correlation reflect different constructs. In this thesis FA was used in an econometric problem. Namely, can utility values obtained with a generic utility measure be attached to the health states from a disease specific quality of life instrument?

The third technique used outside its main field of application is the Discrete Choice Experiment (DCE) and associated Discrete Choice Models (DCM). In a DCE respondents are shown two alternatives and are asked to choose which of those two alternatives they prefer. Because the data collected is typically in the form of a proportion of respondents preferring alternative A to alternative B, a specific class of models (i.e. probability models) is needed to analyse these data. DCEs and DCMs have been extensively applied and refined in the fields of transport economics and marketing. In this thesis DCE and DCM were employed as an alternative to time trade off (TTO) or visual analogue scale (VAS) for the elicitation and modelling of health state valuations.

Besides the use of the various statistical techniques from the social and health sciences, mathematical approaches from seemingly unrelated fields such as (astro)physics can also be used in health economics. This is shown in a study where polar coordinates r and ϑ (i.e. radius and angle) are used rather than the conventional Cartesian coordinates x and y to estimate a utility model for the valuation of health related quality of life. In essence what this means is that the regression model was based on geometry. Even though the use of polar coordinates (and therefore geometry) is highly uncommon in the social and health sciences, it is considered the default approach in many branches of physics and astrophysics.

Lastly, the frequentist approach to statistics and the Bayesian approach to statistics were compared in a study to perform a fixed-effect and a random-effect meta-analysis. Both approaches have their own underlying framework and, for the purpose of data synthesis, both approaches allow for the same type of models. The frequentist approach is the conventional type of statistics and is different from the Bayesian approach. The idea behind Bayesian statistics is that what is known (or believed to be true) about the model parameters before seeing the new data can be captured in a probability distribution called a prior. This prior is then synthesized with the information in the new data to produce a posterior probability distribution, which expresses what we now know about the parameters after seeing the data [6].

Structure of this thesis

This structure of this thesis is as follows. Chapter 2 introduces the concept of quality of life measurement. It examines the impact of age on quality of life of the general population from eleven European countries. The chapter also contains a description of the EQ-5D, the most widely used generic utility measure, which is applied extensively in this thesis. Chapter 3 forms an introduction into economic evaluations. It provides an example of a Markov model for the evaluation of the cost-effectiveness of panitumumab, a new chemotherapy for the treatment of colorectal cancer.

This is followed by three chapters on sources of uncertainty in CUA outcomes. Chapter 4 investigates the impact of four different methods of meta-analysis on the outcomes (i.e. cost-effectiveness) of a probabilistic Markov model for Chronic Obstructive Pulmonary Disease (COPD). Chapter 5 presents a simulation study to quantify the amount uncertainty surrounding utility values and the link between the number of respondents and health states included in the valuation study on the uncertainty surrounding the utilities. Chapter 6 assesses the appropriateness of using a mapping model to attach EQ-5D based utilities to health states from a disease specific quality of life measure for use in hip related disorders, the Oxford Hip Score (OHS).

The last two studies presented in this thesis are on new approaches in utility modelling. Chapter 7 shows how a regression technique based on geometry can be used to estimate a utility model for EQ-5D based health state valuations. Chapter 8 describes a study where EQ-5D health state valuation data was collected using four different elicitation techniques: ranking, VAS, TTO and DCM. The impact of the different elicitation techniques is compared on the basis of the resulting utility models.

Lastly, chapter 9 summarises the conclusions of the different chapters and discusses the implications of the results.

References

1. Eckhardt R. Stan Ulam, John von Neumann, and the Monte Carlo method. Los Alamos Science. 1987;(15), 131-137.
2. Metropolis N, Ulam S. The Monte Carlo method, Journal of the American Statistical Association, 1949;(44), 335-341.
3. Galanti S, Jung A. Low discrepancy sequences: Monte Carlo simulation of option prices. Journal of derivatives. 1997; (5): 63-83
4. Glasserman P. Monte Carlo methods in financial engineering. Springer, New York, 2010
5. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.
6. O'Hagan A, Luce B. A primer on Bayesian statistics in Health Economics and Outcomes Research. Bayesian Initiative in Health Economics & Outcomes Research 2003.

Chapter 2

Age dependency of self-reported health as measured by EQ-5D in Europe

Mark Oppe, MSc[1], Frank de Charro, PhD.[2]

1. Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

2. EuroQol Group Executive Office, Rotterdam, The Netherlands.

Introduction

The impact of age on health is getting increasingly important because of the aging populations in many western countries. Older age is associated with an increase of mortality rate due to poorer health. Poor health in turn is associated with a decrease in quality of life. Therefore both the increase in mortality rate and the decrease in quality of life are important when considering the health of an aging population. Many types of studies use the general population as a frame of reference. For example the mortality of the general population is taken as background mortality on which a particular disease related mortality is superimposed. In such studies it might be important to not only take this background mortality into account, but also the background quality of life of the reference population. It would therefore be helpful to investigate and model the impact of age on quality of life.

In this study we used data collected with the EQ-5D to investigate the impact of age on quality of life. The EQ-5D is a generic measurement instrument for use in adults to describe and value health states [1,2]. The EQ-5D classification describes health states according to five dimensions: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression. Each dimension has three levels: 'no problems'; 'some problems'; and 'severe problems'. Health-state descriptions are constructed by taking one level for each attribute, thus defining 243 (3^5) distinct health states, where '11111' represents the best and '33333' the worst state.

The health states can be converted into single summary index numbers (also known as utilities) using an EQ-5D value set. The value sets are based on statistical models applied to data that reflect the values of the EQ-5D health states. In the valuation studies, a sample of the general population evaluates hypothetical health states (i.e. descriptions of health as defined by the EQ-5D classification system). Commonly used techniques for the valuation of EQ-5D health states are the Visual Analogue Scale (VAS) and Time Trade-Off (TTO). The value sets are anchored on the utility scale where full health = '11111' = 1 and 'dead' = 0. This allows the EQ-5D index values to be used in the calculation of Quality Adjusted Life Years (QALYs).

In addition to the classification system and associated utilities, the EQ-5D instrument contains a visual analogue scale, the EQ VAS. The EQ VAS is a vertical rating scale ranging from "best imaginable health" at the top of the scale (=100) to "worst imaginable health" at the bottom of the scale (=0). Respondents are asked to rate their health today directly on this scale (as opposed to rating an EQ-5D health state on the scale). This means that the EQ-VAS can capture information from aspects of health not included in the descriptive system of EQ-5D.

The aim of this paper is to investigate the relation between age and self-reported health as expressed by the EQ-5D descriptive system and the EQ-VAS.

Methods

Data

The data originated from the EuroQol Group database that was assembled as part of the European Union funded EQ-net project from 2002 [3]. The database was updated with additional data that became available after 2002. For the purposes of this study, data for 10 European countries were used: Belgium [4], Finland [5], Germany [6,7], Greece [8], Hungary [9], The Netherlands [10,11], Slovenia [12], Spain [13-15], Sweden [16,17], and the UK [18,19]. Data from separate studies within the same country were pooled in order to increase the number of observations in each cell. As the current study was undertaken in order to provide models for the population norms, only data from surveys that targeted the general public were used. Data from two of the studies in the EuroQol EQ-net database were based on patient populations rather than the general population and were therefore not included. The final restriction on the data was that, due to the low number of very elderly respondents, age was restricted to the range of 18 to 85 years old.

Models

In order to investigate the relationship between self-reported health and age, OLS regression models were estimated where EQ-VAS was the dependent variable and age was the independent variable. Because the relation between EQ-VAS and age might be different for men than for women, OLS models were also fitted separately for men and women. Since there might be marked differences between the EQ-VAS values of people who classify their own health as perfect (i.e. in 11111) compared to those of people that indicate to have health problems on the descriptive system. About half of the sample reports an EQ-5D state of 11111. This proportion ranges from 75% for age 18 to 20% for age 85. Therefore an analysis was undertaken on for these two groups.

To gain insight in the effect of age on the individual 5 EQ-5D dimensions regression models were estimated for the proportion of reported problems per dimension. Because relatively few level 3 problems were reported, these models were carried out on dichotomized levels of the five EQ-5D dimensions (i.e. no problems; problems). Therefore logistic regression was used. The logistic model is of the form $P = 1/(1+e^{-z})$, where z is a function of age and P = the proportion of problems. The final model was selected based on best model fit as measured by: 1) adjusted R^2 of the models based on data including both the within and the between variance, 2) adjusted R^2 of the models based only on the between variance and 3) Akaike's information criteria (AIC).

When analyzing the data it was found that the EQ-VAS data per age-category was skewed. This could pose a problem since one of the assumptions underlying OLS regression is that the data is normally distributed. Therefore, the models were also fitted on EQ-VAS values that were transformed using a logistic transformation in order to remove this skewness [3]. The transformation that was used was

$$VAS^* = \ln\left(\frac{VAS - \min}{\max - VAS}\right), \text{ where } \min = 0.5 \text{ and } \max = 100.5.$$

All analyses were carried out in SPSS version 19.

Results

Table 2.1 shows an overview of the data that was used in this study. We found that the logistic transformation of the EQ-VAS data did remove some of the skewness. However, the model fits were poorer on the transformed data than on the raw data (results not shown). Therefore it was decided to carry out all analyses on the untransformed data. Table 2.2 presents the differences between the models, when data from all European countries was used.

Table 2.1: Proportion of respondents in the dataset by country, sex and age group.

			Age groups						
Country	N	% Male	18-29	30-39	40-49	50-59	60-69	70-79	80-85
Belgium	1258	47%	9%	23%	22%	17%	15%	11%	3%
Finland	1580	49%	20%	15%	14%	10%	18%	17%	5%
Germany	812	60%	15%	15%	15%	22%	20%	11%	3%
Greece	463	54%	29%	19%	17%	17%	11%	6%	1%
Hungary	5457	45%	22%	16%	20%	17%	13%	10%	2%
Netherlands	823	44%	14%	19%	13%	13%	17%	18%	7%
Slovenia	734	44%	27%	19%	17%	16%	11%	9%	1%
Spain	2701	47%	24%	17%	17%	16%	15%	10%	2%
Sweden	3587	54%	16%	19%	18%	20%	13%	11%	3%
United Kingdom	3349	43%	20%	20%	16%	14%	14%	12%	3%
Total	20764	48%	20%	18%	17%	16%	14%	11%	3%

Table 2.2: regression models for EQ VAS versus age based on data from all European countries.

	adjusted R^2	AIC	constant	age	age^2	age^3
Linear	.106	101924	93.216	–.340		
Quadratic	.107	101902	88.876	–.134	–.002	
Cubic	.108	101885	78.731	.616	–.019	.0001

The fact that the age parameter is smaller than zero indicates that EQ-VAS scores decrease with increasing age. Table 2.2 shows that age partly explains the decrease in reported own health although R^2 is low (about 11%). This is due to the fact that age is only an indicator for health. The remaining 89% of the variance is due to other factors, such as the presence of disease. The proportion of explained variance of the linear model was 10.6%. Adding a quadratic term increased the adjusted R^2 from 10.6% to 10.7%, an increase of 0.1%. Using a cubic model resulted in a value of R^2 of 10.8%. Compared to the R^2 resulting from the linear model this is an increase of 0.2%. Furthermore, Akaike's information criterion showed that there was only a marginal difference in information captured between the 3 models. Lastly, when using country specific data rather than aggregated data, the quadratic and cubic models resulted in regression coefficients that were not significantly different from zero, with a significance level of 5%. Therefore, adding a quadratic and/or a cubic term to the model results in only marginally higher proportions of explained variance compared to the linear model. Based on this information it was decided to use the linear model for modeling the age dependency of EQ-VAS.

In order to investigate the impact of gender on the relation between EQ VAS and age, we also estimated the linear models separately for men and for women. Table 2.3 shows the regression coefficients and values of R^2 for the linear model separate for men and women. A linear model was also estimated in which age, sex and the interaction of age and sex (i.e. age*sex) were introduced stepwise as covariates. This model resulted in the exclusion of sex from the model, but the values of R^2 and AIC were virtually identical. As can be seen in table 2.3 the constant is higher for women than for men, whereas the age parameter is smaller. This means that women start out with slightly higher EQ-VAS values than men, but from age 28 they are lower than those for men.

Table 2.3: Regression coefficients and values of R^2 for the linear model by sex.

	adjusted R^2	AIC	constant	age	age*sex (M=0,F=1)
Men	0.097	47169	92.61	–0.32	
Women	0.115	54688	93.71	–0.36	
Total	0.106	101896	93.12	–0.34	
Total + sex*age	0.107	101880	93.20	–0.35	0.023

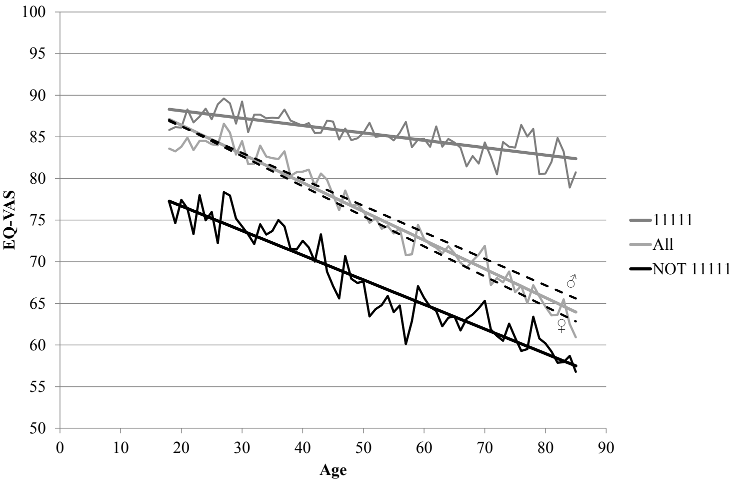


Figure 2.1: Observed EQ VAS scores and predicted EQ VAS scores by age with the linear model.

Figure 2.1 shows the EQ VAS scores predicted by the linear model and the mean observed EQ VAS scores per year of age (each age is used as an age group). This means that all the differences in reported EQ VAS within a single age (i.e. within state variance) have been erased. Figure 2.1 is therefore somewhat misleading because it shows only part of the total variance of the data, and the model seems to fit the data (explaining the total variance) much better than actually is the case. The separate models for respondents indicating no problems on any of the EQ-5D dimensions (i.e. 11111) and respondents with problems on one or more EQ-5D dimensions (i.e. NO 11111) show that in both cases the EQ-VAS decreases with age. Not surprisingly, the EQ-VAS values for 11111 consistently higher than those for the combined data. Furthermore, the EQ-VAS values for NO 11111 are consistently lower. The fact that the EQ-VAS values for 11111 also decrease with age indicates that the EQ-VAS picks up decreases in self assessed quality of life that the EQ-5D descriptive system doesn't pick up.

Separate models were tested for countries with more than 3000 respondents. These countries are Hungary, Spain and the United Kingdom. Even though the two Swedish studies together have more than 3000 respondents, Sweden was not included because only one

Table 2.4: Regression coefficients and values of R^2 for the linear model per country.

	adjusted R^2	constant	age
Hungary	0.214	95.65	−0.53
Spain	0.070	88.35	−0.27
United Kingdom	0.064	93.97	−0.24
Total	0.106	93.12	−0.34

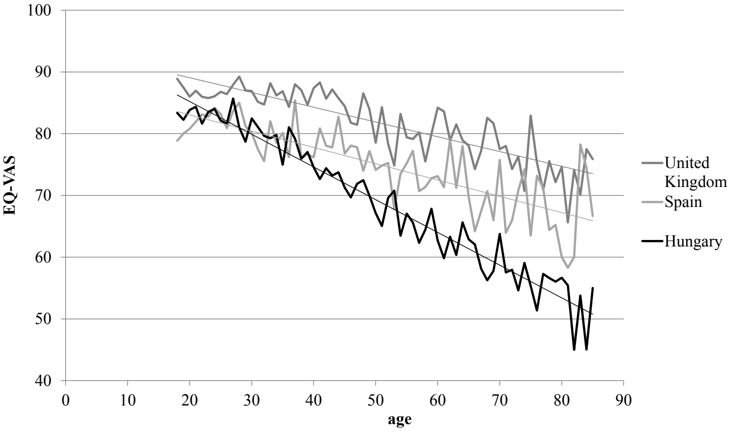


Figure 2.2: Observed EQ VAS scores and predicted EQ VAS scores by age with the linear model for three countries in the dataset.

of the two available Swedish studies reported EQ VAS values (N = 534). The models for Hungary, Spain and the United Kingdom are presented in table 2.4.

Figure 2.2 shows the EQ VAS values predicted by the linear model and the mean observed EQ VAS values per year of age for Hungary, Spain and the United Kingdom. As was the case in figure 2.1, all the differences in reported EQ VAS within a single age were erased. Figure 2.2 is therefore also somewhat misleading because it shows only part of the total variance of the data, so again the model seems to fit the data (explaining the total variance) much better than is actually the case.

Logistic regression was used to model the proportion of reported problems on each of the five EQ-5D dimensions. The models for sex were based on the data from all European countries. The constant regression coefficient sets the offset point at age 0. Small constant (i.e. large negative value) results in a low proportion of reported problems at age 0. The age coefficient is related to the global effects of age on the reported proportion of problems. A (relatively) large value of b_{age} results in a proportion of reported problems that increases faster with increasing age than a small value of b_{age}. Table 2.5 shows the regression coefficients and values of R^2 for the logistic model of the five EQ-5D dimensions.

As can be seen in table 2.5 the proportion of explained variance varies from 1.6% for anxiety/depression to 21.3% for mobility. Also, the values of the regression coefficients, and therefore the shape of the curves from the models, vary greatly between dimensions. For each dimension, the values of the constant for women are smaller than those for men, and the values of the age parameter for women are larger than for men. This means that

Table 2.5: Nagelkerke R^2 and regression coefficients for the logistic model of the proportion of reported problems.

		Nagelkerke R^2	constant	age
Mobility	Men	0.193	−4.733	0.058
	Women	0.231	−4.980	0.064
	Total	0.213	−4.867	0.061
Self-Care	Men	0.113	−6.135	0.056
	Women	0.146	−6.520	0.063
	Total	0.131	−6.347	0.060
Usual Activities	Men	0.107	−3.986	0.042
	Women	0.131	−4.190	0.046
	Total	0.120	−4.095	0.044
Pain / Discomfort	Men	0.114	−2.321	0.036
	Women	0.147	−2.409	0.041
	Total	0.131	−2.369	0.039
Anxiety / Depression	Men	0.010	−1.774	0.011
	Women	0.023	−1.743	0.016
	Total	0.016	−1.764	0.014

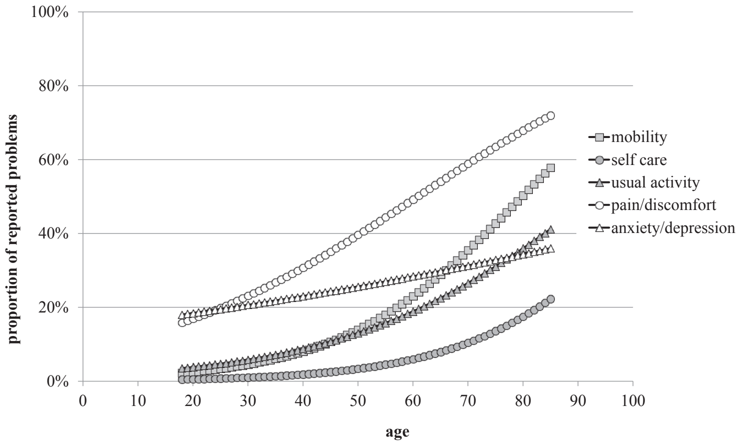


Figure 2.3: Logistic model for the proportion of reported problems by age for all European countries.

younger women report fewer problems than men, but the proportion of reported problems increase faster with increasing age for women than for men. The models of the proportion of reported problems by age for all European countries are presented in figure 2.3.

Table 2.6: Nagelkerke R^2 and regression coefficients for the logistic model of the proportion of reported problems.

		Nagelkerke R^2	constant	age
Mobility	Hungary	0.254	−4.951	0.068
	Spain	0.205	−5.165	0.062
	Sweden	0.179	−5.565	0.063
	United Kingdom	0.190	−4.426	0.055
	Total	0.213	−4.867	0.061
Self-Care	Hungary	0.189	−6.629	0.071
	Spain	0.117	−7.000	0.061
	Sweden	0.070	−6.709	0.049
	United Kingdom	0.072	−5.401	0.042
	Total	0.131	−6.347	0.060
Usual Activities	Hungary	0.193	−4.895	0.060
	Spain	0.093	−4.204	0.040
	Sweden	0.033	−3.814	0.026
	United Kingdom	0.094	−3.585	0.037
	Total	0.120	−4.095	0.044
Pain / Discomfort	Hungary	0.181	−2.713	0.048
	Spain	0.077	−2.283	0.029
	Sweden	0.080	−1.720	0.031
	United Kingdom	0.142	−2.727	0.041
	Total	0.131	−2.369	0.039
Anxiety / Depression	Hungary	0.077	−2.015	0.030
	Spain	0.008	−2.090	0.010
	Sweden	0.000	−0.903	0.000[†]
	United Kingdom	0.023	−2.150	0.017
	Total	0.016	−1.764	0.014

[†] Regression coefficient is not significantly different from 0 with a significance level of 0.01 (b_{age} = 0.00006, significance = 0.98).

Separate models have been made for countries with more than 3000 respondents. Because data from both Swedish studies could be used for determining the models for reported proportion of problems, models for Sweden have been made in addition to models for Hungary, Spain and the United Kingdom. The regression coefficients and Nagelkerke R^2 for the logistic models of the proportion of reported problems are presented in table 2.6.

Comparing tables 2.5 and 2.6 shows that differences between countries were larger than the overall differences between men and women, as was the case for the models estimated for the EQ-VAS. The reason that R^2 is zero for the anxiety/depression dimension for the Swedish data is that since the age parameter is essentially equal to zero, no age dependency was found, and hence the variance cannot be explained using a model that has age as its only explanatory variable.

Conclusions and discussion

In order to investigate the relationship between self assessed health related quality of life and age as measured by EQ-5D, we analysed population data from 10 European countries. First we estimated regression models where the EQ-VAS values were the dependent variable, and age and gender the independent variables.

The linear model resulted in a constant = 93.12 and age parameter = −0.34 with adjusted $R^2 = 0.106$. Extending the model to also include quadratic and cubic age parameters did only not improve the fit of the model. Using the mean observed EQ-VAS scores per year of age instead of the raw data, (i.e. all the within variance is removed), results in a dramatic increase in R^2. The value of R^2 increases from 10.7% to 93.9% while the regressions coefficients stay the same. If not only the raw EQ VAS values are changed into the mean EQ VAS values per year of age, but also only a single observation per year of age is used, all ages receive equal weight in the regression model. The resulting R^2 for this variant of the linear model is 0.95. The corresponding regression coefficients change slightly from the individual-data model results to: constant = 93.26 and age = −0.35.

Differences were found between men and women, where women tend to start out with higher quality of life than men. However, because women show a steeper decline in EQ-VAS scores by age than man, men tend to have higher EQ-VAS values from age 28 and up. However, the differences are small (about 3 points on the VAS at age 85). However, even though they are statistically significant they should be interpreted with care because no adjustments were made for differences in survival between men and women. Lastly, we found that differences between the linear models based on the EQ-VAS between countries are larger than the overall difference between men and women.

Modeling of the proportion of reported problems has been carried out using logistic regression. Again the values of R^2 are low (ranging from 23% to 1%), but all but one of the regression coefficients are significant with a significance level smaller than 0.01. Distinct differences were found between the relation with age of the physical domains mobility, self-care and usual activities, and the domains pain/discomfort and anxiety/depression. The physical domains start with few reported problems (<5%) at age 18, that increase with age at an increasing rate. Pain/discomfort and anxiety/depression start out higher (18% and 19%) and also increase with age but at a constant rate, which is higher for pain/discomfort than for anxiety/depression.

Not all data used for each country is fully representative of that country. Some studies have been conducted in specific geographic regions or other subgroups of the population as a whole. Despite this, we consider the collective dataset has adequate representativeness for Europe and has enabled us to obtain meaningful results, as the ten countries taken together have a good geographical distribution over Europe.

References

1. The EuroQol Group. EuroQol-a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199-208.
2. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53-72.
3. Brooks, Rabin, de Charro (Eds). The measurement and valuation of health status using EQ-5D: a European perspective. Kluwer Academic Publishers 2003.
4. Cleemput I, Kind P, Kesteloot K. Re-scaling social preference data: implications for modelling. In: Kind P, Macran S. (Editors). 19th Plenary Meeting of the EuroQol Group. Discussion Papers. Centre for Health Economics, University of York. 2004:113-123.
5. Ohinmaa A, Eija H, Sintonen H (1996). Modelling EuroQol values of Finnish adult population. In: Badia X, Herdman M, Segura A. (Editors). EuroQol Plenary Meeting. Discussion Papers. Barcelona. Institut Universitari de Salut Publica de Catalunya. 1996:67-76.
6. Schulenburg J.-M. G. V. D, Claes C, Greiner W, Uber A. The German version of the EuroQol quality of life questionnaire. In: Badia X, Herdman M, Segura A. (Editors). EuroQol Plenary Meeting. Discussion Papers. Barcelona. Institut Universitari de Salut Publica de Catalunya. 1996:135-161.
7. Claes C, Greiner W, Uber A, Schulenburg J-M. Graf v.d. An interview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. In: Greiner W, J-M. Graf v.d. Schulenburg, Piercy J. (Editors). (EuroQol) Plenary Meeting. Discussion Papers. Hannover Uni-Verlag Witte. 1999:13-39.
8. Yfantopoulos Y. Quality of life easurement and health production in Greece. In: Greiner W, J-M. Graf v.d. Schulenburg, Piercy J. (Editors). (EuroQol) Plenary Meeting. Discussion Papers. Hannover Uni-Verlag Witte. 1999:100-114.
9. Szende A, Nemeth R. (2003). Health-related quality of life of the Hungarian population. Orv Hetil. 2003;144(34):1667-74.
10. Essink-Bot ML, Stouthard M, Bonsel GJ. Generalizability of valuations on health states collected with the EuroQol questionnaire. Health Economics. 1993;2:237-246.
11. Lamers LM, McDonnell J, Stalmeier PFM, Krabbe PFM, Busschbach JJV. The Dutch Tariff: Results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121-1132.
12. Prevolnik Rupel V, Rebolj M. The Slovenian VAS tariff based on valuations of EQ-5D health states from the general population. In: Cabasés J, Gaminde I. (Editors). 17th Plenary Meeting of the EuroQol Group. Discussion Papers. Universidad Pública de Navarra. 2001:11-23.
13. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Medical Decision Making. 2001;21(1): 7-16.
14. Gaminde I, Cabasés J. Measuring valuations for health states among the general population in Navarra (Spain). In: Badia X, Herdman M, Segura A. (Editors). EuroQol Plenary Meeting. Discussion Papers. Barcelona. Institut Universitari de Salut Publica de Catalunya. 1996:113-123.

15. Gaminde I, Roset M. Quality adjusted life expectancy. In: Cabasés J, Gaminde I. (Editors). 17th Plenary Meeting of the EuroQol Group. Discussion Papers. Universidad Pública de Navarra. 2001:173-183.

16. Bjork S, Norinder A. The weighting exercise for the Swedish version of the EuroQol. Health Econ. 1999;8(2):117-126.

17. Burström K, Johannesson M, Diderichsen F. Swedish population health-related quality of life results using the EQ-5D. Quality of Life Research. 2001;10(7):621-635.

18. Kind P, Dolan P, Gudex C, Williams A (1998). Variations in population health status: results from a United Kingdom national questionnaire survey BMJ. 1998;316(7133):736-41.

19. Kind P, Hardman G, Macran S (1999). UK Population norms for EQ-5D.York Centre for Health Economics. 1999: Discussion Paper 172.

Chapter 3

Economic evaluation of Panitumumab as monotherapy in metastatic colorectal cancer

Mark Oppe, MSc, Ken Redekop, PhD, Maiwenn Al, PhD, Carin Uyl de Groot, PhD.

Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands

This research was sponsored by an unrestricted grant from Amgen, Zug, Switserland

Introduction

Colorectal cancer is the fourth most commonly diagnosed cancer and the second leading cause of cancer-related deaths in the US. In the US, it is estimated that over 153.800 cases of colorectal cancer are diagnosed and over 50.600 patients will die from the disease in 2008 [1]. In the European Union, the most recent update from the GLOBOCAN database reports in 2008 an incidence of approximately 334.000 patients with colorectal cancer and 150.000 deaths associated with these cancers [2]. In the Netherlands, in 2010 approximately 12,700 patients were diagnosed, and over 5.000 patients were reported to have died from it [3].

At the time of diagnosis, almost a quarter to a third of the patients already have metastatic disease (stage IV; mCRC), and approximately a third of the patients diagnosed and resected with early-stage disease subsequently develop metastases [4]. Nearly all patients with metastatic cancer will die of their disease and treatment of metastatic disease is in most cases palliative in intent as no cure is available except in subjects with (potentially) resectable liver or lung metastases. 5-fluorouracil, a thymidylate synthetase inhibitor, introduced in the fifties of the previous century, is still part of the treatment. In the past decades, however, therapies for mCRC have improved dramatically and have shifted from monotherapy to combination therapy and, finally, to sequential combination therapy. Combinations of newer cytotoxic agents, i.e., irinotecan and oxaliplatin, with fluoropyrimidine regimens have improved survival in these patients and have become the new standard of care. The addition of the topoisomerase-I inhibitor irinotecan to the combination of 5-fluorouracil and leucovorin (5-FU/LV; FOLFIRI) in the first-line treatment of mCRC resulted in improved response rates, a longer time to progression, and greater overall survival than 5-FU/LV alone [5,6]. The cytostatic agent oxaliplatin, a third generation platinum compound, was introduced in the same period. In combination with 5-FU/LV (FOLFOX), oxaliplatin resulted in higher response rates, longer progression-free survival, and longer overall survival in the first-line treatment of mCRC than the combination of irinotecan and bolus FU/LV [7,8]. Both of these agents, irinotecan and oxaliplatin, are associated with significant toxicities, which occasionally can be severe, particularly when combined with bolus 5-fluorouracil.

In addition to these chemotherapeutic agents, in the last decade, three biologic compounds have been approved for the treatment of metastatic colorectal cancer: bevacizumab, cetuximab and panitumumab [9,10]. Although all three agents are monoclonal antibodies directed against specific cancer-related targets, they are quite different. Bevacizumab is a recombinant humanized antibody that targets tumor angiogenesis by specifically binding to the vascular endothelial growth factor (VEGF), thereby blocking VEGF binding to its receptor in endothelial cells and subsequent signaling. When in first-line combined with irinotecan and bolus 5-FU/LV (IFL), bevacizumab confers superior response rates, time to progression, and survival compared with IFL alone [11]. Moreover, the addition of bevacizumab to 5-FU/LV has also shown advantages in both overall and progression-free survival times compared with 5-FU/LV or IFL regimens [12].

Both cetuximab and panitumumab target the human epidermal growth factor receptor (HER-1/EGFR), which when activated by various ligands, including the epidermal growth factor (EGF), initiates a signaling cascade [13]. The protein product of the proto-oncogene KRAS (Kirsten rat sarcoma 2 viral oncogene homologue) is a central down-stream signal-transducer of EGFR [14]. In tumors, activation of KRAS by EGFR contributes to EGFR-mediated increased proliferation, survival and the production of pro-angiogenic factors [13,14]. KRAS is one of the most frequently activated oncogenes in human cancers. Mutations of the KRAS gene at certain hot-spots (mainly codons 12 and 13) result in constitutive activation of the KRAS protein, independently of EGFR signalling [15]. In mCRC, the incidence of KRAS mutations is in the range of 30 to 50% [16].

Study data have demonstrated that patients with mCRC and activating KRAS mutations in the tumor are highly unlikely to benefit from treatment with these antibodies as monotherapy or in combination with chemotherapy. Results from a pivotal trial in patients with mCRC who had failed standard chemotherapy [17], comparing panitumumab plus Best Supportive Care (BSC) versus BSC alone, according to KRAS status demonstrated that treatment with panitumumab in patients with a tumor with activating KRAS mutations in codon 12 and 13, no treatment effect could be demonstrated [18]. In patients with a KRAS wild-type tumor, higher disease control (51% versus 12%) and improved progression free survival (median 12.3 versus 7.3 weeks; Hazard Ratio=0.45, 95% Confidence Interval 0.34-0.59) was shown [18]. 76% of patients randomized to BSC alone in first instance were given panitumumab upon progressive disease in a subsequent protocol. Therefore, no effect on overall survival could be demonstrated [18].

Panitumumab is the first fully human monoclonal antibody approved in mCRC targeting the EGFR. Panitumumab is approved in Europe by the European Medicines Agency for the treatment of patients with wild-type KRAS mCRC in first-line in combination with FOLFOX, in second-line in combination with FOLFIRI for patients who have received first-line fluoropyrimidine-based chemotherapy (excluding irinotecan), and as monotherapy after failure of fluoropyrimidine-, oxaliplatin-, and irinotecan-containing chemotherapy regimens. The chimeric antibody cetuximab is approved by the European Medicines Agency for the treatment of patients with EGFR-expressing, KRAS wild-type mCRC in combination with irinotecan-based chemotherapy, in first-line in combination with FOLFOX, and as a single agent in patients who have failed oxaliplatin- and irinotecan-based therapy and who are intolerant to irinotecan. The introduction of the biologicals in the treatment of mCRC has resulted in a near doubling of patient survival, which is a significant accomplishment in the treatment of a disease that was once considered untreatable.

In the Netherlands, panitumumab monotherapy was granted temporary reimbursement in 2008 for KRAS wild-type mCRC after failure of fluoropyrimidine-, oxaliplatin-, and irinotecan-containing chemotherapy regimens. Although combination therapy of cetuximab and irinotecan was approved earlier by the EMA, it was not (widely) used in Netherlands due to a negative reimbursement assessment in 2007. However, in 2009, cetuximab was also granted temporary reimbursement as monotherapy. Currently, reimbursement dossiers are under discussion for both panitumumab and cetuximab combination therapy in earlier lines. In clinical practice, this means that in the Netherlands, EGFR antibodies are currently mostly used in third line treatment as monotherapy, even though it is not limited to third line per se.

The focus of the current cost-effectiveness study is on the use of panitumumab as monotherapy for mCRC patients that failed fluoropyrimidine-, oxaliplatin-, and irinotecan- containing chemotherapy regimens.

Methods

The primary source of effectiveness data for the cost-effectiveness model was an open-label phase 3 clinical trial including 463 patients with mCRC [17,19]. The treatment arm of the trial was panitumumab + Best Supportive Care (BSC; n = 231) and the comparator arm was BSC alone (n = 232). All patients were followed approximately every 3 months for up to 2 years after random assignment. Upon disease progression, patients in the BSC arm were eligible to receive panitumumab under a separate study in a cross-over protocol. In total

Table 3.1: Demographics and baseline characteristics of patients included in the trial.

	Panitumumab	Best Supportive Care
N	231	232
Sex		
% Male	63%	64%
Age		
median (range)	62 (27-82)	63 (27-83)
ECOG		
% 0, % 1, % 2	46%, 41%, 13%	34%, 50%, 15%
K-ras wild type		
N (%)	124 (54%)	121 (52%)

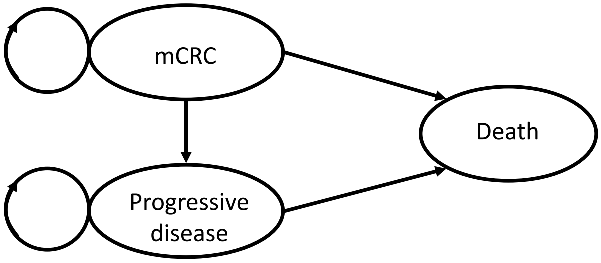


Figure 3.1: Structure of the cost-effectiveness model.

176 patients from the BSC arm (76%) crossed over to panitumumab. Both arms of the trial included patients both with wild-type and with mutated KRAS tumors, as the role of this biomarker was only investigated afterwards. A summary of the demographics and baseline characteristics can be found in table 3.1.

The model that was used to assess the cost-effectiveness of panitumumab plus BSC (Pmab arm) versus BSC alone (BSC arm) was a Markov model with a probabilistic sensitivity analysis (PSA). The three disease states in the model were: mCRC, Progressive disease and Death (figure 3.1). Patients start in the Markov state mCRC. From there they can stay in the mCRC state, enter the progressive disease state, or die. From the progressive disease state patients can remain in this state or die. A graphical presentation of the structure of the cost-effectiveness model is presented in figure 3.1. The cycle length of the model was 14 days and the time horizon was 4 years (i.e. 104 cycles). Through extrapolation of the survival data to this 4 year time horizon it was found that 99.99% of patients will have died after 4 years. Therefore the analyses are effectively life time analyses. One of the characteristics of a Markov model is that the model is memoryless. This Markov assumption implies that the risk of death for patients in the progressive disease state is the same for all patients in that state irrespective of the amount of time that patients spent in that state [20] (i.e., it has no memory of patient history). It is clear that in our case the Markov assumption does not hold (i.e., patient history should be taken into account when determining the transition probabilities) and therefore we used a micro simulation Markov model. In a micro simulation Markov model patients are sent through the model one at a time, so that individual patient histories can be examined when evaluating the model and time-dependency can be included in the transitions from one state to the next [20]. In each analysis 1,000 individual patients were simulated per arm. The probabilistic sensitivity analyses (PSA) used 1,000 iterations. The iteration of the PSA and the simulation of the patients were implemented as a nested Monte Carlo simulation. That is, for every iteration of the PSA, a new set of 1,000 patients per arm were simulated.

Model parameters

The input parameters used in the cost-effectiveness model fall into four categories: Transition probabilities between disease states, Use of panitumumab, Costs, and Health related Quality of Life.

Transition probabilities

In order to model the time dependent transition probabilities, we first analyzed the data using Kaplan Meier survival analyses. Next parametric functions were fitted to the data and compared to the results from the Kaplan Meier analysis. The time dependent transition probabilities from mCRC to death (i.e., patients who die before disease progression) and from progressive disease to death (i.e., patients who die after disease progression) were estimated using Weibull models. For the Weibull model the cumulative survival is: $S(t) = Exp(-L * t^G)$. In the PSA the correlation between L and G was taken into account using Cholesky decomposition. The scale parameters of the Weibull models were not normally distributed. Therefore the natural logarithm of the scale parameter *Ln(L)* was used to calculate the probabilistic transitions in the CE-model. The Kaplan Meier analyses for the transition probabilities from mCRC to progressive disease showed that the survival functions were initially concave, but after a while turned convex. Therefore two models were fitted to each curve, one for the concave part and one for the convex part. For the Pmab arm the cut-off point was at 56 days. The concave part was modeled using an exponential function $S(t) = Exp(-L * G)$. The convex part was modeled using a Weibull function. For the BSC arm the cut-off point was at 60 days. Both parts were modeled using a Weibull function. The cut-off points (i.e. 56 days and 60 days) were based on the results of the Kaplan Meier analyses and indicate that for the transition from mCRC to progressive disease the patients can roughly be divided into two groups, early progression versus late progression. Figure 3.2 shows the results from the estimation of the time dependent transition probabilities.

Panitumumab

Since the CE model uses BSC alone as a comparator, patients in the BSC arm who were treated with panitumumab were censored at the cross-over time. The exposure to panitumumab was therefore limited to patients in the Pmab arm. We accounted for the correlation between the number of treatments with panitumumab and the total time at risk in the mCRC and progressive disease states. According to the trial protocol patients did not receive panitumumab after disease progression was established. However, disease progression was monitored both by the physician treating the patient and separately (and retrospectively) by an expert panel. In our model we used the information on disease progression supplied

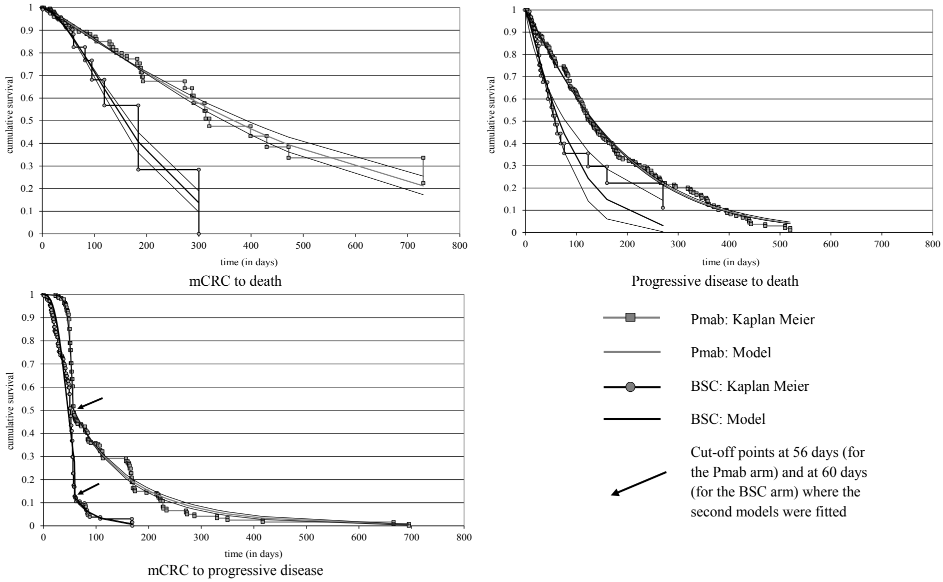


Figure 3.2: Transition probabilities between the 3 states of the model: Kaplan Meier survival curves and modelled survival curves (with 95% CIs) for the Pmab arm and the BSC arm.

by the expert panel in order to avoid effects of differences between individual physicians. In many instances the expert panel concluded that progressive disease occurred at an earlier point in time than the physicians did. Therefore, contrary to the protocol specification, panitumumab had been administered after disease progression and we incorporated this in the model and in the base case scenario. We tested the impact of this aspect of the model in a sensitivity analysis.

Costs

The unit cost that was used for panitumumab was the 2008 pharmacy purchase price of € 4.46 per mg. The total costs related to panitumumab were implemented in the model on the basis of the *mean* dosage of panitumumab that was used in the clinical trial. These were 425 mg (se = 3.2) before progression and 418 mg (se = 7.3) after progression. The uncertainty surrounding the mean dosage before and after progression was included in the PSA. All patients in the Pmab arm received panitumumab at least once before progression.

Apart from the costs of panitumumab, six other cost components were identified and included in the CE model. These were: concomitant medication, hospitalization, visits, medical procedures, radiotherapy, and chemotherapy. The resource use was derived from the trial data. The unit costs were based on data from the Netherlands and are expressed in

Table 3.2: Unit costs for panitumumab, mean costs per patient for the other cost components (in 2008 euro) and utilities for the disease states in the base case model.

	Panitumumab		Best Supportive Care	
	mean	se	mean	se
Unit Costs				
Panitumumab per 440mg	€1,962			
Mean Costs per patient				
Concomitant medication	€45		€18	
Hospitalization	€1,779		€1,248	
Visits	€386		€248	
Medical procedures	€109		€65	
Radiotherapy	€101		€88	
Chemotherapy	€97		€30	
Total (excl Pmab)	*€2,518*	*€273*	*€1,696*	*€269*
Utilities				
mCRC	0.74	0.007	0.68	0.013
Progressive disease	0.70	0.008	0.70	0.008

2008 Euros. The hospitalization costs made up 70% of the costs due to these six components. Therefore, the coefficient of variation of the hospitalization costs was used to estimate the standard error of the total costs associated with these six components. Subsequently, the total costs were incorporated in the model as a gamma distribution. An overview of the costs is presented in table 3.2.

Health related Quality of Life

The clinical trial included regular measurements with the EQ-5D to assess health related quality of life of the patients in the trial. The UK-TTO value set was used to convert the health profiles of the patients into utilities [21]. These utilities were consequently used to determine the mean utility of the mCRC and progressive disease states in the Markov model (table 3.2). The utility for the state dead was zero by definition. In the probabilistic sensitivity analyses beta distributions were fitted to the utilities. As can be seen in table 3.2, the mean utility values before progression in the Pmab arm were higher than those of the BSC arm (0.74 vs. 0.68). The mean utility values after progression were not significantly different at 95% level. Furthermore, because there was no clinical reason to assume that patients in the Pmab arm had different utilities after progression than patients in the BSC arm, the utilities after progression in both arms were set to the mean utility for all patients after progression (i.e., 0.70).

Scenarios

The cost-effectiveness of panitumumab was assessed in three different scenarios. In the first scenario (the base case) the model parameters were estimated using data from all patients included in the trial as described above. In the second scenario all model parameters were re-estimated based on data from the subgroup of patients that had KRAS wild-type mCRC (referred to as the KRAS-WT subgroup in the remainder of this paper). Compared to the base case, the biggest impact for the Pmab arm was in time to progressive disease; the Kaplan Meier estimates showed a median gain of 50 days for the KRAS-WT subgroup. For the BSC arm, the largest impact was on the time from progression to death. Contrary to our expectations, the survival analyses showed that the KRAS-WT subgroup had a median gain of 61 days for the time from progression to death for the BSC arm, compared to only a 7 day median gain for the time from progression to death for the Pmab arm. The data showed that this 61 day median survival gain was associated with only a handful of actual patients. This was due to the high level of censoring (i.e., 76% of patients in the BSC alone arm were allowed to cross-over to panitumumab). When censored data is used in Kaplan Meier analyses it is implicitly assumed that the censored patients do not differ from the non-censored patients. However, from a clinical point of view, it is possible that the patients in worst health cross over to the Pmab arm, whereas the patients in good health remain in the BSC arm. Therefore, it was hypothesised that the censoring due to cross-over lead to an overestimation of the survival for patients in the BSC arm. This hypothesis was confirmed by the direct comparison of the time from progression to death of the BSC arm of all patients to that of the KR-WT patients using a log rank test. This resulted in a Chi square of 0.322 with a significance of 0.57 showing that the 2 curves (and therefore the 61 day survival gain) were not statistically different at the 95% level. Because of this, the second scenario used the transition probability from progression to death from all BSC patients, not just from the KRAS-WT patients. However, we included the model results when using only the data from the KRAS-WT patients in the sensitivity analyses.

The third scenario was also based on the KRAS-WT subgroup, but was extended to include diagnostic testing for KRAS for all patients in the Pmab arm. These costs were assumed to be €200 per test, based on the Dutch Healthcare Authority tariff for complex molecular diagnostics. Patients with a positive test result (roughly one half – see table 3.1) received treatment with panitumumab plus BSC, while the patients with a negative test result received BSC alone.

Sensitivity analyses

Two sensitivity analyses were performed for scenario 1 both related to the inclusion of administration of panitumumab after progression. As stated previously, we used the retrospective assessments of disease progression from the expert panel, rather than the assessments of the physician treating the patient. The result of this is that the base case model also includes administering panitumumab after disease progression. In the first sensitivity analysis (SA1), patients in the Pmab arm received panitumumab only before progression. This implies that no costs were attributed to use of panitumumab after progression. It was assumed that the survival gain in the time from progression to death in the Pmab arm compared to the BSC arm was due only to the effects of panitumumab given *before* disease progression. Apart from the removal of the costs for panitumumab after progression all other model parameters are the same as in the base case.

The second sensitivity (SA2) analysis was a more conservative version of the first. Patients in the Pmab arm again received panitumumab only before progression. This time, however, it was assumed that the survival gain in the time from progression to death in the Pmab arm compared to the BSC arm was due to the effects of panitumumab given *after* disease progression. In order to simulate the survival time of patients in the Pmab arm after progression, we assumed that the survival after progression was the same as in the BSC arm. Therefore, the transition probabilities from the state progressive disease to the state dead of the BSC arm were used in both arms. The costs after progression in the Pmab arm were set to the value of the BSC arm. The other model parameters (i.e. the costs for panitumumab, the utility values and the other resource use costs) were the same as those in the base case.

In scenario 2 we used the transition probability from progression to death from all BSC patients, not just from the KRAS-WT patients to model survival after disease progression for the BSC arm. We included a third sensitivity analysis (SA3) for this scenario using only the data from the KRAS-WT patients.

Results

The base case scenario, which was based on all patients included in the trial, resulted in estimates of €15,502 for the incremental costs and 0.24 for the incremental QALYs (table 3.3). This leads a mean ICER of €64,321/QALY (table 3.4). The difference in number of QALYs between the two arms (i.e. ΔQALY in table 3.3) was larger for scenario 2 (0.24 versus 0.31) but smaller for scenario 3 (0.24 versus 0.16 respectively). In both the Pmab arm and the BSC arm, the mean number of QALYs (and life years) for the KRAS-WT scenario is higher than

Table 3.3: Costs, QALYs and incremental differences for the 3 scenarios.

	Panitumumab		Best Supportive Care		Incremental	
	Costs	QALYs	Costs	QALYs	Δ Costs	Δ QALYs
Base case						
mean	€17,193	0.50	€1,691	0.26	€15,502	0.24
95% CI	[15,503; 19,097]	[0.48; 0.53]	[1,181; 2,250]	[0.23; 0.34]	[13,740; 17,579]	[0.16; 0.28]
CoV †	0.05	0.03	0.16	0.11		
Scenario 2						
mean	€21,527	0.59	€1,346	0.28	€20,181	0.31
95% CI	[18,983; 24,306]	[0.56; 0.62]	[949; 1,826]	[0.24; 0.37]	[17,589; 22,969]	[0.22; 0.37]
CoV †	0.06	0.03	0.15	0.12		
Scenario 3						
mean	€11,611	0.43	€1,346	0.28	€10,264	0.16
95% CI	[10,183; 13,207]	[0.40; 0.50]	[950; 1,803]	[0.24; 0.39]	[9,038; 11,529]	[0.11; 0.18]
CoV †	0.06	0.03	0.15	0.12		

† CoV = Coefficient of Variation.

Table 3.4: mean ICERs (in 2008 Euro per QALY) plus 95% confidence intervals for the 3 scenarios and sensitivity analyses (SA).

	ICER	95% CI
base case	64,321	[52,642; 94,187]
SA 1	50,419	[40,748; 73,418]
SA 2	112,070	[95,386; 129,973]
Scenario 2	64,541	[52,136; 90,706]
SA 3	126,936	*
Scenario 3	66,131	[53,789; 96,316]

* Bootstrapped confidence intervals for this sensitivity analyses can not accurately be determined since the 95% CI of Δ QALYs includes 0.

for the base case scenario. Furthermore, the survival benefit of panitumumab compared to BSC is also higher. However, since the cost difference between the two arms is also higher, the ICER for scenarios 2 and 3 are close to the ICER of the base case ($\text{ICER}_{\text{scen 2}} = 64{,}541$ and $\text{ICER}_{\text{scen 3}} = 66{,}131$ respectively). SA1 resulted in a decrease of the ICER: $\text{ICER}_{\text{SA 1}} = 50{,}419$. This was due to a decrease in ΔCost of €3,351. SA2 – a conservative variant of SA1 – resulted in a similar reduction in ΔCost, but this was offset by a reduction in ΔQALY of nearly 55% leading to an $\text{ICER}_{\text{SA 2}} = 112{,}070$. The sensitivity analysis on the transition probabilities of the BSC arm in scenario 2 also resulted in an ICER about twice as high ($\text{ICER}_{\text{SA 3}} = 126{,}936$). This was due to the fact that in this sensitivity analysis ΔQALY was 50% lower (e.g. 0.16 versus 0.31 for scenario 2).

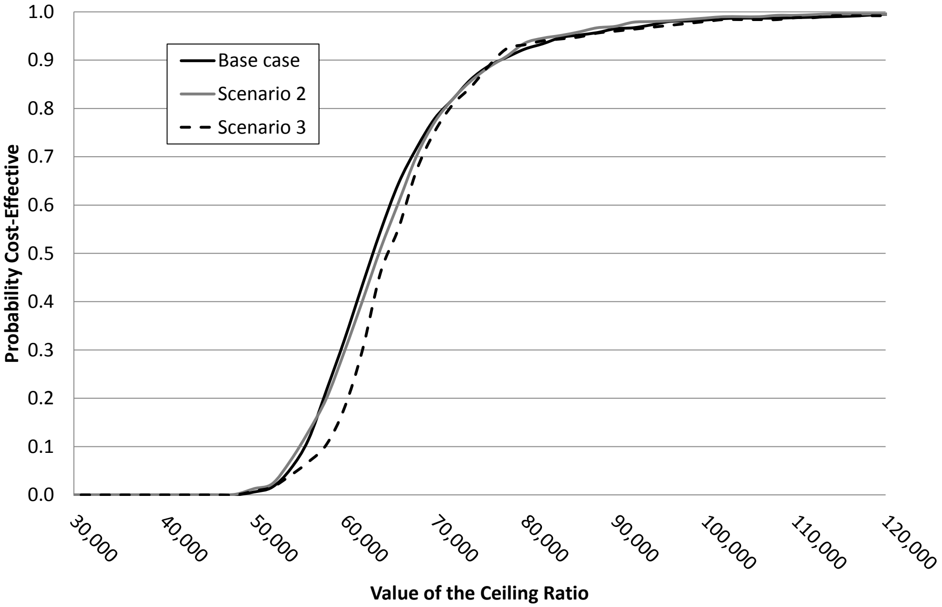


Figure 3.3: Cost Effectiveness Acceptability Curves: The probability that treatment with panitumumab is cost-effective per value of the ceiling ratio for the three scenarios.

As can be seen in table 3.3, the coefficient of variations (CoV) for the BSC arm are about 4 times the size of the CoV for the Pmab arm. This difference in uncertainty illustrates the impact of censoring the BSC patients at the time they cross-over to panitumumab.

Cost-effectiveness acceptability curves for the three scenarios are shown in figure 3.3. This figure shows the differences in the impact that the uncertainty associated with the three scenarios has on the probability that treatment with Pmab is cost effective.

Conclusions and discussion

The cost-effectiveness of panitumumab as monotherapy in mCRC was determined using a probabilistic micro simulation Markov model. Three different scenarios were evaluated with the model. The first scenario was the base case model based on the complete patient population of the pivotal trial in which panitumumab plus BSC was compared with BSC alone in mCRC patients after failure of fluoropyrimidine-, oxaliplatin-, and irinotecan-containing chemotherapy regimens, and resulted in an ICER of €64,321/QALY. The second scenario was based on the subgroup of patients with KRAS wild-type mCRC, but otherwise equal to the first. This scenario resulted in an ICER of €64,541/QALY. In the third scenario, prior testing for the KRAS tumor mutational status was included for patients in the Pmab arm. The resulting ICER was €66,131/QALY.

When averaged over the entire study time, the mean utility values before progression of the Pmab arm were about 9% higher than those of the BSC arm (0.74 vs. 0.68). However, the utility values at baseline were 0.68 for Pmab versus 0.66 for BSC and this difference was not statistically significant at the 5% level. This means that at baseline there was no difference between the two arms with respect to quality of life. Therefore there is both a survival gain and a quality of life benefit associated with the use of panitumumab before progression.

The cut-off points (i.e., 56 days and 60 days) were based on the results of the Kaplan Meier analyses and indicate that for the transition from mCRC to progressive disease the patients can roughly be divided into two groups, early progression versus late progression. This was also found when estimating the transition probabilities of the KRAS-WT subgroup, leading to the conclusion that this division is not associated with the KRAS tumor mutational status.

The uncertainty surrounding the ICERs is related to the uncertainty surrounding the costs and effects used in the model. From the coefficients of variation presented in table 3.3 it follows that the uncertainty related to the outcomes of the BSC arm is about 4 times higher than that related to the Pmab arm. This is because of the cross-over design of the trial. As mentioned earlier, 76% of patients in the BSC arm cross-over to receive panitumumab. All these patients are censored at the moment of cross-over, resulting in a loss of power in the BSC arm. In addition to this, about half the patients had KRAS wild-type mCRC, effectively halving the sample size, for analyses based on this subgroup.

A number of assumptions were made when building the CE model. Firstly, serious adverse events (SAE) were not modeled as separate Markov states. The data showed that the number of SAEs was low (94 for the Pmab arm; 50 for the BSC arm) and the median duration of these events was shorter than the cycle time of the model. Therefore the costs of these SAEs were included in the mCRC and progressive disease states. Secondly, treatment with panitumumab before and after progression was modeled separately and was therefore assumed to be independent of one another. Lastly, costs and QALYs were not discounted. In the Pmab arm, 76% of the patients will have died within 1 year, and in the BSC arm, 97% of the patients. After 2 years the proportions of patients that died increased to 97% and 100%. Therefore only very small differences were found between the discounted and non-discounted results, typically around 0.1% of the magnitude of the ICER. For example, in the base case the discounted ICER would drop from €64,321/QALY to €64,244/QALY. Therefore, compared to the uncertainty associated with the estimation of the ICERs themselves, the impact of discounting is negligible.

We feel confident that the effects of these assumptions are either very small compared to the amount of uncertainty surrounding the model outcomes, or even out when the model is evaluated using a small cycle time and a large number of patients and simulations.

Our results can be compared to the cost-effectiveness of cetuximab. Two economic evaluations have been carried out where cetuximab was compared to best supportive care in mCRC [22,23]. The models in those studies were different from the one we used in this study: they estimated the mean overall survival and mean costs directly from the trial data, whereas we used a decision analytic model. The main advantages of using a micro-simulation Markov model is the ease with which the trial results can be extrapolated beyond the trial time horizon and the possibility to implement and analyse different scenarios

Mittmann and colleagues found an ICER of $299,613/QALY, while Starling and colleagues estimated that the ICER was £57,608/QALY. For patients with KRAS wild-type mCRC, Mittmann obtained an ICER of $186,761/QALY. Comparing these results to our own indicates that panitumumab is associated with a lower ICER than cetuximab. This is the case both when all patients are considered and when only patients with KRAS wild-type mCRC are considered.

A commonly used “rule of thumb” for the maximum ICER threshold in the Netherlands is €80,000/QALY for conditions with a high burden of disease such as cancer [24] (although this is seen more as a guideline and not a fixed threshold). These drugs are put on the “expensive drugs list” where they are temporarily reimbursed, pending availability of new data from real world clinical practice. Our results show that the estimated cost-effectiveness of panitumumab is below this threshold and that temporary reimbursement is therefore warrented. An observational study with a follow-up time of 3 years is ongoing, where the focus is on establishing “real life” effectiveness and resource use instead of trial based efficacy and resource use. The results from that study will ultimately be used to run the economic evaluation with real-life data to facilitate the decision on whether or not panitumumab will be considered for prolonged reimbursement in the Netherlands as monotherapy for KRAS wild-type mCRC patients that have failed fluoropyrimidine-, oxaliplatine- and irinotecan-containing chemotherapy regimes.

References

1. International Agency for Research on Cancer. GLOBOCAN 2008, Cancer Incidence and Mortality. www.IARC.fr. Accessed on 12 November 2012.
2. International Agency for Research on Cancer. GLOBOCAN 2008, Cancer Incidence and Mortality. www.IARC.fr. Accessed on 12 November 2012.
3. IKNL 2012. www.IKNL.nl; Accessed on 12 November 2012.
4. American Cancer Society. www.cancer.org. Accessed on 12 November 2012
5. Saltz LB, Cox JV, Blanke C, Rosen LS, Fehrenbacher L, Moore MJ, et al. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. Irinotecan Study Group. N Engl J Med. 2000; 343(13):905-14.
6. Douillard JY, Cunningham D, Roth AD, Navarro M, James RD, Karasek P, et al. Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: a multicentre randomised trial. Lancet. 2000; 355(9209):1041-47.
7. Tournigand C, André T, Achille E, Lledo G, Flesh M, Mery-Mignard D, et al. FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. J Clin Oncol. 2004; 22(2):229-37.
8. Goldberg RM, Sargent DJ, Morton RF, Fuchs CS, Ramanathan RK, Williamson SK, Findlay BP, et al. A randomized controlled trial of fluorouracil plus leucovorin, irinotecan, and oxaliplatin combinations in patients with previously untreated metastatic colorectal cancer. J Clin Oncol. 2004; 22(1):23-30.
9. Punt CJ. New options and old dilemmas in the treatment of patients with advanced colorectal cancer. Ann Oncol. 2004; 15(10):1453-59
10. Cunningham D, Humblet Y, Siena S, Khayat D, Bleiberg H, Santoro A, et al. Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. N Engl J Med. 2004; 351(4):337-45.
11. Hurwitz H, Fehrenbacher L, Novotny W, Cartwright T, Hainsworth J, Heim W, Berlin J, Baron A, Griffing S, Holmgren E, Ferrara N, Fyfe G, Rogers B, Ross R, Kabbinavar F. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. N Engl J Med. 2004 Jun 3;350(23):2335-42
12. Kabbinavar FF, Hambleton J, Mass RD, Hurwitz HI, Bergsland E, Sarkar S. Combined analysis of efficacy: the addition of bevacizumab to fluorouracil/leucovorin improves survival for patients with metastatic colorectal cancer.J Clin Oncol. 2005 Jun 1;23(16):3706-12.
13. Baselga J. Why the epidermal growth factor receptor? The rationale for cancer therapy. The oncologist 2002;7 (suppl 4):2-8.
14. Schubbert S, Shannon K, Bollag G. Hyperactive RAS in developmental disorders and cancer. Nat Rev Cancer. 2007;7:295-308.
15. Riely GJ, Landanyi M. KRAS mutations: an old oncogene becomes a new predictive biomarker. J Mol Diagnost. 2008;10:493-496.

16. Andreyev HJ, Norman AR, Cunningham D, et al. Kirsten ras mutations in patients with colorectal cancer: the multicenter “RASCAL” study. J Natl Cancer Inst. 1998;90(9):675-684.

17. Van Cutsem E, Peeters M, Siena S, Humblet Y, Hendlisz A, Neyns B, et al. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. J Clin Oncol. 2007; 25(13): 1658-64.

18. Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol. 2008; 26(10):1626-34.

19. Clinicaltrials.gov. Available from http://www.clinicaltrials.gov; ID 20020408; Accessed on 10-04-2010.

20. Briggs A, Sculpher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford: Oxford University Press, 2006.

21. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997; 35(11):1095-108.

22. Mittmann N, Au HJ, Tu D, O’Callaghan CJ, Isogai PK, Karapetis CS, et al. Prospective cost-effectiveness analysis of cetuximab in metastatic colorectal cancer: evaluation of National Cancer Institute of Canada Clinical Trials Group CO.17 trial. J Natl Cancer Inst. 2009; 101(17): 1182-92.

23. Starling N, Tilden D, White J, Cunningham D. Cost-effectiveness analysis of cetuximab/irinotecan vs active/best supportive care for the treatment of metastatic colorectal cancer patients who have failed previous chemotherapy treatment. Br J Cancer. 2007; 96(2):206-12.

24. Raad voor de Volksgezondheid en Zorg. Zinnige en Duurzame zorg. Zoetermeer 2006.

Chapter 4

Impact of methods of data synthesis on the outcomes of cost utility analyses

Mark Oppe, MSc, Maiwenn Al, PhD, Maureen Rutten-van Mölken, PhD.

Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

Publication:

Oppe M, Al M, Rutten-van Mölken M. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics. 2011;29(3):239-50.

This research was sponsored by an unrestricted grant from Boehringer Ingelheim GmbH

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a chronic and progressive lung disease, characterised by predominantly respiratory symptoms that worsen over time, such as breathlessness, and episodes of acute decompensation, called exacerbations [1]. Pharmacological treatment of COPD aims to relief symptoms, improve lung function, reduce frequency and severity of exacerbations and improve quality of life. In order to establish the cost-effectiveness of available treatments, several cost-effectiveness (CE) models have been published [2] one of which co-developed by the authors of this research presented here [3-5].

In order for cost-effectiveness models to be able to inform resource allocations over longer periods of time, they must be suited to incorporate emerging evidence. Some model updates may be rather simple and straightforward, such as including new cost data for resource use types, whereas others will be more complex to handle, such as new patient-level data that can inform mathematically-derived model parameters such as transition probabilities. With the evolution from deterministic to probabilistic models, this task will be even more complex, as not only point estimates of the parameters will need updating but also their associated distributions and potential inter-dependencies.

Meta-analysis has traditionally been used to combine quantitative results of several similar studies into a pooled estimate of the relative treatment effect (e.g. odds ratio, relative risk, proportional difference in change from baseline). It uses the magnitude of the treatment effect and its uncertainty from each individual study to produce a weighted mean of the treatment effect [6-8]. However, in cost effectiveness models, a relative treatment effect like the odds ratio of having an exacerbation in COPD is not the only parameter to be estimated. A cost effectiveness model is made up of a wide range of different model parameters, including absolute treatment-specific transition probabilities between disease states, probabilities of experiencing events, utility values, costs and relative risks of the benefits of one treatment over the other. In other words, the model parameters in a probabilistic CE model have different types of distributions than those used to model a relative treatment effect. Both the distributions for these model parameters and the relative effects between comparators need to be modelled. Therefore there is more heterogeneity in the case of CE models than usual in meta-analyses.

The aim of this study is to illustrate how standard methods of evidence synthesis perform when applied to different types of model parameters and their distributions using an existing cost-effectiveness model. In our study, transition probabilities and exacerbation probabilities were re-estimated incorporating new patient-level data from a 1-year clinical trial. The outcomes of the study will demonstrate how sensitive the estimates of the cost-effectiveness are to the choice of method to combine the data.

Methods

Details of the Markov model for which we re-estimated the parameters were published before [3,4]. In short, the model has three COPD states of increasing severity based on pre-bronchodilator Forced Expiratory Volume in one second as percentage of the predicted value (FEV$_1$ % pred): moderate COPD, severe COPD and very severe COPD. The fourth state in the model is death. In pre-specified time intervals (Markov cycles) patients move between states and have a risk of experiencing an exacerbation. The model adopts a time horizon of 5 years. The cycle length of the model is one month, except for the first cycle where it was 8 days. Transitions between states were assumed to take place halfway through the cycle. During each cycle, there is a risk of getting an exacerbation. That exacerbation can either be severe or non-severe.

The risk of experiencing an exacerbation varies by disease state and treatment group. Given treatment group and disease state, exacerbation probabilities were assumed to be constant over time. Healthcare resource use and quality of life values (utilities) depend upon COPD severity state and the severity of the exacerbation. Given disease state and exacerbation severity, resource use and utilities were assumed to be similar across treatment groups in the model. The model investigated three different bronchodilator therapies, the reference treatment (i.e. the new treatment of interest) and two comparator treatments. Differences in costs and QALYs between those are driven by three factors: 1) the transition probabilities between disease states which themselves depend on the decline in FEV$_1$ % pred, 2) the exacerbation probabilities in each state and 3) the costs of the (study) medications for each treatment group.

All monthly transition probabilities between disease states in the first year and all monthly probabilities to experience exacerbations in the first year and subsequent years were directly obtained from the patient-level data of the three trials used to construct the original model [9-11]. In the first trial the difference between the reference treatment and comparator treatment 1 was assessed ($N_{ref}=356$; $N_{comp1}=179$). In the second trial the difference between the reference treatment, comparator 2 and placebo was assessed ($N_{ref}=402$; $N_{comp2}=405$ $N_{plac}=400$). In the third trial the differences between the reference treatment and placebo were assessed ($N_{ref}=550$; $N_{plac}=371$). The first year transition probabilities and exacerbation probabilities were based on patient-level data of clinical trials of up to one year duration [9-11]. The subsequent probabilities were estimated based on the published decline in lung function (FEV$_1$ decline 52 ml/year [12]), or in case of the exacerbation probabilities, were assumed to remain as observed in the first year. These were first calculated for the reference treatment and then derived for the comparator treatments as the relative differences between the reference and the comparators.

The reason to conduct the current study was that new patient-level data from a clinical trial with a follow-up time of 1 year became available. In this new trial the reference treatment was compared with placebo ($N_{ref} = 670$; $N_{plac} = 653$) [13-15]. These new data triggered us to re-estimate three sets of parameters of our CE model: 1) the probability of getting an exacerbation, 2) the probability that the exacerbation was severe, and 3) the transition probabilities between disease states. These re-estimations were performed with different techniques of data-synthesis because we aimed to investigate the consequences of applying these different techniques for the cost-effectiveness estimates. Costs, resource use, utility values and relative risks of the reference treatment versus the comparators remained unchanged.

The parameters of the reference treatment in the CE model were re-estimated by directly applying the various meta-analysis models (see below). The parameters for the two comparator treatments were re-estimated indirectly using the new results for the reference treatment and the (old) relative risks between the reference and the comparator treatments.

In this study we focused on two model specifications that are generally used for meta-analysis: fixed-effects (FE) and random-effects (RE). In fixed-effect meta-analysis the assumption is made that each of the individual studies aims to estimate the same true parameter value (e.g. the underlying exacerbation probability or transition probability) and that differences between studies are due to random (sampling) error. In other words, it is assumed that all factors that could influence the parameter value are the same in all the study samples, and therefore the effect size is the same in all the study samples. The combined effect is the estimate of this value. In this study, we used the inverse variance method for the fixed-effect meta-analysis. It produces a weighted average of aggregated data across all studies to give a pooled estimate of the transition probabilities and the exacerbation probabilities. The weights are based on the inverse of the uncertainty (standard error squared) of each study, i.e. studies with a large variance get a small weight and vice versa. The variance of the pooled estimate is calculated as the inverse of the sum of the weights [8].

In a random-effect meta-analysis it is assumed that, in addition to sampling error, differences between studies are caused by heterogeneity between studies [16]. In other words, it is assumed that all studies are samples drawn from a pool of all possible studies that differ from each other in ways that could impact on the treatment effect. For example, the intensity of the intervention or the age of the subjects may have varied from one study to the next. The goal is to estimate the mean of all possible studies. The true parameter value may be study specific and varies across studies. The variance in the true underlying parameter value of each study is called the random-effect variance. Random-effect models should be used when there is heterogeneity between study results caused, for example, by different patient populations or different study designs. As in the fixed-effect approach, the true parameter value is calculated as the weighted average of study-specific values. Now, however, the weights are based on a combination of the sampling error (standard error squared) of each study and the random-effect variance (in this case the trial specific effects). We used the method proposed by DerSimonian and Laird for our frequentist random effects analysis [7]. The homogeneity between the studies was tested with the I^2-statistic. The I^2-statistic is the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies [16,17]. When the between-study variance becomes 0, the random-effect meta-analysis becomes identical to the frequentist fixed-effect meta-analysis.

We have used both the frequentist approach and the Bayesian approach to perform a fixed-effect meta-analysis and a random-effect meta-analysis. Both approaches have their own underlying framework and, for the purpose of data synthesis, both approaches allow for the same type of models. The frequentist approach is the conventional type of statistics and is different from the Bayesian approach. The idea behind Bayesian statistics is that what is known (or believed to be true) about the model parameters before seeing the new data can be captured in a probability distribution called a prior. This prior is then synthesized with the information in the new data to produce a posterior probability distribution, which expresses what we now know about the parameters after seeing the data.

The Bayesian fixed-effects models were estimated algebraically (also known as Bayesian updating or a conjugate analysis). Because we were dealing with probabilities based on count data we specified all priors as beta distributions. When the prior distribution of a certain probability is distributed Beta(α_0,β_0) and the newly available data is characterized by a binomial distribution Bin(p,n), with n the sample size and p the probability of success, then the posterior distribution is characterized by the distribution Beta(α_0 + pn, β_0 + (1–p) n) [18].

When specifying random-effect models in a Bayesian framework, prior distributions need to be defined not only for the parameters of interest (i.e. transition probabilities and exacerbation probabilities) but also for the between-study variance (heterogeneity) [18-20]. The probabilities were based on binomial distributions Bin(p,n) just as in the Bayesian fixed-effects model specification. In order to model these, we defined $\mu_i = \text{logit}(p_i) = \ln(p_i/(1\text{-}p_i))$ with μ_i~Normal(μ_0, $1/\tau_0^2$), where τ_0 is the between-study standard deviation, $1/\tau_0^2$ is the precision, and the subscript i indicates the different trials. Because we had no strong preconceptions regarding the priors, we used non-informative priors in all cases. The prior distributions for μ_0 and $1/\tau_0^2$ were defined as μ_0~Normal(0, 10^{-6}) and $1/\tau_0^2$~Gamma(0.001,0.001). Generally, Bayesian RE models result in greater uncertainty surrounding the means than the frequentist RE models. This is because the uncertainty about the between study variance is captured in Bayesian RE models, while this is assumed to be known with certainty in frequentist RE models.

As stated previously, we used the I^2-statistic to assess homogeneity of the data in the frequentist approach and used that as a basis for the choice between FE and RE model specifications. Since there is no Bayesian equivalent of the I^2-statistic we used residual deviance to assess the model fit of Bayesian models [21]. Residual deviance is a measure of fit linked to the Deviance Information Criterion (DIC), the (Bayesian) hierarchical modeling generalization of the Akaike Information Criterion. Deviance measures the fit of the model to the data points using the likelihood function. The larger the likelihood, the closer the model fit. The best fit is where the model predictions equal the observed data. Such a model is called a saturated model. The residual deviance is the deviance for the model minus the deviance for the saturated model: $D_{res} = -2(\text{loglike}_{model} - \text{loglike}_{saturated})$ with posterior mean $\bar{D}_{res}$. If the model is an adequate fit, we expect $\bar{D}_{res}$ to be roughly equal to the number of data points (in our case the number of data points is 4, one for each study). The total number of iterations used to estimate the Bayesian RE models was 50,000. The burn in comprised the first 20,000 iterations. We checked convergence of the Bayesian RE models through inspection of the autocorrelation, sampling history, posterior density distributions and MCMC errors.

Summarizing, we compared the following four methods: frequentist fixed-effects meta-analysis, frequentist random-effects meta-analysis, Bayesian fixed-effects meta-analysis and Bayesian random-effects meta-analysis. The first three of these methods were implemented using algebraic calculations, but the Bayesian RE model required a Markov Chain Monte Carlo process. In order to investigate the impact of these four meta-analysis methods on the cost-effectiveness results, we filled the cost-effectiveness model with the parameter estimates that were obtained with each of the four methods and studied the differences between the three treatments in total costs, in total number of QALYs and in the Net Monetary Benefits that were used to calculate the Cost-Effectiveness Acceptability Frontiers [22]. We used MS Excel 2003 for the frequentist models and the Bayesian fixed-effects analyses and WinBUGS 1.4 for the Bayesian random-effects analyses.

Results

Exacerbation probabilities

Table 4.1 shows the mean and SE of the exacerbation probabilities in the reference group before and after incorporation of the data from the new trial, with either one of the four methods of meta-analysis. We use the term SE to indicate the uncertainty of the parameter estimates for all 4 methods. However, strictly speaking this is not the correct terminology since the uncertainty surrounding the Bayesian estimates is not the standard error of the mean, but the standard deviation of the posterior distribution. The different methods of meta-analysis resulted in different estimates of exacerbation probabilities and their standard errors. The fixed-effect meta-analysis produced the lowest means and standard errors for the exacerbation probabilities and the probabilities of the exacerbation being severe. The Bayesian random-effect meta-analysis resulted in the highest estimated probabilities of getting an exacerbation and in the highest standard errors. The estimated probabilities were at most 23% higher and the standard errors were up to 9 times higher than the Bayesian fixed-effect meta-analysis. Except for patients with moderate COPD, this method also resulted in the highest estimated probability that an exacerbation would be severe. In patients with moderate COPD, Bayesian fixed-effect meta-analysis resulted in a 20% higher estimated probability of an event being severe than the standard fixed-effect meta-analysis. It must be remarked here that when normal distributions would be used for the probabilities, the frequentist and Bayesian FE model would yield the same results. However, normal distributions are not appropriate to model probabilities, because probabilities are necessarily restricted to the interval [0, 1]. Therefore we assumed beta distributions for the

Table 4.1: Mean (SE†) monthly exacerbation probabilities of the reference treatment, before and after re-estimation.

	Before re-estimation	Frequentist		Bayesian	
		Fixed effects	Random effects	Fixed effects	Random effects
P (exacerbation)					
Moderate COPD	.051 (.004)	.050 (.003)	.050 (.003)	.050 (.003)	.051 (.005)
Severe COPD	.075 (.003)	.070 (.003)	.075 (.007)	.072 (.003)	.077 (.018)
Very severe COPD	.096 (.005)	.089 (.004)	.107 (.016)	.095 (.005)	.109 (.044)
P (severe exacerbation)|(exacerbation)					
Moderate COPD	.097 (.024)	.101 (.019)	.109 (.029)	.121 (.021)	.118 (.040)
Severe COPD	.136 (.018)	.103 (.012)	.117 (.026)	.118 (.013)	.120 (.045)
Very severe COPD	.192 (.027)	.176 (.021)	.176 (.021)	.178 (.021)	.179 (.028)

† For the Bayesian analyses the standard deviations of the posterior distributions are shown

probabilities, yielding slightly different results for the Bayesian model and the frequentist inverse variance method.

Figure 4.1 graphically displays the differences between the meta-analyses in the probability of getting an exacerbation when having moderate, severe and very severe COPD in the reference treatment arm. The top three estimates are the estimates obtained from the trials that were the basis for the original model while the fourth estimate reflects the new data. Clearly, the probability of having an exacerbation in the second trial was found to be markedly higher than in the other three trials. However such a clear discrepancy between trial 2 and the other trials was not found in other parameters, such as the transition probabilities or the probability that an exacerbation was severe. It can be seen in figure 4.1 that the random-effects specifications put more weight on the second trial when estimating the combined mean value than the fixed-effects specifications do. As we expected, the Bayesian RE models resulted in greater uncertainty surrounding the means than the frequentist RE models.

To assess the fit of the Bayesian models we calculated the residual deviance for the Bayesian fixed effects and random effects models for the probability to experience an exacerbation. The results are presented in table 4.2.

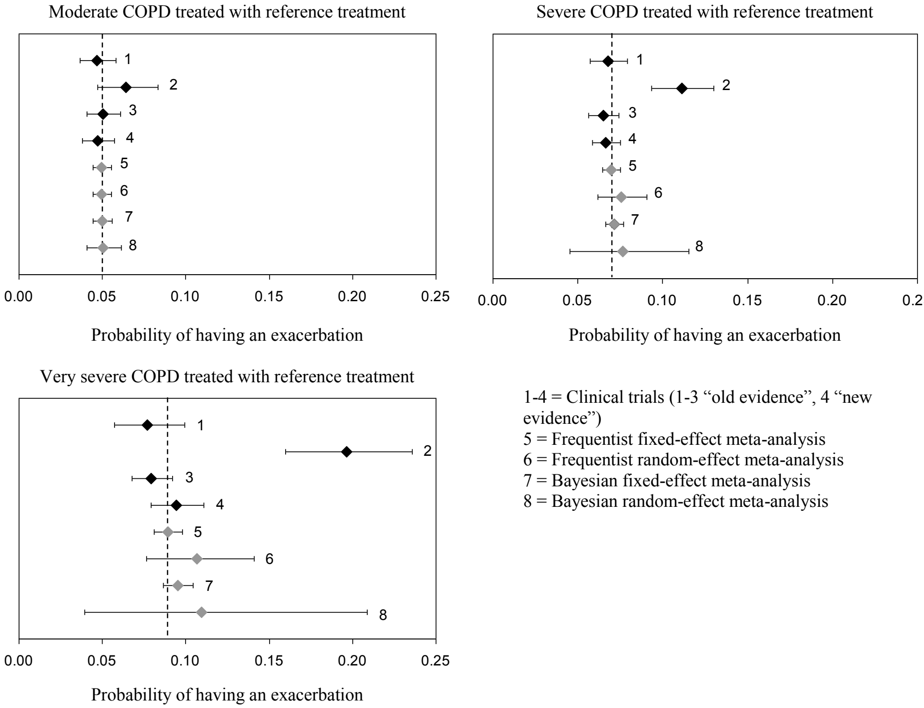


Figure 4.1: Meta-analyses of exacerbation probabilities.

Table 4.2: Residual deviance of the Bayesian models for the monthly exacerbation probabilities of the reference treatment.

	Moderate COPD	Severe COPD	Very severe COPD
P (exacerbation)			
fixed effects	4.24	27.70	48.83
random effects	3.86	4.41	4.18
P (severe exacerbation)|(exacerbation)			
fixed effects	5.55	13.61	2.03
random effects	4.40	5.02	2.24

For moderate COPD the residual deviance of the Bayesian fixed effects and random effects models are very similar and close to the number of data points, indicating that both models fit equally well. However, for severe and very severe COPD the random effects model clearly outperforms the fixed effects model. The Bayesian random effects models also perform better for the probabilities of the exacerbation being severe.

Transition probabilities

The monthly transition probabilities for months 2 to 12 from the first year for the frequentist and Bayesian analyses are presented in table 4.3. As indicated in the table, there were a number of transition probabilities for which the between-study heterogeneity was too small for the Bayesian random-effects model to converge. These probabilities were entered as fixed-effects in the model. In other words, for these transitions the random-effects model collapsed into a fixed-effects model.

Table 4.3: Mean (SE[†]) 1st year monthly transition probabilities between the states in the Markov model for the reference treatment from the frequentist and Bayesian meta-analyses.

	Frequentist Fixed-Effects				Frequentist Random-Effects			
	Moderate	Severe	Very sev.	Death	Moderate	Severe	Very sev.	Death
Moderate	.961 (.008)	.037 (.008)	.000 (.000)	.002 (.002)	.962 (.008)	.037 (.008)	.000 (.000)	.002 (.002)
Severe	.019 (.004)	.961 (.006)	.017 (.004)	.003 (.002)	.019 (.004)	.962 (.006)	.017 (.004)	.003 (.002)
Very sev.	.000 (.000)	.037 (.009)	.960 (.010)	.003 (.003)	.000 (.000)	.037 (.009)	.960 (.010)	.003 (.003)
	Bayesian Fixed-Effects				Bayesian Random Effects			
	Moderate	Severe	Very sev.	Death	Moderate	Severe	Very sev.	Death
Moderate	.957 (.008)	.039 (.008)	.002 (.002)	.002 (.002)	.957 (.013)	.039 (.012)	.002 (.002)	.002 (.002)*
Severe	.020 (.005)	.958 (.007)	.020 (.005)	.003 (.002)	.020 (.007)	.957 (.017)	.020 (.016)	.003 (.002)*
Very sev.	.005 (.003)	.039 (.009)	.953 (.010)	.003 (.003)	.005 (.003)*	.039 (.016)	.953 (.016)	.003 (.003)*

* Parameters in the Bayesian random-effects model that are entered as fixed-effects due to small between-study heterogeneity

[†] For the Bayesian analyses the standard deviations of the posterior distributions are shown

Table 4.4: Residual deviance of the Bayesian models for 1st year monthly transition probabilities between the states in the Markov model for the reference treatment.

	Bayesian Fixed-Effects			
	Moderate	Severe	Very sev.	Death
Moderate		2.02	4.40	1.13
Severe	2.33		4.17	1.06
Very sev.	3.48	1.69		1.11
	Bayesian Random-Effects			
	Moderate	Severe	Very sev.	Death
Moderate		2.41	4.44	1.15
Severe	2.46		3.68	1.07
Very sev.	3.48	2.05		1.12

Just as for the exacerbation probabilities, the fixed-effect meta-analysis also produced the lowest means and SEs for the transition probabilities. However, there is little difference in mean transition probabilities between the fixed-effect and random-effect meta-analysis because the lack of heterogeneity meant that the same method was used for the random effects procedure as for the fixed effect procedure. Regarding the standard errors, the most notable result is the great increase in standard errors when applying Bayesian random-effect meta-analysis. These standard errors are up to 2.8 times higher than in the other approaches.

The residual deviance for the Bayesian fixed effects and random effects models for the 1st year monthly transition probabilities of the reference treatment are presented in table 4.4. As can be seen in this table, the residual deviances are very similar and therefore the random effects model fits the data only marginally better.

As stated previously, the re-estimated model parameters for the comparator treatments were obtained indirectly via relative risks. These indirectly obtained re-estimated probability distributions were jointly used with the directly re-estimated probabilities for the reference treatment in the analysis on the impact of method of data synthesis on cost-effectiveness.

Cost-effectiveness

Table 4.5 shows the impact of using different types of meta-analysis on the estimated differences in cost and outcomes between treatment options. In all meta-analyses except the Bayesian random-effect meta-analysis, the reference treatment remained dominant compared to the two comparators as was the case in the original model. The results for

Table 4.5: Results for the three treatment arms in the CE model: Mean (SE) Costs and QALYs before and after re-estimation

		Reference		Comparator 1		Comparator 2	
		Costs	QALYs	Costs	QALYs	Costs	QALYs
Before re-estimation		7386 (515)	3.340 (.107)	7606 (841)	3.253 (.171)	8326 (1227)	3.251 (.226)
Frequentist	FE	6975 (417)	3.351 (.091)	7006 (613)	3.272 (.145)	7652 (976)	3.254 (.192)
	RE	7290 (504)	3.348 (.090)	7456 (793)	3.245 (.191)	8295 (1234)	3.245 (.191)
Bayesian	FE	7261 (461)	3.338 (.145)	7284 (682)	3.267 (.141)	8214 (1123)	3.244 (.196)
	RE	7407 (724)	3.344 (.099)	7392 (1185)	3.296 (.160)	8387 (1699)	3.260 (.283)

FE = Fixed-effects
RE = Random-effects

the reference treatment in terms of costs, QALYs and costs per QALY are slightly less favorable in the re-estimated models compared to the original model. In all meta-analyses, the probability that an exacerbation is severe increases in moderate COPD, whereas it decreases in severe and very severe COPD. In comparator 1 the effect of the latter decrease is relatively greater than in the reference treatment because in comparator 1 more patients move towards severe and very severe COPD. In the reference treatment the effect of the increasing probability of experiencing a severe exacerbation when having moderate COPD is relatively greater than in comparator 1 because more patients remain in moderate COPD. In addition, the matrix of transition probabilities that is generally more favorable for the reference treatment than the other two treatments becomes slightly less favorable when adding the new data. Overall, these effects lead to a slight worsening of incremental results of the reference treatment compared to comparator 1. This worsening is most pronounced when updating the model input using Bayesian random-effect meta-analysis. The much larger standard errors for the transition probabilities and exacerbation probabilities in the Bayesian random-effect meta-analysis are the reason for the more pronounced effect. These large SEs cause skewness in the beta distributions used, causing a shift in the point estimate for outcomes such as QALYs, even when the input point estimates differ little between synthesis approaches. This effect is larger for comparator 1 than for the reference treatment, since the SEs for comparator 1 are larger than the SEs for the reference treatment.

Figure 4.2 shows the cost-effectiveness acceptability frontiers resulting from each type of meta-analysis. In the base case the reference treatment has the highest expected net benefit for all values of the willingness to pay for a QALY above 0. The same is true for the meta-analyses except for the Bayesian random-effect meta-analysis where the reference treatment has the highest expected net benefit for threshold values above €320 per QALY. If the threshold is below this value, comparator 1 has the highest expected net benefit.

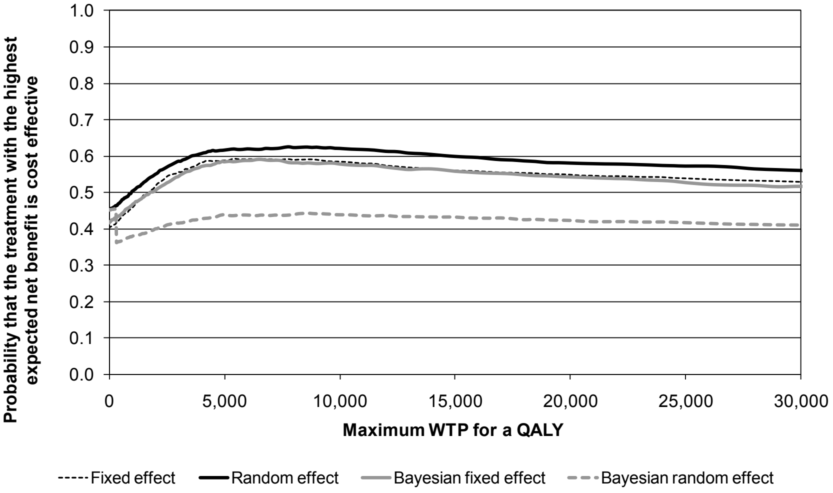


Figure 4.2: Cost effectiveness acceptability frontiers based on the four meta-analysis methods.

The cost-effectiveness acceptability frontier was about 15% lower for the Bayesian random-effect meta-analysis than for the other meta-analyses. This is due to the greater amount of uncertainty around the parameter estimates. As a result of this uncertainty, there will be more variability in the results of the model runs, i.e. there will be more runs where comparators are favored over the reference in terms of expected net benefit. The acceptability frontier resulting from the frequentist random-effect meta-analysis lies above the acceptability frontier of the other meta-analyses because the point estimate of the ICER was driven by the favorable exacerbation pattern for the reference treatment in this analysis.

Conclusions and discussion

In this study we have compared four different methods of meta-analysis and found that the estimates of three groups of model parameters, i.e. the probabilities of having an exacerbation, the probabilities that the exacerbation is severe and the transition probabilities, can vary considerably depending on the method used. Not only the estimates of the mean parameter values were affected but also, and more prominently, the estimates of the standard errors. We found up to nine-fold differences in standard errors of the exacerbation probabilities and up to almost three-fold differences in standard errors of the transition probabilities. These differences were found for the Bayesian random-effect meta-analysis, the method that was most different from the other methods.

Nevertheless, in this particular study, the impact of the different methods on the estimated differences between bronchodilators in costs and QALYs is relatively limited. This is partly because the probabilities that are most sensitive to the choice of meta-analysis are the 'small' probabilities (e.g. the probability of moving from moderate to very severe COPD), which by themselves do not drive the cost-effectiveness outcomes. It is also due to the fact that we only had new evidence for the reference treatment to add and no new evidence on the relative treatment effect of the reference treatment versus the comparators. It is likely that if we would have had new evidence on the difference between the reference treatment and the comparators, or if the new data would have been less well in line with the previous studies, the impact of the choice of meta-analysis method would have been more substantial. However for cost-effectiveness models it will be quite a common occurrence that data synthesis will only be required for parts of model as new evidence may only affect selected aspects.

Compared to the original CE model results, re-estimating the model using the Bayesian random-effect meta-analysis led to the greatest change in cost-effectiveness estimates. Not only were the point estimates most different, but also the uncertainty surrounding these estimates. The cost-effectiveness acceptability frontier, showing the probability that the expected Net Monetary Benefit of the reference treatment is below the maximum willingness to pay for a QALY (ceiling ratio or threshold) was roughly about 15% lower for the Bayesian random-effect meta-analysis than for the other meta-analyses. The reference treatment always had the highest expected net benefit, except in the Bayesian random-effect meta-analysis, where for very small values below €320, comparator treatment 1 had the highest expected net benefit. The uncertainty around the cost-effectiveness is greatest when using the Bayesian random-effect meta-analysis. This is because in Bayesian RE meta-analysis the total uncertainty is based on the combination of uncertainty from three sources as opposed to one or two sources: 1) the between study heterogeneity of the data, 2) the uncertainty associated with the priors for the model parameters and 3) the uncertainty associated with the priors for the between study heterogeneity.

In general, there are more model specifications possible than the four we have used in the current study. All four of our models were univariate, as opposed to multivariate, which are models in which the outcome measures (exacerbation and transition probabilities) are analyzed jointly, thereby also revealing information about the correlations between the multiple outcome variables. To apply multivariate meta-analysis the estimated vector of outcome measures along with the corresponding estimated covariance matrix per trial is needed [23]. Since the overall database used in our study and the number of underlying patient samples was limited, it was not possible to perform a multivariate meta-analysis. Also, the data did not allow the specification of hierarchical models beyond random-effects or meta-regression models. In a meta-regression adjustments are made for characteristics of the different trials that could be associated with differences in the observed parameter values. With such small numbers it is impossible to ascertain whether the study-specific covariates really explain the heterogeneity between the studies and the likelihood to find false positive results is high.

The results of this study show that the choice of method to derive a pooled estimate of the model parameters is an additional source of uncertainty that is commonly not parameterized and therefore not included in a probabilistic sensitivity analysis and a value of information analysis. The question is: "Should it be parameterized, should it be part of a separate sensitivity analysis or is there a way to decide what method is best suitable for a given set of studies?" Although doing a sensitivity analysis on method of meta-analysis is wise, the answer is probably the latter.

If studies are homogeneous, the choice between using a frequentist or a Bayesian method is unimportant. If they are not, random-effect meta-analysis accounts for the heterogeneity, but it may be more important to examine the reasons for lack of homogeneity in order to decide whether pooling in itself is legitimate. Homogeneity should be considered jointly for all parameters that need to be estimated. For example, based on figure 4.1 one could conclude that trial 2 is an outlier and that therefore a fixed effects model would be most appropriate to model the exacerbation probabilities. However, trial 2 is not an outlier for the other model parameters such as the probability of having a severe exacerbation conditional on having an exacerbation, so the assumption that trial 2 is an outlier does not hold. Therefore allowing the choice between fixed effects and random effects model specifications to be made on a per parameter basis, instead of on all information captured within the trials can lead to erroneous assumptions.

Besides this more conceptual argument on the choice for fixed effects or random effects, techniques are available for assessing which method to use based on the data itself. In this study we used the I^2-statistic for the frequentist approach and residual deviance for the Bayesian approach. They indicated that for the probability of experiencing an exacerbation and the probability of the exacerbation being severe the random effects models perform better. However this difference was not found for the transition probabilities.

In our current study data pooling was legitimate given the similarity between studies with respect to study design and patient population. To illustrate this, the small difference between the fixed-effect and random-effect meta-analysis of transition probabilities is explained by the very small random-effect variance, i.e. the very small variance between the true underlying transition probabilities in each study.

The choice between the frequentist and Bayesian approach might be driven by ones preference, since frequentist and Bayesian statistics can be seen as two opposing philosophies about statistics. In practice, health economists have already more or less adopted the Bayesian philosophy with the probabilistic decision analytic framework in which parameters are thought of as having probability distributions [24]. As such, standard probabilistic sensitivity analysis is essentially Bayesian. Thus, unless one is a convinced frequentist, we would argue that a Bayesian approach is in general preferable. In addition, the Bayesian approach offers great flexibility in analysing data from various distributions, and when performing a random-effect analysis the Bayesian approach captures the between study heterogeneity more completely.

When opting for the Bayesian approach it is important to carefully address the choice of priors, especially when using an informative prior. When applying the Bayesian random-effect model, it is particularly important to realize that when the number of studies combined is small, the choice of the prior for the between-study variance may critically affect the analysis. This is not a weakness of the Bayesian approach but merely a reflection of the true uncertainties inherent in the problem of combining information from diverse sources. For example, the fixed-effect approach to combining information is equivalent to assuming that the prior distribution of the between-study variance is concentrated at or very near 0 and any justification for applying the fixed-effect model in a given situation is exactly equivalent to justifying the use of such a prior distribution. The use of a prior distribution for the between-study variance is best thought of as a compromise between opposing philosophies about meta-analysis: those who believe that that variance is near 0 (the philosophy of a fixed-effect meta-analysis) and those who believe that the between-study variance is large and borrowing strength is hopeless in most cases (the "you can't combine apples and oranges" philosophy). An open-minded prior distribution should assign significant prior probabilities that either philosophy could be right for any given problem. Additionally, the meta-analysis should be tested for sensitivity to alternative specification of the priors. The appropriateness of the use of a Gamma(0.001,0.001) distribution as a prior for the precision ($1/\tau^2_0$) has been debated in the literature [25,26]. Therefore we assessed the use of this prior specification in our study by comparing the results with those where uniform(0,10) was used as a prior for τ_0. It was found that in our case the Gamma(0.001,0.001) priors on the precision behaved better than the uniform(0,10) priors on the between-study standard deviation. This was because some of the probabilities were relatively small.

In conclusion, this study has demonstrated that the choice of method for the meta-analysis can affect resulting model parameter updates considerably. This can in turn affect the estimates of cost-effectiveness and the uncertainty around them, although in the current study the impact on the preferred treatment was limited and the results remained qualitatively the same as in the original model, before the re-estimation. In general, Bayesian methods are preferable. For fixed effect models, the Bayesian approach offers great flexibility in analysing data from various distributions. If a random effects specification is warranted, a Bayesian approach is in general more appropriate because the between study heterogeneity is captured more completely.

References

1. GOLD. Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for Diagnosis, Management, and Prevention of COPD. Bathesda MD: National Institutes of Health, National Heart, Lung and Blood Institute, Retrieved July 6, 2006 from http://www.goldcopd.com; 2005.
2. Rutten-van Mölken M, Lee TA. Economic modeling in chronic obstructive pulmonary disease. Proc Am Thorac Soc. 2006;3(7):630-4.
3. Oostenbrink JB, Rutten-van Molken MP, Monz BU, FitzGerald JM. Probabilistic Markov model to assess the cost-effectiveness of bronchodilator therapy in COPD patients in different countries. Value Health. 2005;8:32-46
4. Rutten-van Molken MP, Oostenbrink JB, Miravitlles M, Monz BU. Modelling the 5-year cost effectiveness of tiotropium, salmeterol and ipratropium for the treatment of chronic obstructive pulmonary disease in Spain. Eur J Health Econ. 2007;8:123-135
5. Oostenbrink JB, Al MJ, Oppe M, Rutten-van Mölken MP. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value in Health. 2008;11:1070-1080
6. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60:431-439
7. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986;7:177-188
8. Sutton AJ, Abrams K, Jones DR, al e. Methods for Meta-analysis in Medical Research. London: Wiley, 2000
9. Vincken W, van Noord JA, Greefhorst AP, Bantje TA, Kesten S, Korducki L, Cornelissen PJ. Improved health outcomes in patients with COPD during 1 year's treatment with tiotropium. Eur Respir J. 2002;19:209-216
10. Casaburi R, Mahler DA, Jones PW, Wanner A, San PG, ZuWallack RL, Menjoge SS, Serby CW, Witek TJ, Jr. A long-term evaluation of once-daily inhaled tiotropium in chronic obstructive pulmonary disease. Eur Respir J. 2002;19:217-224
11. Brusasco V, Hodder R, Miravitlles M, Korducki L, Towse L, Kesten S. Health outcomes following treatment for six months with once daily tiotropium compared with twice daily salmeterol in patients with COPD. Thorax. 2003;58:399-404
12. Scanlon PD, Connett JE, Waller LA, et al. Smoking cessation and lung function in mild-to-moderate chronic obstructive pulmonary disease. The Lung Health Study. Am J Respir Crit Care Med. 2000;161:381-390.
13. Bateman E, Singh D, Smith D, Disse B, Towse L, Massey D, Blatchford J, Pavia D, Hodder R. Efficacy and safety of tiotropium Respimat SMI in COPD in two 1-year randomized studies. Int J Chron Obstruct Pulmon Dis. 2010;5: 197–208
14. Clinicaltrials.gov. Available from http://www.clinicaltrials.gov; NCT ID NCT00168844; Accessed on 22-07-2009

15. Clinicaltrials.gov. Available from http://www.clinicaltrials.gov; NCT ID NCT00168831; Accessed on 22-07-2009

16. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11): 1539-58

17. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557-560

18. Gelman A, Carlin JB, Stern HS ea. Bayesian Data Analysis. London: Chapman & Hall, 1995

19. Carlin BP, Louis TA. Bayes and emperical Bayes methods for data analysis. London: Chapman & Hall, 1996

20. Kreft IGG, Leeuw JD. Introducing multilevel modeling. London: Thousand Oaks, Calif.: Sage, 1998

21. Spiegenhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc B. 2002;64:1-34

22. Fenwick E, Briggs A. Cost-effectiveness acceptability curves in the dock: case not proven? Med Decis Making 2007;27:93-95.

23. Arends LR. Multivariate meta-analysis: modelling the heterogeneity. Mixing apples and oranges: dangerous or delicious? Ph.D. Thesis. Alblasserdam: Haveka BV, 2006

24. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, Lu G. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics. 2006;24:1-19

25. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006;1:515-533

26. Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis. 2006;1:473-514

Chapter 5

Statistical uncertainty in TTO derived utility values

Mark Oppe, MSc[1], Siem Oppe, MSc[2], Frank de Charro, PhD[3]

1. Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

2. Dutch national road safety research institute SWOV, Leidschendam, The Netherlands

3. EuroQol Group Executive Office, Rotterdam, The Netherlands.

Submitted for publication

Introduction

Generic utility measures have been developed and used to capture health related quality of life in a single summary index score, called a utility. These utilities are anchored on full health = 1 and death = 0. The quality weights can then be multiplied with life years gained to form the QALY, a measure of health that includes both quantity and quality of life [1]. Using the QALY as an outcome measure allows policy makers to compare the effectiveness (or – more precisely – cost utility) of treatments across different diseases. This is necessary when making budget allocations for the health care sector as a whole under budget constraints.

Three of the most widely used utility measures are the EQ-5D [2], the SF-6D [3,4] and the HUI [5,6]. Valuation studies have been performed for each of these instruments where utility weights were derived for the health states described by the instruments. For all instruments, the utility values for health states are in general presented (and subsequently used) as point estimates [1].

However, utilities are based on empirical valuation studies in which statistical models are used to estimate the utilities associated with health states on the basis of subsets of states. Because of the empirical nature of these valuation studies it is not clear to what extent the uncertainty in the parameter estimates is due to random error and model imprecision. In this paper we focus on two issues. Firstly, how can we quantify the effect of these two error components? Secondly, what is the impact of the sample size and number of observed health states on this uncertainty in the valuation studies used to derive the utilities? These questions are important for the design of new valuation studies Three components of uncertainty were identified and investigated in this study:

1. The error due to the different responses for a specific health state given by the respondents in the valuation study (i.e. response heterogeneity). This should NOT be included in the error of the utilities, because this is random error (i.e. people will give different answers and we are interested in the mean value of the general population, not the individual answers of respondents). In ANOVA terms this is the within-variance [7].
2. The error/uncertainty due to differences between the sample means and the population means. A possible source of this error is the amount of bias of the sample. This type of error can only be investigated by comparing the outcomes to those of another representative random sample. E.g., if the sample is drawn from a sub-group of patients with a particular disease, there is a possibility of bias. However, even for an unbiased sample there will be a difference between the mean health state valuations and the mean population values. The standard error of the mean for the observed

health states could be used as a measure for this difference, under the assumption of a non-biased sample. This standard error depends on the sample size and will be larger, the smaller the sample.

3. The error due to the use of a misspecified or imprecise model. Whether a model is precise enough can be investigated by assessment of the percentage of the between-variance[†] that is explained by the model [7]. One possibility is that the model as such is correct and the error/uncertainty is solely due to the fact that not all health states were observed but that interpolation was used to estimate the values for the unobserved health states. Comparisons between models with varying numbers of observed states allows for quantification of this error. Furthermore, the within-variance can be used to estimate the variance of the health state mean for each health state and to check whether the differences between observed and predicted means are within acceptable limits, under the assumption of a perfect model.

In addition to the within and between variances, the mean absolute error (MAE) and the goodness of fit statistic (R^2) can be used as measures for the performance of a model. MAE is the mean of the absolute values of the differences between observed and the estimated values. R^2 indicates the proportion of error in the individual valuations reduced by the model, that is the ratio of explained and total variance in the individual scores. It is less appropriate for a check on the correctness of the model, because it includes the within variance described under point 1. I.e. a model that fits (almost) perfectly on the observed mean values of the states might still have a low value of R^2. The same model applied to the mean values instead of the individual values will show this.

Of course, other sources of uncertainty exist. For example, in the regression analyses that are commonly used to estimate the utility models it is assumed that the observed states have normal distributions, which might not be the case (e.g. floor and ceiling effects may be present). Furthermore there is the possibility of interviewer bias, and the valuations themselves are susceptible to change over time [8]. However, these biases as well as the possible bias in the sample will not be taken into account here, as they are associated with the valuation task and format and are not statistical in nature.

In this study we focus on the uncertainty associated with EQ-5D based utilities. The EQ-5D is a generic measurement instrument to describe and value health states [2]. The EQ-5D classification describes health states according to five attributes: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression. Each attribute has three levels: 'no problems'; 'some problems'; and 'severe problems'. Health-state descriptions are con-

[†] Within variance is the variance of the individual health states originating from the different answers of respondents. Between variance is the variance of the mean values over all observed health states

structed by taking one level for each attribute, thus defining 243 (3^5) distinct health states, where '11111' represents the best and '33333' the worst state.

We used data from two EQ-5D valuation studies, the UK MVH study and the Dutch TTO study [9,10]. Both studies used Time-Trade-Off as elicitation technique for obtaining utilities. The UK study included 42 observed health states and had a sample size n = 2997. The 43rd state that was evaluated in the MVH-study was "unconscious". This state was omitted in our analyses as it is not part of the EQ-5D classification system. The Dutch study included 17 states (a subset of the 42 from the UK study) and n = 298. We assumed that samples used in both valuation studies were unbiased representations of the national populations.

Methods

The quality of the regression model that resulted from the MVH study, which model was also used in the Dutch valuation study, is the main object of investigation in this paper. The regression equation and parameter estimates that resulted from the MVH study are taken as starting point for this investigation. In order to find out to what extent the model is wrong or imprecise, an ANOVA check was carried out to see whether the parameter estimates are within the error bounds under the assumption of a perfect model. Firstly, the between-within variances of the health state valuations are estimated without any model assumptions. Secondly, using these estimates, we investigated to what extent the between variance is reduced by the model.

Next a Monte Carlo simulation was carried out to find out to what extent uncertainty in the value sets is influenced by the sample sizes and the number of health states used in the studies. Starting from the same health states as used in the original study, for each health state random samples are drawn from a normal distribution with a mean health state value deduced from the Dutch and MVH models and variance equal to those of the observed health state valuations of each model. In other words, we combined the modelled means with the observed variances to generate samples in our simulation. Each health state was evaluated an equal number of times. This number was equal to 25, 50, 100, 298 (the number of participants in the Dutch study), 500 and 1000. Random sample values outside the range from +1 to −1 were truncated.

To each simulated sample the same linear model was applied as in both original studies. The parameters were estimated and the model estimates were computed for each health state. There were one hundred simulations for each sample, resulting in distributions of values for each health state. Because the model predictions are used as 'population' means for each sample, a 'perfect' model is guaranteed for the outcomes. This way, the uncertainty in the mean health state values could be investigated as a function of sample size. Lastly, we varied the number of health states to study the impact that number of observed health states has on the uncertainty of the model parameters. The sets included 14, 17, 21, 28, 35 and 42 health states with n = 1000. They were selected to be a subset of the original 42 health states from the MVH study chosen in such a way that they were spread out over the full spectrum of health states. Data were analysed using SAS v8.2.

Results

For the UK data the R^2 of the original model was 0.43. At first sight this may seem low. However, the F value[$] was high at 801 and the mean absolute error (MAE) was low at 0.039 [9]. This implies that the model fit was better than one would expect on the basis of the R^2 value. When the model was fitted to the mean values only (thereby removing the within variance) the R^2 was 0.98. This shows e.g., that addition of interaction parameters will not improve the model significantly. The residuals, using the Dolan model as a perfect model (i.e. using the same standard errors to determine the CI's where the residuals should be between) are shown in figure 5.1. As can be seen the residuals lie mostly within the CI,

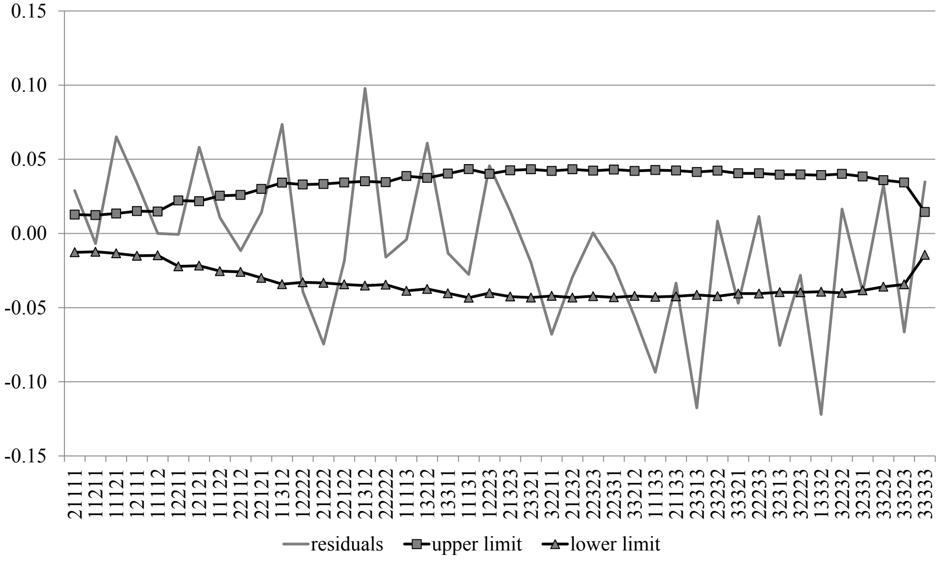


Figure 5.1: Residuals with 95% CI's using the Dolan regression model as a perfect model. EQ-5D states are ranked from best (left) to worst (right).

[$] This is the ratio of the within and between variance

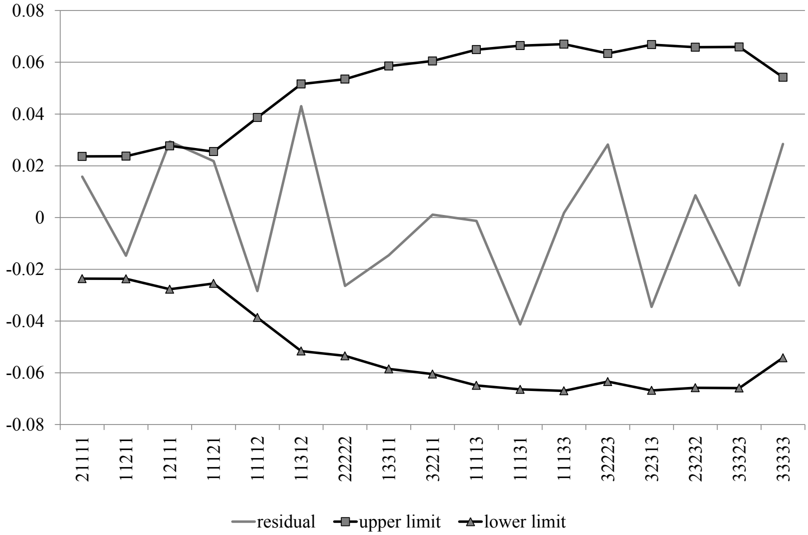


Figure 5.2: Residuals with 95% CI's using the Lamers regression model as a perfect model. EQ-5D states are ranked from best (left) to worst (right).

indicating that the model fits the data well. Figure 5.2 shows the same situation for the Dutch data. Because the sample size is smaller the 95% CI is wider, and the residuals all fall within the boundaries.

The standard deviation over the 100 Monte Carlo simulations was calculated for each health state included in the simulation. The uncertainty of the utility values in a single simulation varied with varying sample sizes used. Therefore, the standard deviation over the 100 simulations also varied with sample size and is an estimate of the uncertainty of the utility values due to sample size.

In figure 5.3 the number of states in the models for the UK was 42 and for the Netherlands 17. The mean SD of the index values was taken over all states. As can be seen the uncertainty for the UK is lower than for the NL and in both cases the uncertainty decreases with increasing sample size.

The impact of the number of health states on the parameter estimates and uncertainty surrounding the parameter estimates at an N = 1000 of the model is shown in figure 5.4. In figure 5.4 it can be seen that the uncertainty decreases substantially with increasing number of health states, however, the percentage of change is also in the worst case of 14 health states still relatively small.

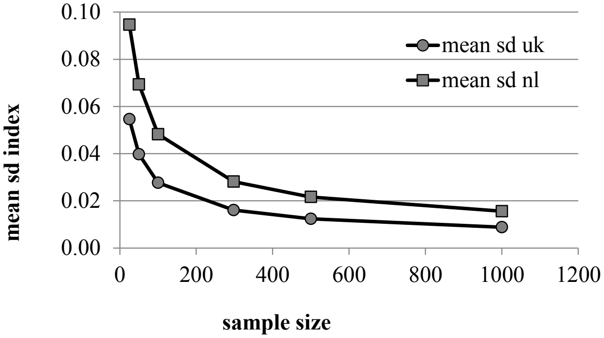


Figure 5.3: Effect of sample size on the mean uncertainty surrounding the utilities (N = 25, 50, 100, 298, 500, and 1000) for the UK with 42 observed states and the Netherlands with 17 states.

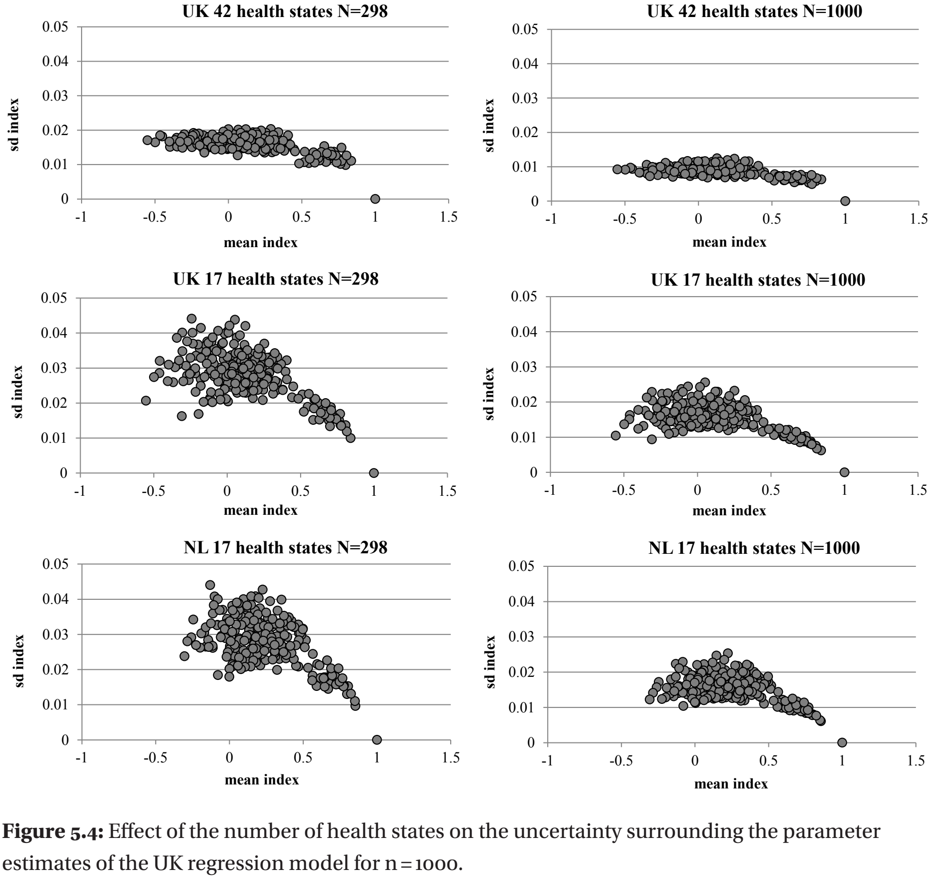


Figure 5.4: Effect of the number of health states on the uncertainty surrounding the parameter estimates of the UK regression model for n = 1000.

In figure 5.5 a comparison is made between the UK and the NL for individual health states. This figure shows the impact of the ceiling effect on the uncertainty surrounding the health states (the uncertainty is smaller for better health states). In the 2 graphs with 17 health

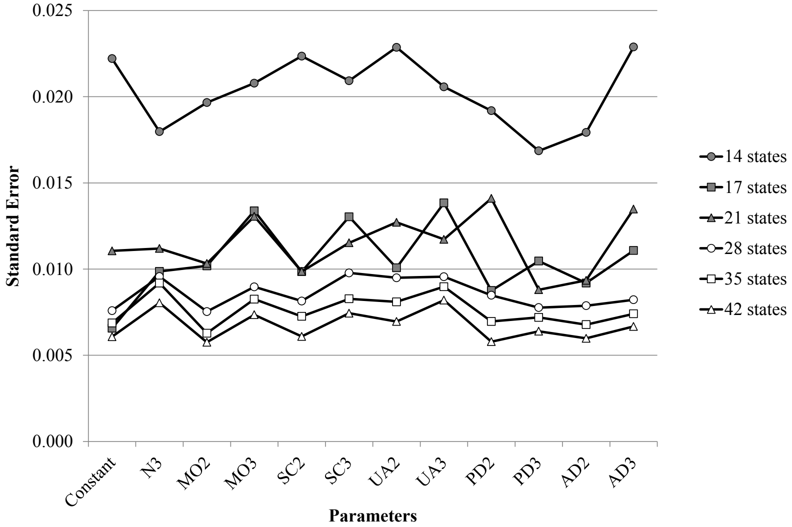


Figure 5.5: Comparison of uncertainty surrounding all 243 EQ-5D health states based on the Dutch and UK utilities for 17 and 42 health states and 298 and 1000 observations per state.

states and N = 1000 a floor effect can be observed, although this is smaller than the ceiling effect. Finally, from this figure can be concluded that increasing the number of health states from 17 to 42 reduces the uncertainty more than increasing the number of observations per health state from 298 to 1000.

Conclusions and discussion

In this study we analysed the uncertainty bounds of the utilities from two of the EQ-5D value sets: The Dutch TTO study and the UK MVH study. In order to assess the uncertainty of the estimate for each health state, the within variance was removed from the models. Also, the impact of the number of respondents and health states on the uncertainty was investigated using Monte Carlo simulation.

For the MVH-study, as well as the Dutch study the chosen values for the number of states and respondents resulted in outcomes that were well within the error bounds. That is, the observed means did not differ significantly from the means predicted by a perfect model. Therefore, the choice of a perfectly fitting model in the Monte Carlo simulation seems acceptable.

As expected, the Monte Carlo results show that increasing the number of respondents per state decrease the uncertainty surrounding the utility values. Also, increasing the number of health states in a valuation study will lessen the uncertainty margins of the utilities. The standard error of the mean utilities, averaged over all health states, ranged from 0.095 when 17 states were valued by 25 respondents to 0.009 when 42 health states were valued by 1000 respondents. The uncertainty for individual health states using the simulated Dutch data for these numbers of respondents and states used in the Dutch valuation study ranged from 0.044 to 0.010. The mean uncertainty over all health states was 0.028. The uncertainty for individual health states using the simulated MVH data at the number of respondents and states used in the MVH valuation study ranged from 0.013 to 0.006. The mean uncertainty over all health states was 0.009.

In probabilistic cost utility analyses the uncertainty surrounding the utilities originates from the fact that not all patients indicate that they are in the same health state. It is therefore only associated with the heterogeneity of the patient population in the study. Our findings imply that apart from this heterogeneity the inherent uncertainty of the utility estimates should also be taken into account in cost utility analyses. Looking at the magnitude of the uncertainty it becomes clear that for small valuation studies the uncertainty surrounding the utilities can get large enough to have a possible impact on CUA results if the differences in utility values between the treatment arms under consideration are small. When these value sets are used in a CUA the differences in effect might prove to be not significant and could influence policy makers' decisions on whether or not to reimburse certain procedures or drugs.

We recognise that this study has several limitations. In the Monte Carlo simulation the uncertainty was modelled using the magnitude of the standard error from the observed data. It was assumed that the uncertainty followed a normal distribution and was truncated at +1 and –1. This assumption might not reflect the actual uncertainty distribution close to the ends of the scale. Also the number of simulations that was run was 100, which is on the small side. We feel confident that these issues might only have a small impact on our results. However, we only looked at uncertainty that was statistical in nature. Other sources of uncertainty such as interviewer bias, bias due to the fact that the sample might not be representative of the population and the change over time of respondents values were not taken into account. Therefore our estimates should be considered as conservative.

In a saturation study (i.e. a study where all 243 possible EQ-5D health states are valued by respondents) the uncertainty that should be incorporated is the standard error of the mean of each health state. In the MVH study with $N \approx 750$ per state the mean standard error was 0.017. For the Dutch study with $N = 298$ the mean observed standard error was 0.026. These values are of similar magnitude as the uncertainty from the models. Therefore, for a saturation study you would need far more observations in total than for a modelling study while obtaining a similar level of uncertainty for each health state.

Based on our results we suggest that for EQ-5D TTO valuation studies one would need between 35 and 42 health states (i.e. about 15% to 17% of the total number of possible states) to be valued by between 300 and 500 respondents in order to obtain uncertainty margins surrounding the utilities smaller than 0.01 on average. This is about 1/3 of the magnitude of the smallest mean parameter estimate of both the Dutch and the UK models. For both models this is for the dimension usual activities at level 2 (UA2). For the Netherlands UA2 = 0.032 and for the UK UA2 = 0.036. Also this is well below the minimum important difference (MID) obtained with EQ-5D when utilities were used as point estimates. Pickard et al. found an MID of 0.08 when EQ-5D is used in cancer [11] while Walters and Brazier found slightly lower values in other disease areas, 0.074 on average [12].

Expending the number of levels of the EQ-5D from 3 to 5 will increase the number of possible health states from 243 to 3125. Care has to be taken to include enough health states and enough observations per health state in a 5 level valuation study. If too few states or observations per state are used the uncertainty surrounding the utilities might become so big that the utility values themselves will not be significantly different from one another. This would mean that the full range of 3125 possible health states would not be used. Other valuation techniques than TTO such as VAS and discrete choice experiments, have other underlying mechanisms of generating utility values and therefore of creating the associated uncertainty. This should be taken into account when considering the use of such techniques to generate utilities.

As Szende and Schaefer [13] indicate in their paper, mapping techniques are expected to be used more frequently to obtain utility values. When mapping from a disease specific questionnaire to for instance EQ-5D is done, an additional source of uncertainty is generated. This additional uncertainty should also be included in probabilistic models and could prove to be large compared to the uncertainty from the utilities from EQ-5D and the uncertainty from trial data itself.

Lastly, the whole issue of uncertainty does not limit itself to the effect side of cost utility analyses. The same arguments could be made for the cost side, where commonly the uncertainty is included only from differences in resource use of individual patients while it is assumed that the unit costs are deterministic, and therefore perfectly known.

References

1. Drummond MF, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programmes. Oxford University Press, 3rd edition 2005.
2. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53-72.
3. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21(2):271-92.
4. Kharroubi S, Brazier JE, O'Hagan A. Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. Soc Sci Med. 2007; 64(6):1242-1252.
5. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system: Health Utilities Mark 2. Medical Care. 1996;34:702-722.
6. Feeny D, Furlong W, Torrance GW, Goldsmith CH,Zhu Z, DePauw S, et al. Multiattribute and single-attribute utilityfunctions for the Health Utilities Index mark 3 system. Medical Care. 2002;40:113-128.
7. J.P.Guilford, Fundamental statistics in psychology and education, p.278vv, McGraw-Hill Inc, New York, 1965.
8. Macran S, Kind P. Valuing EQ-5D health states using a modified MVH protocol: Preliminary results. In: Badia X, Herdman M, Roset M, editors. 16th Plenary Meeting of the EuroQol Group. Sitges, 6-9, November 1999. Discussion Papers. Institut de Salut Pública de Catalunya, Spain, 2000:205-240.
9. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-1108.
10. Lamers LM, McDonnell J, Stalmeier PFM, Krabbe PFM, Busschbach JJV. The Dutch Tariff: Results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121-1132.
11. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes. 2007;5:70.
12. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14(6):1523-1532.
13. Szende A, Schaefer C. A taxonomy of health utility assessment methods and the role for uncertainty analysis. Eur J Health Econ. 2006;7(2):147-151.

Chapter 6

Comparison of EQ-5D and Oxford Hip Score: implications for mapping

Mark Oppe, MSc[1], Nancy Devlin, PhD[2], Nick Black, MD[3]

1. Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

2. Office of Health Economics, London, UK

3. London School of Hygiene & Tropical Medicine – Department of Health Services Research & Policy, London, UK

Publication:

Oppe M, Devlin N, Black N. Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. Value Health. 2011;14(6):884-91.

Introduction

There is increasing interest in the routine use of patient reported outcome measures (PROMs) to audit the impact of health care and compare the performance of health care providers [1]. There are two principal types of PROM: generic and disease-specific. One of the most commonly used generic measures is the EQ-5D which describes health in terms of five domains each with three response levels. This results in potentially 243 health states [2]. It was developed for the purpose of combining descriptions of health states with information about their values based on social value sets, which show the index value (also referred to as the weight, or 'utility') for each state, anchored at 1 for full health and 0 for dead. Although anchored on dead, the EQ-5D does allow states to be considered worse than dead (i.e. utility <0). Utilities are generated using stated preferences techniques, such as time trade off (TTO) and are elicited from members of the general public [3]. They are used to calculate the incremental Quality Adjusted Life Years (QALYs) gained by patients as a result of a treatment. This is used as the basis for comparing the cost effectiveness of different treatments in terms of the incremental cost per QALY gained.

In contrast, disease-specific measures (DSM) have been developed to provide more detailed information on the condition of patients in a specific patient group. For example, the Oxford Hip Score (OHS) has 12 items to assess symptoms and functional status (disability) with each item having five possible answers, resulting in over 244 million possible health states [4]. The summary score for a health state described by OHS is obtained by adding the levels of each item resulting in a score between 12 and 60, where 12 is the best outcome.

There are three principal differences between the EQ-5D and disease-specific measures. First, DSMs can include items related to domains that are not included in EQ-5D and hence will not be reflected in changes in the patient's EQ-5D utility score. Second, the availability of more items per domain in DSMs might result in greater sensitivity to change in health status. And third, while the scale properties of the items have been made explicit in the EQ-5D, this may not be true for DSMs.

When cost effectiveness analysis is needed and no generic utility measure was included in the clinical study, utilities may still be obtained by linking the DSM data collected to a generic utility measure. The utility weights from the EQ-5D might be linked to health states derived from a disease-specific measure in a mapping study. The most common way of estimating a mapping function is by comparing data from each instrument collected for the same population and to estimate the relationship between the two via regression, although other methods are also used [5,6]. One of the most straightforward techniques employed is to estimate a regression model where utility is the dependent variable and the disease-specific items are the independent variables [6] – this is referred to as a direct approach to mapping. An alternative approach is to use an indirect method. In this case, the DSM items are mapped on to the items of the generic measure [7]. When estimating a mapping model ideally the goal should be to provide a single universal model applicable in all situations. At the very least it should be generalisable to include 'out of sample predictions'. The appropriateness of a mapping approach therefore hinges on both the representativeness of the data and on the comparability of the information captured by both types of instruments [5].

The aim of this study was to assess the comparability of the information captured by the OHS and the EQ-5D and investigate the validity of obtaining utilities for the OHS via mapping.

Methods

We made use of data obtained from a prospective cohort of NHS patients in England undergoing unilateral hip replacement [8] recruited at 11 health care providers. Data were collected from 512 patients before undergoing hip replacement and from 444 patients six months after surgery. There were 37 missing values for the OHS and 23 missing values for the EQ-5D. The UK-TTO value set (MVH-A1 tariff [9]) was used to calculate the utilities for the responses of the EQ-5D descriptive system. This is the value set that was elicited in a UK general population sample in 1993 and is the most widely used. Because the pre-operation (pre-op) and post-operation (post-op) data were expected to vary according to disease severity, we defined disease severity categories based on the utilities from the UK-TTO value set. These categories comprised steps of 0.1 in the utility scale i.e. the top category was 1 to (but not including) 0.9 and the bottom category (most severe states) was −0.3 to (but not including) −0.4. Although the UK-TTO value set has a minimum value of −0.59, values below −0.35 were absent in our data set (i.e. none of the patients reported to be in the worst EQ-5D states).

We started our analyses by exploring the data to find the (dis)similarities of the pre- and post-op data and the instruments using the combined correlation matrix of the OHS and EQ-5D. The correlation matrix comprised the inter-item correlations for all items of both questionnaires. In addition the combined pre- and post-op data was investigated as was the change in health status between the two time points. Principal component analysis (PCA) was applied to explore and compare the underlying dimensional structure of the OHS data and EQ-5D evident in these data. The basic idea behind PCA is to investigate whether a number of items generate information about a more general underlying construct [10].

PCA determines these factors and the way items are associated by analysing the pattern of correlation. Items with relatively high inter-correlation are assumed to reflect the same construct, and items with low inter-correlation reflect different constructs. Eigenvalues are used to condense the variance in a correlation matrix. The eigenvalues of a construct represent the relative share of variance accounted for by this construct. The sum of the eigenvalues is equal to the number of items. In our case this was 17: five items of the EQ-5D plus 12 items of the OHS. If the items do not correlate with each other the eigenvalues will reflect only the variance in the original items and be equal to one for each item.

We carried out both exploratory and confirmatory PCAs. For the former we selected those constructs that had an eigenvalue >1 [11]. Since this is based solely on the inter-item correlations the meaning of the resulting constructs can be difficult to interpret. The number of constructs in the confirmatory PCA was derived from inspection of the items. We chose five constructs reflecting the five items of the EQ-5D. In order to obtain a more interpretable set of factors, varimax rotation – an orthogonal rotation of the factor axes – was used to rotate the factors of both the exploratory and the confirmatory PCA [10]. Varimax rotation is effectively a change of coordinates of the factor solution, which allows for improved differentiation and interpretability of the extracted factors. Apart from giving insight in the dimensional structure of both instruments, the PCA also provided a basis for the choice between a direct or an indirect mapping approach.

We assessed the predictive performance of mapping models between the OHS and EQ-5D. Three main effects OLS regression models were estimated, one based on the pre-op data, one based on the post-op data and one based on the combined data. In all three models the EQ-5D index was the dependent variable and OHS items were the independent variables. The aim of mapping is to find a single model with which to map all OHS responses onto EQ-5D utilities, irrespective of when they were collected. Therefore, the performance of the three models was tested on the pre-op data, the post-op data and the combined data. Finally, we also estimated a fourth model based on the combined data where the dependent variable was again the EQ-5D index, but where the independent variables included the 12 OHS items, the 12 OHS items squared and all 66 two way interactions from the 12 OHS items. This full model was proposed by Rowen et al. for mapping SF-36 to EQ-5D in order to test for non-linearity of the mapping function [12]. In order to determine which of the 90 hypothesised model parameters were statistically significant at the 95% level, we followed a two step schema. Firstly we did a forward, a backward and a stepwise regression on all 90 parameters. Next we removed the items that were excluded in all three regressions and ran the forward, backward and stepwise regression for a second time on the reduced set of items. Finally, we selected the best predictive model from this set of three. The mean absolute error (MAE) was used as the measure of predictive validity. The MAE is the absolute value of the difference between the observed EQ-5D index and the predicted EQ-5D index.

All analyses were carried out in SPSS version 16.0.

Results

Distribution of EQ-5D utilities

Of the 243 potential EQ-5D states, 52 were reported by the patients, covering most of the utility scale as defined by the UK-TTO value set for EQ-5D. The distribution of these 52 states over the disease severity categories (based on steps of 0.1 on the utility scale) for pre- and post-op data is shown in table 6.1.

None of the patients reported level 3 problems with mobility on the EQ-5D, either before or after their hip replacement surgery. Furthermore, only 1% of patients reported 'extreme

Table 6.1: Observed EQ-5D health states and range of EQ-5D utilities per disease severity category (pre- and post-op data combined).

Severity category	EQ-5D Health States†								Observed utilities			
									n	mean	range	
1	11111								139	1.00	1.00	1.00
0.9	11112	11211	12111	21111	21211				61	0.85	0.81	0.88
0.8	11121	11122	11221	12112	12211	21121	21212	22211	91	0.77	0.71	0.80
0.7	11222	12121	12221	21122	21221	21222	22121	22212	187	0.67	0.62	0.69
0.6	11311	12222	22221	22222					174	0.56	0.52	0.59
0.5	11113								2	0.41	0.41	0.41
0.4	21321								2	0.36	0.36	0.36
0.3	11231	21322	22321	23221					20	0.26	0.21	0.29
0.2	21131	21223	21231	21331	22322				50	0.17	0.10	0.20
0.1	21232	21332	22131	22223	22231	22323			63	0.06	0.02	0.09
0	21233	22232	22331	22332					136	−0.03	−0.08	0.00
−0.1	21333	22233	23331						11	−0.16	−0.18	−0.11
−0.2	22333								6	−0.24	−0.24	−0.24
−0.3	23333								1	−0.35	−0.35	−0.35

† 1 = no problems, 2 = some problems, 3 = extreme problems.
1st digit = Mobility, 2nd = Self-care, 3rd = Usual Activity, 4th = Pain/Discomfort, 5th = Anxiety/Depression.

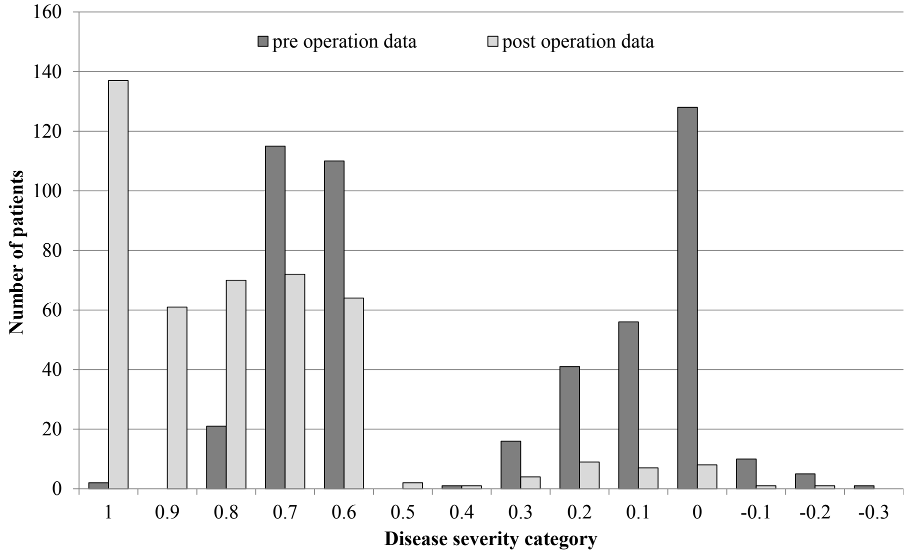


Figure 6.1: Distributions of the disease severity of patients (based on EQ-5D) pre- and post-operation.

problems' for self care and 2% for anxiety/depression. In contrast, 17% of the patients indicated extreme problems for 'usual activities' and 44% for 'pain/discomfort'.

The difference between the EQ-5D data collected before and after surgery is shown in figure 6.1. Before surgery 44% of patients reported a utility between 0.5 and 0.7, and 44% of patients had a utility between −0.1 and 0.3. Following surgery, 31% of patients were in perfect health with a utility of 1 and 61% reported a utility between 0.5 and 0.9. The mean utility gain because of hip replacement was found to be 0.42 (SD: 0.34).

Correlation between EQ-5D and OHS

For the pre-op dataset, the Spearman correlations between the OHS items and the EQ-5D items tended to be moderate (mean = 0.33, range = 0.08; 0.69). The highest correlation (0.691) was found between OHS item, *How would you describe the pain you usually had from your hip,* and the *Pain/discomfort* dimension of EQ-5D. The mean of the correlations between the EQ-5D index and the OHS items was 0.54 (range 0.41; 0.64). For the post-op data the inter-item correlations were higher (mean = 0.43, range = 0.26; 0.63) as was the correlation between the EQ-5D index and the OHS items (mean = 0.58, range = 0.49; 0.68). In both data sets the lowest correlations were those between any of the OHS items and the *Anxiety/depression* dimension of EQ-5D.

Exploratory principal component analysis

The underlying dimensional structure of the pre-op data was investigated using principal component analysis with varimax rotation. The extraction process was based on the correlation matrix. When the number of components was limited to those with an eigenvalue >1, three constructs emerged which explained 56% of the total variance: pain, self care and mobility (table 6.2).

Four out of 17 items (6, 7, 8, 11) had a loading >0.4 on more than one construct. The EQ-5D dimensions of 'mobility' and 'usual activity' loaded on the same construct. The 'anxiety/depression' dimension did not load onto any of the three constructs – this was because of the poor correlation of this dimension with any of the OHS items and other EQ-5D dimensions. OHS items 1, 6, 8, 10, 11, and 12, all related to pain, loaded together on the EQ-5D dimension 'pain/discomfort'. However, OHS item 6 loaded with an almost equal weight onto 'mobility' presumably because this item describes the time a patient is able to walk in combination with the pain from the hip becoming severe. OHS items 7 and 8 loaded onto 'pain' and 'self care'. These items describe an activity which relates to self care in combination with mention of the word painful. OHS item 11 loaded on all three constructs. Similarly OHS items 2, 3, 4, 5, 7, 8, 11 loaded onto 'self care' with items 7 and 8 also loading onto 'pain' and item 11 onto all three constructs. OHS items 6, 9 and 11 plus the EQ-5D dimensions 'mobility' and 'usual activity' loaded onto 'mobility'.

Confirmatory principal component analysis

In the confirmatory analysis the number of components to be extracted was set to five (table 6.3). As can be seen in table 6.3, no two EQ-5D items loaded onto a single factor. Furthermore, 'Anxiety/depression' can be seen to be a distinct construct, unrelated to any of the OHS items. Four OHS items (3,6,9,11) had factor loadings in excess of 0.4 on two constructs. Item 3 loaded both on 'self care' and on 'usual activity', items 6 and 9 on 'pain' and 'usual activity', and item 9 on 'mobility' and 'usual activity'. It is plausible that these items loaded on more than one construct (e.g. item 11 relates to the limitation in usual activity because of pain). Therefore, the OHS gives detailed information on three domains: pain, mobility and usual activity.

Comparison of pre-op data with other datasets

The results from both the exploratory and confirmatory analyses of the post-op data differed from those of the pre-op data. This difference was not in the number of constructs (i.e. three in the exploratory and five in the confirmatory analysis), but in the distribution

Table 6.2: Overview of the items associated with the three constructs derived from the exploratory principal factor analysis of the pre-operation data.

Item	Construct 1: Pain	Factor loading
OHS 1	How would you describe the pain you usually had from your hip?	0.76
OHS 6	For how long have you been able to walk before pain from your hip became severe? (with or without a stick)	0.46
OHS 7*	Have you been able to climb a flight of stairs?	0.42
OHS 8	After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?	0.60
OHS 10	Have you had any sudden, severe pain – 'shooting', 'stabbing' or 'spasms' – from the affected hip?	0.71
OHS 11	How much has pain from your hip interfered with your usual work (including housework)?	0.50
OHS 12	Have you been troubled by pain from your hip in bed at night?	0.72
EQ 4	Pain/Discomfort	0.72
Item	Construct 2: Self Care	
OHS 2	Have you had any trouble with washing and drying yourself (all over) because of your hip?	0.74
OHS 3	Have you had any trouble getting in and out of a car or using public transport because of your hip? (whichever you tend to use)	0.67
OHS 4	Have you been able to put on a pair of socks, stockings or tights?	0.72
OHS 5	Could you do the household shopping on your own?	0.54
OHS 7	Have you been able to climb a flight of stairs?	0.50
OHS 8*	After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?	0.48
OHS 11*	How much has pain from your hip interfered with your usual work (including housework)?	0.46
EQ 2	Self Care	0.77
Item	Construct 3: Mobility	
OHS 6*	For how long have you been able to walk before pain from your hip became severe? (with or without a stick)	0.42
OHS 9	Have you been limping when walking, because of your hip?	0.57
OHS 11*	How much has pain from your hip interfered with your usual work (including housework)?	0.47
EQ 1	Mobility	0.78
EQ 3	Usual Activity	0.63
Item	No construct	
EQ 5	Anxiety/Depression	≤ .340 on all factors

* Item is associated more strongly with one of the other constructs.

Table 6.3: Overview of the items associated with the five constructs used in the EQ-5D, based on the confirmatory principal factor analysis of the pre-operation data.

	Construct 1: Pain/Discomfort	Factor loading
OHS 1	How would you describe the pain you usually had from your hip?	0.77
OHS 6*	For how long have you been able to walk before pain from your hip became severe? (with or without a stick)	0.41
OHS 8	After a meal (sat at a table), how painful has it been for you to stand up from a chair because of your hip?	0.55
OHS 10	Have you had any sudden, severe pain – 'shooting', 'stabbing' or 'spasms' – from the affected hip?	0.67
OHS 11*	How much has pain from your hip interfered with your usual work (including housework)?	0.41
OHS 12	Have you been troubled by pain from your hip in bed at night?	0.71
EQ 4	Pain/Discomfort	0.72
	Construct 2: Usual activity	
OHS 3*	Have you had any trouble getting in and out of a car or using public transport because of your hip? (whichever you tend to use)	0.49
OHS 5	Could you do the household shopping on your own?	0.72
OHS 6	For how long have you been able to walk before pain from your hip became severe? (with or without a stick)	0.51
OHS 7	Have you been able to climb a flight of stairs?	0.69
OHS 9*	Have you been limping when walking, because of your hip?	0.40
OHS 11	How much has pain from your hip interfered with your usual work (including housework)?	0.62
EQ 3	Usual Activity	0.65
	Construct 3: Self Care	
OHS 2	Have you had any trouble with washing and drying yourself (all over) because of your hip?	0.76
OHS 3	Have you had any trouble getting in and out of a car or using public transport because of your hip? (whichever you tend to use)	0.56
OHS 4	Have you been able to put on a pair of socks, stockings or tights?	0.67
EQ 2	Self Care	0.77
	Construct 4: Mobility	
OHS 9	Have you been limping when walking, because of your hip?	0.45
EQ 1	Mobility	0.88
	Construct 5: Anxiety/Depression	
EQ 5	Anxiety/Depression	0.91

* Item is associated more strongly with one of the other constructs.

of items over the different constructs. The most noticeable difference from the exploratory analysis was that four of the five EQ-5D dimensions loaded strongest onto a single construct ('self care' being the exception). Furthermore, the confirmatory analysis on the pre-op data showed that the EQ-5D domains all loaded onto their own construct. This is not the case in the results from the post-op data. Here three of the five EQ-5D domains loaded onto a single construct on which none of the OHS items loaded. This would suggest that the constructs underlying the post-op data are not the same as those underlying the pre-op data.

The combined pre and post-op data resulted in the highest inter-item correlations. On average, the correlations of OHS items and EQ-5D items were 0.23 higher for the combined data than for the pre-op data. The factor analysis on the combined data resulted in a different distribution of the items over the factors than the analysis on the pre-op data or on the post-op data (results not presented).

Analysis based on the individual patient differences between pre- and post-op data (i.e. the changes in health of a patient across the two points in time) resulted in the lowest inter-item correlations. On average, the correlations were 0.05 lower compared to the pre-op data.

Mapping models

The PCA showed that 'Anxiety/depression' was a distinct construct, unrelated to any of the OHS items. Therefore, we could only use the direct mapping approach. In the three main effects mapping models (table 6.4) the observed differences between the pre-op data, post-op data and combined data were reflected in the different OHS items included in each model and the parameter estimates. The performance of the models varied both between dataset and disease severity category (figures 6.2 and 6.3). Predictably, the model estimated on the pre-op data (pre-op model) gave the best fit on the pre-op data with an overall MAE of 0.16, followed by the model based on the combined data (combined model, overall MAE = 0.18) and the model based on the post-op data (post-op model, overall MAE = 0.20). However, the combined model fitted the post-op data best with an overall MAE of 0.10. The post-op model resulted in an overall MAE of 0.11 and the pre-op model in an overall MAE of 0.34. In the pre-op data there were no observations with disease severity category 0.5 and 0.9. In the post-op data these were present and the MAE for category 0.5 was higher than for other categories for all three models.

Of the 90 parameters of the full mapping model only 10 were significantly different from 0 at the 95% confidence level (table 6.4). The full model performed marginally better than the combined model on the combined data with overall MAE = 0.139 compared to 0.147 for the combined model and adjusted R^2 of 0.748 versus 0.714 (figure 6.4). We found that

Table 6.4: Parameter estimates of the three main effects and the full mapping models.

OHS item	pre-op model	post-op model	combined model	OHS item	Full model
1	−.1498	*	−.0266	4	−0.020
2	−.0424	*	−.0333	12	0.049
3	*	*	*	squared terms	
4	*	*	−.0169	7 sq	−0.018
5	*	−.0358	−.0281	11 sq	−0.017
6	−.0305	−.0279	−.0347	interactions	
7	−.0369	−.0433	−.0287	1 x 7	0.023
8	−.0393	*	*	1 x 8	−0.006
9	*	−.0255	*	1 x 12	−0.036
10	*	−.0311	−.0303	2 x 5	−0.011
11	−.0522	−.0681	−.0386	6 x 10	−0.007
12	−.0372	*	−.0301	11 x 12	0.020
Intercept	1.8081	1.1815	1.2602	intercept	0.956

* Parameter estimates not significantly different from 0 at the 95% confidence level.

the mapping models underestimated the utilities for the mild health states while they overestimated those for the more severe health states. Furthermore the predicted utility gains from the full model differed from the observed utility gains obtained from the EQ-5D

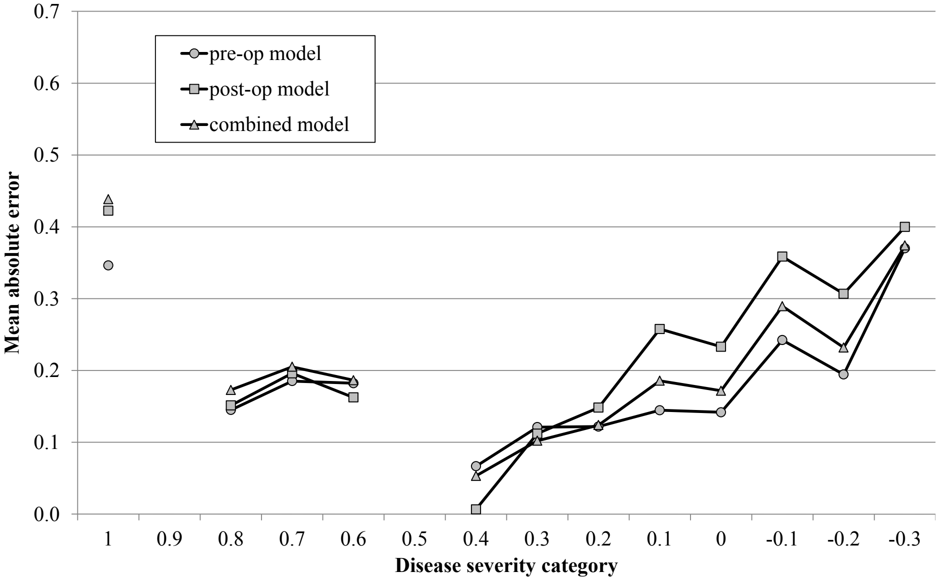


Figure 6.2: Predictive performance of the three mapping models on the pre-operation data.

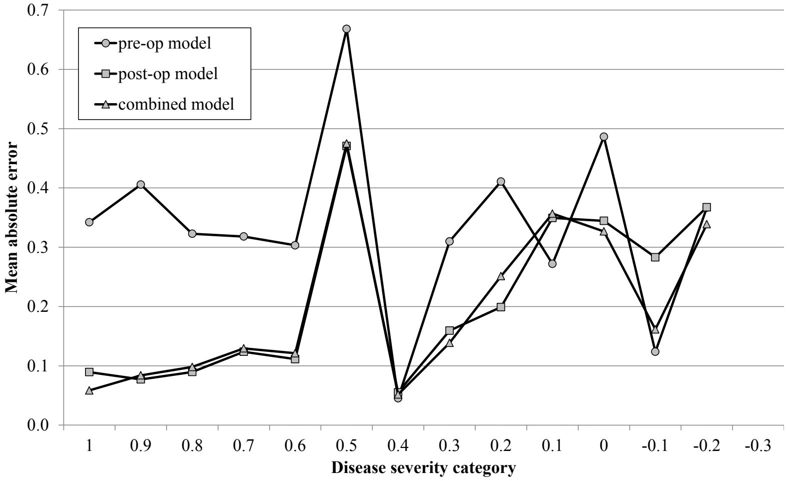


Figure 6.3: Predictive performance of the three mapping models on the post-operation data.

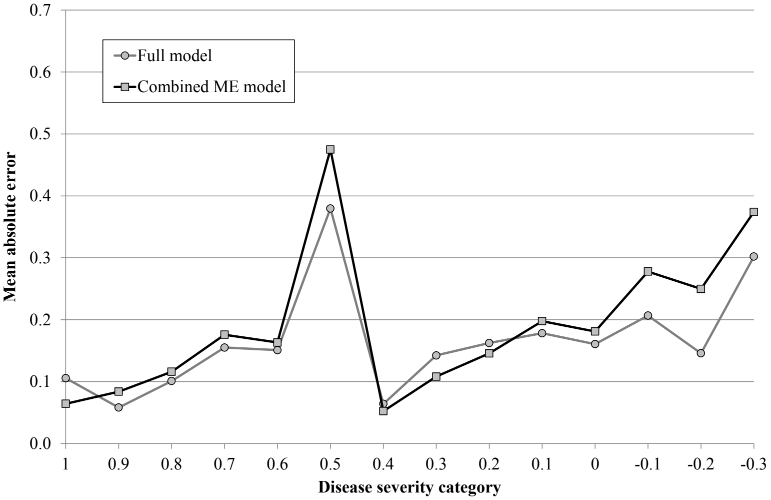


Figure 6.4: Predictive performance of the combined main effects and the full mapping models on the combined data.

data itself (figure 6.5). The number of patients reporting utility gains between 0.3 and 0.7 was higher for the predictions from the full model than for the observed data, while the predicted numbers of patients above or below these boundaries were lower than observed.

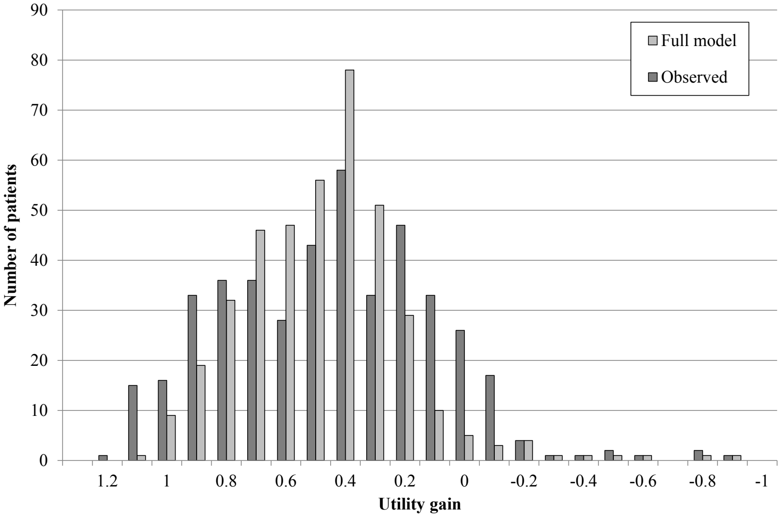


Figure 6.5: Distribution of utility gains due to hip replacement surgery: Observed gains from EQ-5D vs. predicted gains from the full mapping model.

Conclusions and discussion

Our analysis of patients' self-reported health on the OHS and EQ-5D in a sample of NHS patients undergoing hip surgery shows clear differences in the underlying constructs of these two instruments. The exploratory principal components analysis suggests that at least three distinct constructs, relating to pain, mobility and usual activity, can be used to summarize the data. The confirmatory principal components analysis shows that the 12 items of the OHS relate to four of the five dimensions of EQ-5D, the exception being 'anxiety/depression'. The correlations between the OHS and EQ-5D index (0.54 for pre-op, 0.58 for post-op) were slightly lower than the correlations found in other studies. Ostendorf and colleagues found a correlation of 0.64 pre-operation [13], while Dawson and colleagues found correlations of 0.67 for pre-operation and 0.77 for post-operation data [14].

Similar to other studies [12,15,16], we found that the mapping models underestimated the utilities for the mild health states while they overestimated those for the more severe health states. The two mapping models based on the combined data resulted in the best predicted performance. This was expected to be the case since the amount of data on which these models were estimated was almost two times that of the other models. All four models, however, perform poorly in using OHS data to predict utilities of patients with severe states (utility < 0.3), as indicated by MAE values larger than 0.14 for these values. As almost half the patients before surgery fall into this category, this is a major limitation. Because of the large MAE values indicating a minimum prediction error of 10% of the entire utility scale, none of the models could be recommended as an acceptable basis for calculating utilities from the OHS responses for use in cost utility analyses. Further, finding a mapping model that results in an acceptably low MAE is unlikely because of differences in the underlying constructs of the OHS and EQ-5D, such as the absence of 'anxiety/depression' from the OHS. These same differences also preclude the use of indirect mapping approaches when mapping OHS to EQ-5D.

The OHS items are multidimensional such that the same aspect of health status is picked up by different items. Therefore the OHS items more often show responsiveness to change than the items of the EQ-5D. Moreover since the OHS items have five response categories the potential for respondents to indicate an improvement is higher. For each item of the OHS, 75-90% of respondents indicate an improvement following surgery, while 10-20% indicates no change. For the EQ-5D items, only 30-50% indicates an improvement while 40-60% indicates no change. However, after aggregating over items using the Paretian Classification of Health Change approach [18] (in the Paretian Classification of Health Change approach an improvement in health status is defined as an improvement in at least one dimension of health with no deterioration in any of the other dimensions of health), the proportion of respondents indicating an improvement on EQ-5D increased to 82% (with 5% of respondents indicating no change on the aggregate level).

A key difference between the instruments is that the EQ-5D separates out the changes in health over separate dimensions, whereas the OHS combines information on several dimensions in a number of items. Also, in the OHS a mixture of response categories is used. Some of the response categories are similar to those of EQ-5D (i.e. they range from 'no problems' to 'impossible to do') and thus describe levels of severity. Other items in the OHS have response categories based on frequency or quantity (e.g. ranging from 'never' to 'all of the time'). Such items are not present in EQ-5D; hence improvements in these attributes as measured by OHS may not be reflected by a corresponding change in EQ-5D.

The response categories used in the EQ-5D have implications for the way patients can describe their health. The response for the most extreme level of problems with mobility is 'confined to bed'. Although such a definition might be useful to reveal the state of health for some types of patient, changes in those less severely ill cannot be adequately reflected. In effect, hip patients only have the choice between *some* and *no problems* with mobility. None of the respondents report themselves as confined to bed and, as a result, only small improvements in mobility could be detected. This problem may be resolved by the new version of the EQ-5D (EQ-5D-5L [18]) which has five response categories, and which has replaced 'confined to bed' with 'extreme problems with mobility'.

Some items of the OHS relate to more than one construct. This results in inter-item correlations hampering the creation of a parsimonious value set (i.e. interaction effects will be important). The summary score for the OHS is calculated simply by adding the responses for each item. This means that the numerical labels of the response categories are assumed to have intrinsic numerical values with interval scale properties (implying that the items have equal weight and that the distances between the levels are equal). Furthermore, the different types of response categories are treated equally. From past valuation research it is clear that these assumptions do not hold [3]. Therefore the numeric values represented by the summary scores have limited meaning. For quality assessment of services, this limits the use of the summary scores of OHS to comparisons of performance over time and between providers, but does not allow assessment of the absolute impact.

Differences between the OHS and the EQ-5D do not undermine the merits of either instrument when used for their own purposes. However, our results suggest that, because of the conceptual differences between these instruments, it is not possible to produce a viable mapping model for estimating utilities for the OHS based on OLS regression.

References

1. Devlin N, Appleby J. Getting the most out of PROMs: putting health outcomes at the heart of NHS decision-making. London: King's Fund. Available from: www.kingsfund.org.uk/publications/proms.html. [Accessed September 23 2010].
2. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37:53-72.
3. Szende A, Oppe M, Devlin N, (eds.). EQ-5D valuation sets: an inventory, comparative review and user guide. Dordrecht: Springer; 2007.
4. Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br. 1998;80:63-9.
5. Brazier J, Yang Y, Tsuchiya A. Review of methods for mapping between condition specific measures onto generic measures of health. Background paper to: OHE Commission on NHS Productivity and Performance; London; 2008.
6. Mortimer D, Segal L. Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. Med Decis Making. 2008;28:66-89.
7. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. Med Decis Making. 2006;26:18-29.
8. Browne J, Jamieson L, Lawsey J, et al. Patient Reported Outcome measures (PROMs) in elective surgery. Report to the Department of Health. 2007. Available from: www.lshtm.ac.uk/hsru/research/PROMs-Report-12-Dec-07.pdf. [Accessed September 23 2010].
9. Dolan P. Modelling valuations for EuroQol health states. Medical Care. 1997;35:1095-108.
10. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.
11. Johnson RA, Wichern DW. Applied multivariate statistical analysis. (4th ed.). Upper Saddle River NJ: Prentice Hall; 1998.
12. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? Health Qual Life Outcomes. 2009;7:27.
13. Ostendorf M, van Stel HF, Buskens E, et al. Patient-reported outcome in total hip replacement. A comparison of five instruments of health status. J Bone Joint Surg Br. 2004;86:801-8.
14. Dawson J, Fitzpatrick R, Frost S, et al. Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. J Bone Joint Surg Br. 2001;83:1125-9.
15. Brazier J, Yang Y, Tsuchiya A: Review of methods for mapping between condition specific measures onto generic measures of health. Report prepared for the Office of Health Economics; 2007.
16. Franks P, Lubetkin EI, Gold MR, et al. Mapping the SF-12 to the EuroQol EQ-5D Index in a National US Sample. Medical Decision Making. 2004;24:247-54.
17. Devlin N, Parkin D, Browne J. Patient Reported Outcomes in the NHS: new methods for analysing and reporting EQ-5D data. Health Econ. 2010;19:886-905.

18. Herdman M, Sanz L, Lloyd A, Badia X, Gudex C. Qualitative testing of two new 5-level versions of the EQ-5D in Spain: preliminary study results. Proceedings of the 26th EuroQol Group plenary meeting.

Chapter 7

A geometric approach to health state modelling

Benjamin M. Craig, PhD[1], Mark Oppe, MSc[2].

1. Health Outcomes & Behavior Program,Moffitt Cancer Center, Tampa, Florida & Department of Economics, University of South Florida, Tampa, Florida, USA

2. Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

Publication:

Craig BM, Oppe M. From a different angle: a novel approach to health valuation. Soc Sci Med. 2010 Jan;70(2):169-74.

Introduction

Can you imagine a health state where you would rather die than be in that state? If not, might another? Respondents with infinitely negative values for specific health states are commonly encountered in health valuation studies. When accumulating preferences across individuals, such extreme values negate the validity of summary statistics, such as population means and variances. A single infinite value causes the statistic to become infinite itself, losing all information about the other values within the population.

Even if such an extreme value is impossible, the potential of an extreme response (i.e., stated value) hinders survey research. When asked in a survey, respondents may state that they would give anything to have a baby, drink, or good night's sleep (not necessarily in that order). Whether such an extreme response is credible or not, sample means including these infinite responses are not defined. The potential for an infinite response or infinite value is a challenging aspect to all forms of preference-based measurement, and is acutely important in health valuation.

Currently, three trade-off techniques dominate the literature, each of which involves varying quantities of life (i.e., time, risk, and persons). For example, the time trade-off technique (TTO) might ask whether the respondent prefers ten years in a disease state or eight years in optimal health. By raising and lowering the time in optimal health (a.k.a., quality-adjusted life years or QALYs), the interviewer can identify the respondent's indifference point (e.g., ten years in the health state equal to eight QALYs). A state may have an extremely negative value: a respondent may be indifferent between a minute with a disease and the loss of one QALY (one minute with disease equals negative one QALY), which implies that a year with disease is worth –525,949 QALYs. Such an extreme TTO response would overwhelm typical summary measures. Therefore, this paper introduces an application of directional statistics in health valuation studies that may replace the more common practices [1,2].

The classical approach in health valuation remains highly controversial: (1) value is expressed as a ratio, representing the trade-off between two goods (e.g. –1 year/1 minute = –525,949 QALYs); (2) the summary measure is the mean ratio; and (3) because means with outliers behave badly, extreme ratios are arbitrarily transformed to make the estimates look more credible [1,3]. We show that the assumption of angular error or "wavering preference" motivates the use of directional statistics as an alternative approach to ratio statistics, and negates the impetus toward arbitrary transformations.

All trade-offs may be expressed using Cartesian coordinates (x,y), where a person is indifferent between x and y. In time trade-off (TTO), x is time in a disease state and y is time

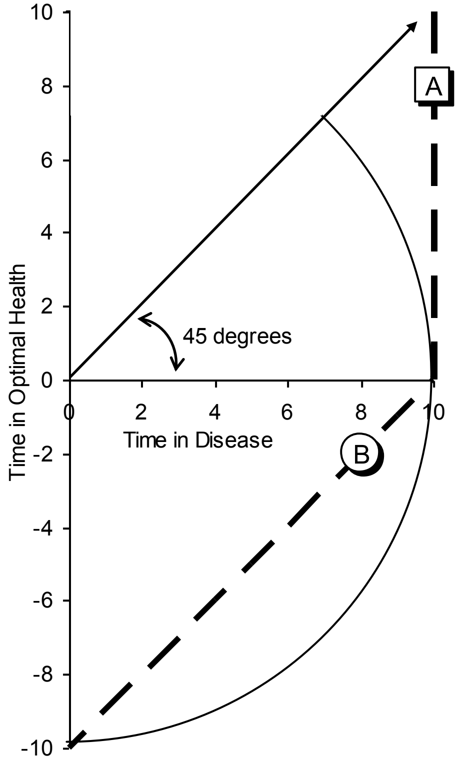


Figure 7.1: Time Trade-off Responses.

in optimal health (a.k.a. quality-adjusted life years (QALYs)). Value may be expressed by the ratio, y/x (See figure 7.1). For example, from a maximum choice of ten health years, a respondent may equate eight years in optimal health to ten years of disease time (Point A; value = 8/10 = 0.8). Or, a respondent may consider a scenario of two years of optimal health followed by eight years of disease, and equate it to "immediate death." Because the value of death is zero on a QALY scale, this response suggests the eight years of disease is equal to a loss of two years in optimal health (Point B; value = –2/8 = –0.25). All TTO responses (x,y) may be arranged on the dashed line in figure 7.1, and values on a QALY scale are bounded between one and negative infinity.

Using a sample of trade-off responses, the conventional approach to valuation is to estimate a ratio statistic, μ, by minimizing the sum of squared error:

$$\underset{\mu}{Min}\, \tfrac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2 \quad \text{where } \varepsilon_i = y_i / x_i - \mu \tag{1}$$

Values, y_i/x_i, vary across individuals. Additive variation around a ratio statistic, μ, may be expressed using an error term, ε_i, representing randomness in value and measurement error. Typically, additive error distributions have finite variance in addition to the expected value zero and independence. However, error in a ratio statistic can be poorly defined, because if one or more x values are zero, the error becomes infinitely large. In the UK valuation of EQ-5D states, Dolan addressed the infinite variance problem by arbitrarily replacing all negative ratios (i.e., worse-than-death or WTD ratios) with y/10 [1]. Because y in the TTO varies from negative 9.75 to 10 years, the range of the adjusted values is bounded, from −0.975 to one.

Another concern with the application of ratio statistics is that x and y are not interchangeable. In other words, μ(x,y) is not equal to the inverse of μ(x,y). This is particularly evident if one or more y values are zero. However, in a more general conjoint analysis, the trade-off of x for y may or may not be equal to the inverse of the trade-off of y for x, particularly in the case of complementary goods (e.g., shoe strings and shoes); however, this is an advantageous attribute in health valuation and other applications, like monetary exchanges (e.g., dollars for yen).

Drummond and colleague discuss similar difficulties in the estimation of incremental cost-effectiveness ratios [4,5]. On a cost-effectiveness plane, the y-axis reflects incremental costs (y) and the x-axis reflects incremental effectiveness (x). The convention is to divide the mean cost by the mean effectiveness (a ratio known as an incremental cost-effectiveness ratio or ICER) as an alternative to a ratio statistic.

The application of directional statistics in health valuation addresses the problems of extreme values and interchangability, and motivates an estimator nearly identical to the ICER (i.e., ratio of means).

Methods

Directional Statistics in Health Valuation

Every point in a Euclidean space can be uniquely mapped to a set of polar coordinates (θ, r) described by an angle and a radius:

$$x = r * \cos(\theta) \quad y = r * \sin(\theta) \quad r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \arctan(y/x)$$

Specifically, each ratio (y/x) is the tangent of an angle, θ. The radius, r, represents the size of the trade-off. Instead of a ratio statistic as the value estimator, we propose estimating the tangent of the mean angle.

In figure 7.1, we show that the QALY angles are bounded between 45 degrees and negative 90 degrees. For example, a non-trader's response of a negative infinite QALY value is a negative 90 degree QALY angle. Similarly, the value of optimal health (ratio = 1) is a 45 degree angle and the value of dead (0) is a zero degree angle.

Instead of expressing randomness in value, $y_i/x_i = \mu + \varepsilon_i$, we may express randomness in direction, $\theta_i = \theta + \varepsilon_i$. Direction error has been examined in many settings, such as adjustments on a dial, readings on a compass or clock, or the variability of wind directions or seasonality [6-8]. As in the aforementioned examples, respondent preferences in health valuation may waver in a directional fashion (e.g., feeling up beat or downtrodden).

Our solution of changing the coordinate system so that problems can be circumvented (or calculations be made more easily) is commonly used in physics. For example, obtaining the equation of motion of a system of coupled oscillators can be done easier using Langrangian mechanics with spherical coordinates than using Newtonian mechanics with Cartesian coordinates [9].

Because angles are bounded, directional statistics are finite by construction and interchangeable. However, two well-known issues prevent the use of ordinary least squares (OLS) (equation 1) as a directional loss function for the estimation of mean angles: the crossover problem and circular variance. The crossover problem is related to the circular nature of angles. For example, on a compass, where north is zero degrees, the arithmetic mean of 45 degrees (northeast) and 315 degrees (northwest) is 180 degrees (south), not 0 degrees (north), even though zero may be a more accurate representation of central tendency. The potential of crossing over north prevents the use of arithmetic means in directional applications. The QALY angles lie between 45 degrees (the values of optimal health) and negative 90 degrees (value of negative infinity), not throughout the entire circle. Therefore, crossover (i.e., angles beyond 180 or negative 180 degrees) is not possible.

Because the sum of squared error does not represent circular variance, OLS (equation 1) is inappropriate to use as a directional loss function. The largest possible error in QALY angles is 145 degrees; yet, the OLS specification allows for error beyond 145 degrees, and the square of this error may reach beyond the crossover point (180 degrees). OLS is inappropriate for the estimation of a linear probability model for similar reasons.

In directional statistics, circular variance is represented by

$$1 - \sin\hat{\theta}\,\sin\theta - \cos\hat{\theta}\,\cos\theta \rightarrow \underset{\hat{\theta}}{Min}\tfrac{1}{N}\sum_{i=1}^{N} 1 - \sin(\hat{\theta})\sin(\theta_i) - \cos(\hat{\theta})\cos(\theta_i) \qquad (2)$$

Mean angle, $\hat{\theta}$, is the estimate that minimizes circular variance, which is a directional loss function analogous to OLS (equation 1). Mardia and Jupp refer to this measure of dispersion as one minus the mean resultant length, $\bar{R}$ [10]. Unlike the error in ratio statistics, each element in the circular variance expression is finite, ranging from zero to two, with an overall mean ranging between zero and one. If the angles are widely dispersed (i.e., discordance in health state value), circular variance approaches one, $\bar{R}\cong 0$, and if the angles are concentrated (i.e., concordance in health state value), circular variance approaches zero, $\bar{R}\cong 1$.

To clarify the estimator of the tangent, we take the derivative of equation 2 and set it to zero:

$$\tfrac{1}{N}\sum_{i=1}^{N} -\cos\left(\hat{\theta}\right)\sin\left(\theta_j\right) + \sin\left(\hat{\theta}\right)\cos\left(\theta_j\right) = 0$$

$$\cos\left(\hat{\theta}\right)\left(\tfrac{1}{N}\sum_{i=1}^{N} -\sin\left(\theta_j\right) + \tan\left(\hat{\theta}\right)\tfrac{1}{N}\sum_{i=1}^{N}\cos\left(\theta_j\right)\right) = 0$$

$$\tan\left(\hat{\theta}\right) = \tfrac{1}{N}\sum_{i=1}^{N}\sin\left(\theta_j\right) \Big/ \tfrac{1}{N}\sum_{i=1}^{N}\cos\left(\theta_j\right)$$

$$\tan\left(\hat{\theta}\right) = \tfrac{1}{N}\sum_{i=1}^{N}\frac{y_i}{r_i} \Big/ \tfrac{1}{N}\sum_{j=1}^{N}\frac{x_j}{r_j}$$

The tangent of the mean angle is the mean of y/r over the mean of x/r, where r is the radius. In figure 7.1, each TTO response has a radius as measured from the distance to the origin. If all responses (x, y) were rescaled by dividing by their radii, they would lie along the semicircular line. The tangent of the mean angle would be the mean of the rescaled y over the mean of the rescaled x. In other words, the mean angle estimator ignores the distance from the origin of each response.

Instead, radii may be included in the loss function as weights for each element of equation two. It follows that the tangent of a radially weighted mean angle is the mean of y over the mean of x:

$$\underset{\hat{\theta}_r}{Min}\,\tfrac{1}{N}\sum_{i=1}^{N} r_i\left(1-\sin\left(\theta_i\right)\sin\left(\hat{\theta}_r\right)-\cos\left(\theta_i\right)\cos\left(\hat{\theta}_r\right)\right) \rightarrow \tan\left(\hat{\theta}_r\right) = \tfrac{1}{N}\sum_{i=1}^{N} y_i \Big/ \tfrac{1}{N}\sum_{j=1}^{N} x_j \qquad (3)$$

Radial weighting the loss function suggests that angular error far from the origin is more important than error near the origin. In valuation, trade-offs with lengthy radii may be given more weight, because greater quantities are involved. For example, a monetary exchange involving millions of Euros may receive greater attention than a typical money exchange at an automatic teller machine. On the other hand, in trade-off response, which represents a single respondent's valuation of a single state (i.e., one person, one vote), variability in the radii (see figure 7.1) is viewed as an artefact of the experimental design, and the mean angle removes this arbitrary noise. The point (9, 10) is farther from the origin than the point (8, 10), but this does not necessarily suggest that it is more or less important.

In health valuation, directional statistics are appealing for their technical simplicity and plain intuition (i.e., individual preferences waver). Instead of the ratio statistic (i.e., the mean of y/x), the approach entails a ratio of means, $\bar{y}/\bar{x}$. Radially weighted or not, the mean of x is non-zero by construction; therefore, the directional statistics may be more robust than their ratio counterparts. If x and y were switched, the resulting estimate would be the inverse of the original (i.e., interchangeability). When Dolan replaced WTD responses with y/10, the adjusted ratio statistic became $\bar{y}/10$, which is similar to the radially weighted estimator, $\bar{y}/\bar{x}$ [1]. Although Dolan's transformation has no theoretical basis, estimates under the classical approach approximate those based on directional statistics by construction.

Circular Regression

Valuation studies typically examine trade-offs between hypothesized health scenarios to predict the values of scenarios that were not directly incorporated into the sample. Out-of-sample predictions can be accomplished using a linear combination of state-specific variables, Z'β, known as a multi-attribute utility (MAU) regression model. Using OLS (equation 1), the classical approach is to estimate the MAU regression model, $y_i/x_i = Z_i'\beta + \varepsilon_i$, where the dependent variable is the ratio, y/x. To improve the face validity of these predictions, Dolan arbitrarily replaced the dependent variable with y/10.

The circular regression approach is to estimate a linear MAU model by minimizing circular variance (equation 2), where $\theta_i = \arctan(y_i/x_i)$ and $\hat{\theta} = \arctan(Z_i'\hat{\beta})$. Similarly, the radially weighted directional loss function (equation 3) may be minimized to estimate $\hat{\theta}_r = \arctan(Z_i'\hat{\beta}_r)$. The MAU regression coefficients, β and β_r, are on the same scale as the ratio statistic estimates; yet, the circular regression approach avoids the problems of ratio statistics and the arbitrary transformations of WTD responses.

The Measurement and Valuation of Health Study in the United Kingdom

To demonstrate the application of directional statistics in health valuation, we examine data from the seminal Measurement and Valuation of Health Study [1,11]. In 1993, the University of York administered 3395 interviews with a response rate of 64% and collected values of 42 EQ-5D health states and the state of unconsciousness. During the TTO exercise, respondents placed a value on up to 13 states. As mentioned earlier, the MVH protocol bounded the lower end of the loss in years to be greater than negative 9.75 (See figure 7.1); therefore, the ratio, y/x, is bounded between 1 and −39 (or −9.75/0.25).

For the TTO analytical sample (N = 3,355), respondents were excluded (1) if only one or two states were valued (other than 11111, "immediate death," and "unconscious"); (2) if all states were given the same value; or (3) if all states were valued worse than "immediate death." The three criteria motivated the exclusion of 1.2% of the TTO respondents. Across the 3,355 respondents, each of the 39,673 TTO responses described an equivalence of time in optimal and non-optimal health, BTD (10,y) or WTD (10+y, y), where y is time in optimal health between 10 and negative 9.75 years.

In this analysis, the values of the 42 hypothesized EQ-5D states were estimated using ratio statistics with and without Dolan's transformation of WTD responses, and using directional statistics with and without radial weights. This allowed the comparison of the four methods (i.e., mean ratio, Dolan adjusted, Unweighted, and Radially-Weighted) without the distraction of state-specific attributes. Likewise, four MAU regression models were estimated to predict the values of the 243 EQ-5D states.

For both the 42 state values and the regression coefficient, 95% confidence intervals were estimated using the percentile method by applying bootstrap sampling with respondent-specific cluster replacement. For each iteration of the bootstrap, a sample of respondents was extracted from the analytical sample with replacement and the analysis was re-run with the bootstrap sample. After 1,000 iterations, the parameter estimates were sorted and the top and bottom 24 estimates of each parameter removed. The 25th and 975th estimates represented the 95% confidence interval under the percentile bootstrap approach [12].

In complement to visual inspection, concordance between the predictions made in this study was measured using Lin's coefficient of agreement and mean absolute difference. Because of its prominence in the literature, Dolan's published value set was compared to these regression predictions. Using the same source data and variables as the original analysis of the MVH data, ratio statistic estimates with Dolan's transformation of WTD responses were nearly identical in this analysis to published estimates. Minor deviations between the published and the re-estimated values may be attributable to differences in sample selection criteria.

Like the original analysis of the MVH data, the MUA regression model includes twelve indicator variables: five for second level domains, five for the third level domains, one for any second or third level domains (i.e., constant); and one for any third level domains (i.e. N3). The EQ-5D descriptive system has five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) each with three possible levels [13]. The MUA regression model captures the detrimental effects of each level on each domain as well as the multiplicative effects of any one second or third level domains (i.e., constant), and of any one third level domains in the state vector (i.e., N3). All database work was conducted on SAS 9.1, and the analyses were conducted using Stata 10 [14,15].

Results

After stratifying the analytic sample into 42 state-specific subsamples, we estimated the mean ratio with and without Dolan's transformation of WTD responses (i.e., $\frac{1}{N}\sum y_i / x_i$ and $\bar{y} / 10$) and the tangent of the mean angle with and without radial weights (i.e., $\bar{y}_r / \bar{x}_r$ and $\bar{y} / \bar{x}$). Without the Dolan transformation, the mean ratio is significantly positive for only ten out of the 42 states, which suggests that few states are better than "immediate death". These estimates clearly lack face validity, and motivate the Dolan transformation. Transformed estimates are significantly greater than or equal to mean ratios for all 42 states, arbitrarily increasing values. Of the 42 states, 28 of the transformed estimates are significantly positive and 14 significantly negative, suggesting a third of the states are worse than "immediate death". Because the Dolan transformation is the conventional approach to health valuation, these estimates (i.e., Dolan ratios) are compared with directional results.

Based on figure 7.2, directional statistics produce values similar to the Dolan ratios. As a measure of concordance with the arbitrarily adjusted estimates, Lin's coefficient of agreement is 0.916 for the tangent of the mean angles and 0.954 for the radially weighted

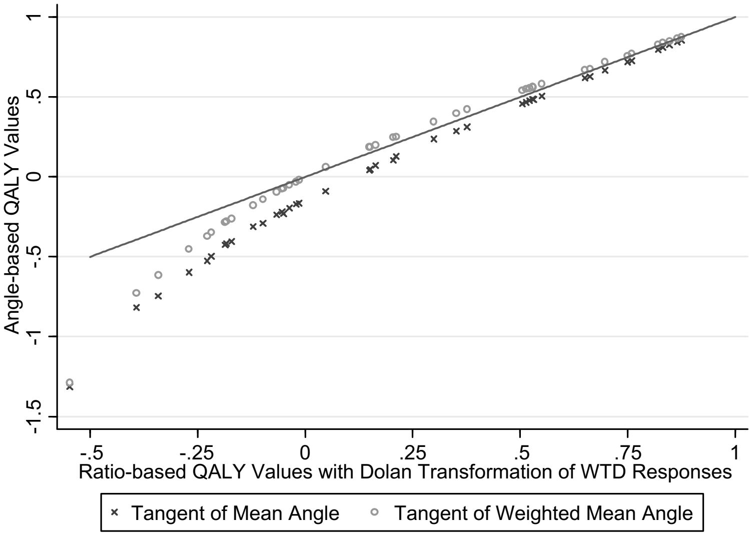


Figure 7.2: Comparison of Angle-based and Ratio-based QALY values for the 42 Hypothesized EQ-5D Health States.

estimates. Similarly, the absolute mean difference is 0.136 for the unweighted and 0.067 for the weighted.

While the estimates are similar, the unweighted estimates are significantly less than the Dolan ratios for all 42 states (figure 7.2). The weighted estimates are more balanced; significantly less than the Dolan ratios for 26 states and greater than for 14 states. If the purpose is to produce estimates similar to the Dolan ratio predictions, the tangent of the radially weighed angle is a preferred estimator.

Figure 7.2 further illustrates the negative relationship between the Dolan ratios and the angle-based QALY values. For BTD states, the differences between the Dolan ratios and the directional estimates appear small. For WTD states, the difference increases as states grow more severe. For example, the Dolan ratio of the "pits" state (33333) is X, and the weighted estimate is Y.

Health Valuation of the Entire EQ-5D Descriptive System

To assign values to all 243 possible EQ-5D health states, we estimated four regression models (table 7.1). Each coefficient reflects a decrement from optimal health (1.00); therefore, based on the 95% bootstrap confidence intervals, it is expected to be significantly negative.

Table 7.1: Multi-Attribute Utility Regression Models for EQ-5D Health States.

N = 3,355 respondents	Mean Ratio*						Tangent of Mean Angle					
with 39,673 responses	Untransformed			Transformed			Unweighted			Radially Weighted		
State Attributes	Coef.	95% C.I.		Coef.	95% C.I.		Coef.	95% C.I.		Coef.	95% C.I.	
Mobility, 2	−0.234	−0.378	−0.092	−0.069	−0.081	−0.058	−0.055	−0.073	−0.038	−0.040	−0.056	−0.025
Self-Care, 2	−0.085	−0.227	0.051	−0.105	−0.116	−0.094	−0.098	−0.115	−0.081	−0.082	−0.097	−0.068
Usual Activity, 2	−0.246	−0.372	−0.114	−0.034	−0.047	−0.022	−0.048	−0.065	−0.031	−0.050	−0.065	−0.037
Pain/Discomfort, 2	−0.240	−0.400	−0.095	−0.120	−0.131	−0.108	−0.124	−0.142	−0.105	−0.106	−0.122	−0.090
Anxiety/Depression, 2	−0.213	−0.346	−0.081	−0.071	−0.082	−0.061	−0.088	−0.104	−0.072	−0.087	−0.101	−0.073
Mobility, 3	−2.627	−2.900	−2.365	−0.311	−0.328	−0.293	−0.499	−0.529	−0.469	−0.498	−0.527	−0.469
Self-Care, 3	−1.614	−1.837	−1.405	−0.217	−0.232	−0.201	−0.327	−0.350	−0.305	−0.323	−0.344	−0.301
Usual Activity, 3	−1.202	−1.450	−0.977	−0.084	−0.102	−0.068	−0.183	−0.210	−0.160	−0.204	−0.228	−0.181
Pain/Discomfort, 3	−2.710	−2.962	−2.468	−0.374	−0.390	−0.358	−0.590	−0.618	−0.563	−0.575	−0.602	−0.549
Anxiety/Depression, 3	−2.050	−2.284	−1.800	−0.234	−0.250	−0.219	−0.386	−0.413	−0.359	−0.395	−0.420	−0.370
Any 2's or 3's	−0.073	−0.163	0.029	−0.086	−0.096	−0.075	−0.114	−0.129	−0.100	−0.090	−0.103	−0.078
Any 3's	0.834	0.583	1.075	−0.279	−0.297	−0.260	−0.125	−0.150	−0.098	−0.028	−0.051	−0.005

* In the TTO, better than death (BTD) response is (10,y) and worse than death (WTD) response is (10+y, y) where y is years in optimal health equal to x years in the health state. Dolan transformed the WTD responses from (10+y,y) to (10,y), arbitrarily inflating the ratios, y/x.

For example, the Dolan ratio coefficient of any one second or third level domain is –0.086, which suggests that non-optimal health states have a maximum value of 0.914 (or 1-0.086). This decrement is known as the non-optimal gap.

The first five coefficients represent the decrement associated with "some problems" on each of the five domains. For these coefficients, the 95% confidence intervals overlap. The second set of coefficients represents decrements associated with "severe problems." With the exception of the N3 coefficient, the Dolan ratio coefficients are significantly lower than the directional coefficients. The value of the N3 coefficient, being larger in the Dolan ratio model (–0.279) than in the directional models (–0.125 and –0.028), suggests that the directional models better differentiate the third level domains than the Dolan ratio model.

The replication of the Dolan ratio model is nearly identical to the published estimates (figure 7.3). Lin's coefficient is 0.999, and the mean absolute difference across the 243 predicted values is 0.006. The small difference is likely attributable to rounding error and changes in the sample selection criteria. Figure 7.3 illustrates the relationship between the published Dolan estimates and the angle-based predictions, including predictions for all 243 EQ-5D states [1]. A negative relationship is illustrated where the greatest difference appears in the prediction of WTD values. For the 243 predictions, Lin's coefficient of agreement between Dolan's values and the unweighted values is 0.85, and the mean absolute difference is 0.164. Greater agreement is found in the weighted estimates, where Lin's coefficient is 0.922 and

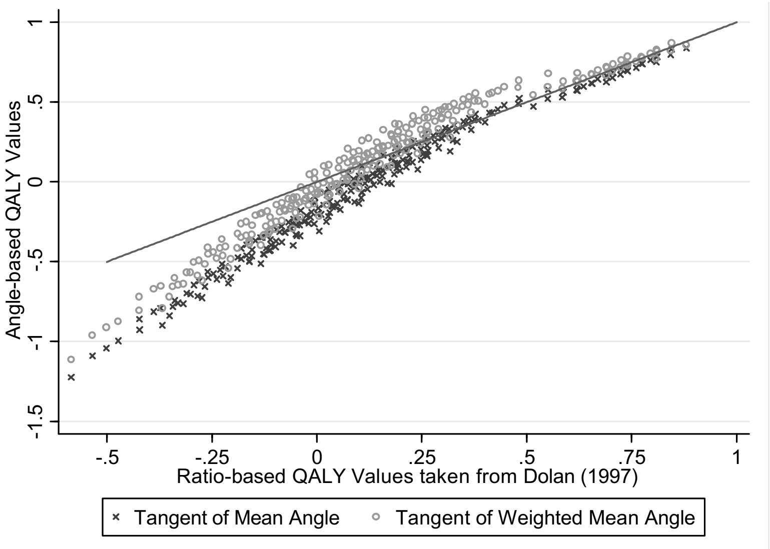


Figure 7.3: Comparison of Angle-based and Ratio-based QALY values for All 243 possible EQ-5D Health States.

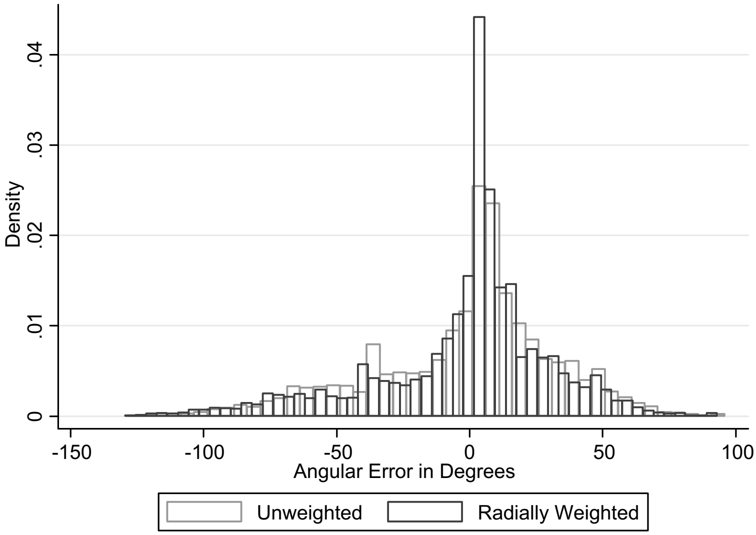


Figure 7.4: Histogram of angular error for circular regression models.

the mean absolute difference is 0.109. For reference, figure 7.4 is a histogram of angular error for the circular regression models with or without radial weights.

Conclusions and discussion

In this paper, we introduce the concept of wavering preferences, and two directional statistics for use in the valuation of health states (i.e., $\bar{y}_r/\bar{x}_r$ and $\bar{y}/\bar{x}$). Each estimator addresses well known issues in the ratio statistics, specifically infinite values and interchangeability, and negates the impetus behind the transformation of outlying responses (e.g., Dolan or Shaw's transformation of WTD responses). The estimator is nearly identical to an incremental cost-effectiveness ratio (ICER). The resulting predictions are similar to those commonly applied in health policy; however, differences occur in the more severe health states. We focus on health valuation; however, multiple areas of conjoint analysis in health and medicine may benefit from this approach, particularly those where trade-offs seem unfathomable.

Given that the QALY scale is bounded between one and negative infinity, QALY angles are bounded between 45 degrees and negative 90 degrees. The directional estimator does not impose these bounds, which leads to two possible limitations. First, the predicted angles may be outside the QALY scale, which is similar to the problems faced in linear probability models. The ratio of means, $\bar{y}/\bar{x}$, are naturally bounded to the interval, but out-of-sample predictions may be off the scale. This is unlikely to occur at the upper bound where 45 degrees is optimal health, because all descriptive systems describe decrements from this point. WTD values may seem more likely to extend past negative 90 degrees, but few states are WTD. A second limitation is that the confidence intervals may span outside the QALY scale. To address this limitation, we apply bootstrap techniques to estimate the confidence intervals around estimates instead of assuming symmetric standard error. Because bootstrap intervals are empirical and rely on resampled predictions, the confidence intervals remain within the QALY scale.

Alternative statistical methods in health valuation have been proposed to analytically accommodate infinite ratios, all of which have underlying assumptions with arbitrary elements. The most common is the transformation of WTD responses. Lamers and colleagues investigated three such transformations: 1) the monotonic transformation, y/10, proposed by Patrick and used by Dolan; 2) the linear transformation, (y/x)/39, proposed by Shaw and colleagues; and 3) truncation at –1 [1-3,16]. Lamers shows that each renders a different value set, and all transformations lack a sound theoretical underpinning.

A second class of alternative methods involves changing the estimator, not the data. In the current paper, we recommend the use of direction statistics; however, Craig and Busschbach recommend regressing y on x, using a coefficient, instead of a ratio statistic, as the estimator [17]. More recent work has investigated changing the measure of central tendency: instead of mean ratio, median or mode ratios may be estimated [18]. Median and mode statistics mitigate the effects of potentially infinite distribution tails, but are less relevant for economic evaluations.

Choosing directional statistics over Craig and Busschbach's coefficient approach may appear arbitrary; nevertheless, each has a clear utility framework (i.e., wavering and episodic utility) [17]. No theoretical framework has yet been proposed to motivate the manipulation of data or the use of medians or modes for decision analyses, so these more pragmatic alternatives may be less justified.

Changing the estimator does not resolve issues inherent to trade-off experimental protocols. TTO responses are collected on two scales, one for BTD responses and another for WTD responses [11]. Scale separation may psychometrically influence TTO responses, which is not addressed by the proposed directional statistics. Secondly, we examine the value of a health state by varying time as a quantity of life, not risk or persons. Even though the problem that we present in this paper is essentially two dimensional, this does not mean that the use of directional statistics is limited to two dimensional problems. In principle the methodology can be extended to include three or more dimensional problems, just like in physics (e.g. in relativistic mechanics a four dimensional coordinate system is typically used) [19-21]. Lastly, the TTO task involves only gains in time. Prospect theory suggests that respondents value losses distinctly from gains, and adjusting for these differences would be analogous to adding a reverse gear to the directional approach [22,23].

The tangent of the radially weighted mean angle, $\bar{y}/\bar{x}$, provides consistent estimates without the arbitrary transformation of WTD responses, and the estimator has a clear underlying theoretical framework (i.e., wavering preferences). Its predictions are nearly identical to Dolan's estimates, except that they have a wider range. To understand this difference, it is noteworthy that the two estimators, $\bar{y}/10$ and $\bar{y}/\bar{x}$, are the same except for the difference between $\bar{x}$ and ten. Because time in disease (x) is ten years or less, no simulation is required to show that the proportional difference between the estimates is always $\bar{x}/10$ by construction. The more difficult questions concern the implications of this wider range in QALY estimates for economic evaluations.

References

1. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-1108.
2. Shaw J, Johnson J, Coons S. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005;43(3):203-220.
3. Lamers L. The transformation of utilities for health states worse than death: consequences for the estimation of EQ-5D value sets. Med Care. 2007;45(3):238-244.
4. Drummond MF, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programmes. Oxford University Press, 3rd edition; 2005.
5. Stinnett A, Paltiel A. Estimating CE ratios under second-order uncertainty: the mean ratio versus the ratio of means. Med Decis Making. 1997:17(4):483-489.
6. Gao F, Seah S, Foster P, Chia K, Machin D. Angular regression and the detection of the seasonal onset of disease. J Cancer Epidemiol Prev. 2002:7(1);29-35.
7. Gao F, Nordin P, Krantz I, Chia K, Machin D. Variation in the seasonal diagnosis of acute lymphoblastic leukemia: evidence from Singapore, the United States, and Sweden. Am J Epidemiol. 2005:162(8);753-763.
8. Gao F, Chia K, Krantz I, Nordin P, Machin D. On the application of the von Mises distribution and angular regression methods to investigate the seasonality of disease onset. Stat Med. 2006:25(9);1593-1618.
9. Landau L, Lifshitz E. Mechanics. Oxford, New York: Pergamon Press; 1989.
10. Mardia K, Jupp P. Directional statistics. Chichester, New York: J. Wiley; 2000.
11. Gudex C. Time Trade-Off User Manual: Props and Self-Completion Methods. Report of the Centre for Health Economics. York, United Kingdom: University of York; 1994.
12. Efron R, Ribshirani B. An Introduction to the Bootstrap: Chapman and Hall; 1993.
13. Szende A, Oppe M, Devlin N, (eds.). EQ-5D valuation sets: an inventory, comparative review and user guide. Dordrecht: Springer; 2007.
14. SAS. SAS 9.1 for Windows. Cary, NC, USA: SAS Institute Inc; 2007.
15. StataCorp. Stata Statistical Software: Release 10. College Station, Texas, USA: StataCorp LP; 2007.
16. Patrick D, Starks H, Cain K, Uhlmann R, Pearlman R. Measuring preferences for health states worse than death. Med Decis Making. 1994;14(1):9-18.
17. Craig B, Busschbach J. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Popul Health Metr. 2009;7(1);3.
18. Shaw J, Shengsheng Y, Chen S, Iannacchione V, Johnson J, Coons S. There's nothing "Average" about these weights: Development and testing of a median model of US EQ-5D Health State Preferences. 14th Annual Meeting of the International Society for Quality of Life Research. Toronto, ON; 2007.
19. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002;11(5):447-456.

20. Craig B. The duration effect: a link between TTO and VAS values. Health Econ. 2009;18(2): 217-225.

21. Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. Med Care. 2009; 47(6):634-41.

22. Oliver A. The internal consistency of the standard gamble: tests after adjusting for prospect theory. J Health Econ. 2003;22(4):659-674.

23. van Osch S, Stiggelbout A. The construction of standard gamble utilities. Health Econ. 2008; 17(1):31-40.

Chapter 8

Discrete choice modelling for the quantification of health states

Elly Stolk, PhD[1], Mark Oppe, MSc[1], Luciana Scalone, PhD[2], Paul Krabbe, PhD[3].

1. Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University Rotterdam, The Netherlands.

2. Center for Health Technology Assessment and Outcomes Research, University of Milan, Milan, Italy

3. Department of Epidemiology, Biostatistics & Health Technology Assessment, Radboud University Nijmegen Medical Centre, The Netherlands

Publication:

Stolk EA, Oppe M, Scalone L, Krabbe PF. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010;13(8):1005-13.

Introduction

Composite measures of health outcomes such as 'quality-adjusted life years' (QALYs) require weights or values attached to different health states that reflect the levels of health associated with these states. The standard gamble (SG) and time trade-off (TTO), which have emerged from health economics research, are frequently used to assign values to health states [1]. Psychology has contributed another technique, the visual analogue scale (VAS) [2]. Unfortunately, there are theoretical and empirical drawbacks to all of these techniques [3]. Responses to the SG and TTO are likely to be influenced by factors extraneous to judgments about health levels, such as risk aversion or time preference. Moreover, empirical violations of the normative axioms supporting the use of these techniques have been noted. Regarding VAS, critics question its interval properties and point to its lack of a relation to economic theory. In the literature on health-state valuation, arguments are raised for and against different techniques, but this debate has not led to consensus [4]. Therefore, but also in light of the diverging empirical results, continued work on improving the methods is warranted.

Probabilistic discrete choice (DC) modelling offers an alternative approach for exploring people's values. Such DC models can be used to analyse data obtained through approaches involving choices, ranks, or matches between alternatives, as defined by attributes and levels [5]. This strategy was first developed in transport economics and marketing. There, instead of modelling people's actual choices (revealed preferences), Louviere and others modelled the choices made by subjects in carefully constructed experimental studies based on stated preferences: discrete choice experiments (DCE) [6]. This approach also made it possible to predict values for alternatives that could not be judged in the real world (i.e., hypothetical situations or conditions). Recently, DC modelling has attracted much attention in the area of health evaluation. The framework offers a conceptual basis for the evaluation of the benefits of health programs. DCE was introduced into health economics to evaluate health-care products going beyond the QALY paradigm. The technique is used to evaluate aspects of health care like waiting time, location of treatment, and type of care [7-9] but also to quantify health outcomes [10-14].

DC modelling has good prospects for health-state valuation. The statistical literature classifies it among the probabilistic choice models that are grounded in modern measurement theory and consistent with economic theory (i.e., the random utility model). All DC models have in common that they can establish the relative merit of one phenomenon with respect to others. If the phenomena are characterized by specific attributes with certain levels, extended probabilistic choice models would permit estimating the relative importance of the attributes and their associated levels, and even estimating overall values for different combinations of attribute levels. A promising feature of DC models is that the derived values only relate to the attractiveness of a health state; they are not expressed in trade-offs between improved health and something else, as in TTO and SG. Bias due to these extraneous factors may therefore be prevented. Moreover, DC models have a practical advantage: when conducting DCEs, health states may be evaluated in a self-completion format. The scope for valuation research is thereby widened as compared to existing TTO protocols for deriving values for health-state measurement instruments such as EQ-5D.

But DC models are not without problems when used for health-state valuation. The analytical procedure on which analysis of DCE data is based assumes that the difference in values between choice options (e.g., two health states) can be inferred from the proportion of respondents that chose one option over the other. This implies that the relative position of all health states on the latent scale would lie between the 'best' and the 'worst' health states. For the estimation of QALYs, however, those values need to be scaled on the full-health – dead scale. If DC modelling is used to value health, a way must be found to link the derived values under this model to the scale required to calculate QALYs. Yet there is no consensus on what is the best way to handle the arbitrarily scaled DC values obtained, so it remains uncertain just how valid and informative DC-based values are.

The first step in any applied procedure for rescaling DC values may be to rescale by anchoring them on values obtained for the best and worst health state using other valuation techniques, such as TTO or SG. This approach might not be ideal, however, since part of the motivation to explore the DC model as a potential candidate to produce health-state values comes from the limitations of existing valuation methods. Alternatively, the DCE may be designed in such a way that the derived health-state values can be related to the value of the state 'dead'. A simple manner to achieve this is to design a DCE in which respondents are presented one health state at a time and asked if they consider it better or worse than being dead. The difference between a health state and being dead can then be estimated from the observed probability that the respondents would prefer to be dead. However, the precision of the estimates will critically depend on the proportion of people who prefer each state over being dead and on the consistency of choices for each state, as explained by Flynn et al. [15]. Both problems are less likely to arise in studies comparing health states to each other rather than to being dead. By mixing these designs, the ability to relate the health-state values to being dead may be maintained, as demonstrated by McCabe et al. [13] and Salomon [16]. These authors mixed the state 'dead' in the choice set as a health state, so that a parameter for the state 'dead' is estimated as part of the model. Doing so provides the information needed to rescale the values while limiting (not omitting) the effect of the aforementioned biases.

It is hard to say beforehand which method of deriving QALY weights with the DC model would produce optimal results: anchoring on external TTO values; or anchoring the DCE-derived values on being dead. Experimentation with DC modelling is therefore required to see how these difficulties may be resolved in a particular situation.

This paper considers the application of DC modelling for deriving health-state values. Research on novel, enhanced, and feasible measurement tools is conducted by the EuroQol Group to support improvement of the Group's health-status measurement instrument, the EQ-5D. This work is motivated by the perceived limitations of the traditional valuation techniques and by the prospects of DC models for health-state valuation. We analysed congruence across methods (DC, Rank, VAS, and TTO) and across samples with the aim of determining whether DC modelling produces value estimates that are comparable to traditional methods. The main focus of the study was to compare DC values to values elicited with the standard TTO technique.

Methods

EQ-5D states

The EuroQol EQ-5D is a generic measurement instrument to describe and value health states [17]. The EQ-5D classification describes health states according to five attributes: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression. Each attribute has three levels: 'no problems'; 'some problems'; and 'severe problems'. Health-state descriptions are constructed by taking one level for each attribute, thus defining 243 (3^5) distinct health states, where '11111' represents the best and '33333' the worst state. An EQ-5D health state may be converted to a single summary index by applying a formula that essentially attaches weights to each of the levels in each dimension. This formula reflects the values of EQ-5D health states as obtained from respondents in a sample of interest. Usually this is a representative sample of the general population, but in the current study both a student sample and a general population sample were used.

Not all EQ-5D states were included in the experiment. We constructed a discrete choice experiment (DCE) of 60 pairs of EQ-5D states, following the methodology described below. For the three other judgmental tasks in our study protocol, a set of 17 EQ-5D health states was selected. The set comprised five very mild, four mild, four moderate, three severe states, and state '33333'. The 17 states are: 11112, 11113, 11121, 11131, 11133, 11211, 11312, 12111, 13311, 21111, 22222, 23232, 32211, 32223, 32313, 33323, and 33333. The same 17 states were used in the Dutch EQ-5D TTO valuation study [18].

Respondents

For practical reasons, this study included a general population sample (target N = 400) and a student sample (target N = 200). Methodological issues were studied only in the student sample to reduce cost, but the validity and feasibility of DC modelling for health-state valuation were eventually explored in the general population, in line with the societal perspective taken in most economic evaluations.

Students were recruited at Erasmus University in Rotterdam, The Netherlands. Each student was offered € 20 for participating. The general population sample consisted of members of an Internet panel. This panel included approximately 104,000 people. Stratified sampling was used to select a research sample from the panel that was representative for the Dutch general population in terms of age, gender, and education. The stratified sampling procedure was performed in three rounds, so the final round allowed for over- or under-sampling of specific groups if the desired distribution over the strata had not been attained yet. The incentive offered to the panel members consisted of a € 2.50 donation to a charity chosen by the respondent and a chance to win gift certificates or other prizes in a lottery.

People in the general population sample were only administered the DCE. The students completed (in this order) the DCE, Ranking, VAS, and TTO task in the presence of one of the researchers or a research assistant. To become familiar with the type of health-state descriptions, all respondents were administered the EQ-5D prior to the judgmental tasks.

Judgmental tasks

DCE – In the DCE, all respondents were presented with a forced choice between two EQ-5D states. After this paired comparison task, students were prompted to answer a second question related to each of the two health states separately. This extra question offered 'dead' as a choice, phrased as "would you rather be dead than living in this health state?" In the remainder of the paper, we will refer to the two outcomes as DCE data and DCE_{dead} data respectively.

The DCE was programmed as a computer experiment. Respondents logged in to a website where they were presented with a number of choices between two EQ-5D states that were randomly selected from the choice set. Our general population sample received nine discrete choices; students received 18 discrete choices, and thus compared 36 states to being dead.

Ranking, VAS, and TTO – The Ranking, VAS, and TTO tasks were performed as described in Lamers et al., 2006 [18]. The valuation procedure may be summarized as follows. First, students rank-ordered the 17 EQ-5D states selected for these tasks, supplemented with 'dead' and state '11111', by putting the card with the 'best' health state on top and the 'worst' one at the bottom. Next, students valued the rank-ordered health states on the EuroQol visual analogue scale (EQ-VAS) using a bisection method that specified the order in which various states needed to be valued. The TTO valuation task followed the VAS valuation. TTO was executed using a Computer Assisted Personal Interviewing (CAPI) method that followed standard TTO protocols based on the original UK study protocol [19]. This implies that the health states were presented in random order, that the TTO task was facilitated by a visual aid, and that the respondents were led by a process of outward titration to select a length of time *t* in state '11111' (perfect health) that they regarded as equivalent to ten years in the target state (for states better than dead) or to select a length of time (10 – *t*) in the target state followed by *t* years in state '11111' (for states worse than dead).

Experimental design of the DCE

The DCE design was constructed using a Bayesian efficient approach, which to our knowledge has not been applied in health economics before. Most DCEs in health economics have applied orthogonal designs. These allow the uncorrelated estimation of main effects, assuming that all interactions are negligible. A limitation of orthogonal designs is that orthogonality is compromised if, for the purpose of data analysis, categorical multi-level variables need to be transformed into a set of dummy variables. Moreover, in optimal orthogonal designs the efficiency of the design is optimized for the situation that choices are made randomly. This is true under the restrictive assumption that the estimates of the parameters in the utility model are equal to zero ($\beta = 0$). This implies that two choice options within a pair have a 50% probability of being preferred, irrespective of their attribute levels. If $\beta = 0$ does not hold, the design will not be optimally efficient for producing information in regard to the true parameter effects [20,21]. Both issues with orthogonal designs apply to EQ-5D valuation, so we decided to look elsewhere.

To construct a Bayesian efficient design, a computer algorithm was used (see Appendix). The algorithm entailed an iterative procedure whereby a great many designs, each with the desired number of 60 choice situations, were randomly selected from the full factorial design and compared by their D-error, which was computed on the basis of expected values of the model parameters. In the Bayesian framework these expected values are known as priors. Because the priors were not perfectly known, they were included as distributions from which they were sampled rather than as point estimates in the design algorithm. This way, when priors deviate from their expected values, the impact on the efficiency of the design is minimized. To that end, the Bayesian efficient design algorithm uses nested Monte Carlo simulation. The best design remaining after 2000 iterations, each containing

1000 draws for the priors, was selected for this study. The probability that this design is the optimal one is small since a more efficient design is likely to exist. Even if not optimal, the design will still be efficient, given the large number of iterations in the Monte Carlo simulation.

The DC model we intended to estimate included main-effect terms for the five categorical three-level EQ-5D domains (transformed into a set of ten dummies) and the so-called N3 term. This is a non-multiplicative interaction term that is frequently used in EuroQol valuation models. It allows for measuring the 'extra' disutility when reporting severe (level 3) problems on at least one EQ domain [17]. Accordingly, a minimum number of 11 pairs is required to estimate all model parameters. It was decided to increase this number to 60 pairs to allow for extension of the model with interaction terms, if relevant.

The priors for the main effects were obtained by taking the weighted average of the parameter estimates from three TTO-based EQ-5D studies [18,19,22]. We used a standard error of 20% surrounding these priors to account for the possibility that parameter estimates modelled on the basis of DCE data might be different from those elicited with TTO. The prior parameter estimates of the interactions were set to 0 (table 8.1).

Table 8.1: Model parameters for the Bayesian efficient design.

Main effects*	Priors for main effects	Interactions (priors = 0)			
MO2	–0.108	MO2*SC2	SC2*UA2	UA2*PD2	PD2*AD2
MO3	–0.434	MO2*SC3	SC2*UA3	UA2*PD3	PD2*AD3
SC2	–0.140	MO2*UA2	SC2*PD2	UA2*AD2	PD3*AD2
SC3	–0.346	MO2*UA3	SC2*PD3	UA2*AD3	PD3*AD3
UA2	–0.090	MO2*PD2	SC2*AD2	UA3*PD2	
UA3	–0.240	MO2*PD3	SC2*AD3	UA3*PD3	
PD2	–0.147	MO2*AD2	SC3*UA2	UA3*AD2	
PD3	–0.463	MO2*AD3	SC3*UA3	UA3*AD3	
AD2	–0.119	MO3*SC2	SC3*PD2		
AD3	–0.354	MO3*SC3	SC3*PD3		
		MO3*UA2	SC3*AD2		
		MO3*UA3	SC3*AD3		
		MO3*PD2			
		MO3*PD3			
		MO3*AD2			
		MO3*AD3			

* The abbreviations MO2 to AD3 represent the five categorical three-level EQ-5D domains transformed into a set of ten dummies. The first level (no problems) was used as reference category.

The algorithm produced a design of 60 pairwise comparisons of two EQ-5D states. To further improve the design, we identified and altered dominant choices in which logical consistency predicts that one alternative will always be preferred. Nine dominant choices were identified. In five pairs, the worst state was improved to escape from dominance; in

Table 8.2: Final set of 60 pairs of EQ-5D health states for the DCE (asterisk marking the 9 states that were manually altered).

Choice	Option 1	Option 2
1	21231	22323
2	23223	31113
3	11112	12221
4	33322	23312
5	22331	23233
6	32133	22312
7	33123*	22233*
8	23212	32121
9	32322	33131
10	11231	32111*
11	33222	11312
12	13122	21212
13	22221	13212
14	22312	11212
15	22132	12321
16	12332	31333
17	22333	33332
18	31222	12112
19	31131	13111
20	12233	13132
21	31131	12121
22	33131	21323
23	33122	31132
24	11133	32211*
25	12231	21121
26	12312	13131
27	21111*	11311
28	11223	12313*
29	13231	31231
30	31123	12212
31	13211	21233
32	33311	22133
33	32112	23312
34	21112	22111
35	32211	13333
36	13131	13113
37	22313	23231
38	31313	32231
39	12123	33321
40	22311	32123
41	11133	21123
42	31311	21313
43	21212	32213
44	11121	22112*
45	13313	31221
46	21321*	12111
47	33323	23122
48	11223	32321
49	23313	32222
50	31323	22321
51	33113*	32332
52	22131	21212
53	23222	31113
54	12222	33121
55	31132	21333
56	12213	31232
57	23312	13123
58	21211	32313
59	31133	21331
60	13321	13231

the other four, the best state was made worse. The alterations were made randomly, but in accordance with the following rules: 1) the D-efficiency of the design was improved with the alterations; and 2) the new health state was not included yet in the choice set. This strategy resulted in a choice set of 60 pairs including 106 unique health states (94 states were included once, 10 twice, and two were included three times).The final set of 60 states is presented in table 8.2. The D-error of this design was 1.11.

Analysis

The rank data was analysed using the 'law of comparative judgment' (LCJ) model, as introduced by Thurstone [23,24]. To model the rankings within the Thurstonian framework, the rankings are transformed ('exploded') into paired comparisons. The analytical procedure assumes that the difference in value between two health states can be inferred from the proportion (i.e., probabilities) of respondents who preferred one health state to another. The resulting matrix of probabilities is subsequently transformed into Z-values (i.e., normal distribution). The LCJ values are obtained by taking the mean of all the columns of the Z-matrix, as described by Krabbe [24].

Mean VAS and TTO values were obtained with approaches commonly used in EQ-5D valuation studies [described, for example, in 17-19]. Observed VAS values were obtained on a scale with the endpoints 'best imaginable health' (=100) and 'worst imaginable health' (=0). To use these values in health-state valuation, they need to be rescaled such that state '11111' has a value of 1 and being dead has a value of 0. Rescaling was performed at the respondent level on the basis of the observed VAS scores for the various health states and the scores that were recorded for 'dead' and 'perfect health,' using the following equation [17]:

$$\text{VAS}_{\text{health state-rescaled}} = \frac{\text{VAS}_{\text{health state-raw}} - \text{DEAD}_{\text{raw}}}{11111_{\text{raw}} - \text{DEAD}_{\text{raw}}}$$

The same procedure that was applied in the Dutch valuation study [18] was used for estimating values from TTO responses. For states regarded as better than dead, the TTO value is $t/10$; for states worse than dead, values are computed as $-t/(10-t)$. These negative health states were subsequently bounded at minus 1 with the commonly used transformation $v' = v/(1-v)$. Linear regression analysis was used to interpolate values for all EQ-5D states from the values for the 17 states that were observed.

For the TTO task, the predicted values for all 243 EQ-5D states were derived after interpolation from the values for the 17 states that were included in the TTO task. The TTO model included an intercept, interpreted as any deviation from full health, as well as dummy variables for the ten main effects and for the N3 parameter.

We modelled and rescaled DCE-derived values in two different ways. The applied DC models were a conditional logit model (estimated only on the DCE data) and a rank-ordered logit model (estimated on DCE *and* $\mathrm{DCE}_{\mathrm{dead}}$ data), as explained below.

First we used the conditional logit model to analyse the DCE data obtained from the 60 pairwise comparisons of EQ-5D states. The model included dummy variables for the ten main effects and the N3 parameter. The values derived from this model are on an undefined scale. To link the DCE-derived health-state values to the QALY scale, we used TTO values for the worst health state (33333) and the best health state (11111) as anchor points for rescaling. For the general population we used TTO values obtained from the Dutch EQ-5D valuation study. For the student sample, we used the empirical TTO values derived in this study. We will refer to the resulting values as the DC values.

Alternatively, we derived health-state values from the DCE data on the QALY scale by anchoring the values on the value for being dead (thus: 0). For this purpose we modelled the information obtained from both the DC and $\mathrm{DC}_{\mathrm{dead}}$ data. This information was used to deduce how the respondent would have rank ordered the two EQ-5D states and 'dead' from most to least preferred. These rank orderings were analysed using a rank-ordered logit model. Besides the dummy variables for the ten main effects and the N3 parameter, this model also includes a parameter for the state of being dead, which can be used to rescale the values and put them on the full-health – dead (1 – 0) scale, as demonstrated by McCabe et al. [13]. The value for being dead is anchored at zero by dividing all coefficients by the coefficient for 'dead.' By additionally restricting the value of full health to 1, values are produced in the 0 to 1 range for states better than dead and negative values for states worse than dead. We will refer to the resulting value set as $\mathrm{DC}_{\mathrm{dead}}$

The two DC models are both variants of the multinomial logit model that is frequently used for analysis of DCE data [5,25]. The latter makes the simplifying assumptions that the error terms are independently, identically distributed (the IID assumption) and that the ratio of the probabilities of two alternatives *i* and *k* does not depend on any alternatives other than *i* and *k* (the IIA assumption). Several other models relax the IIA assumption, thereby alleviating concerns about bias due to its violation. Examples include the mixed logit model, the generalized extreme value model, and the probit model [25]. The first of these is considered the most promising for discrete choice analysis [26]. While mixed logit models are arguably more powerful, they also require higher data quality. For practical reasons, we decided to power the current study just for conditional logit estimation and not for mixed logit estimation. Our aim was to make a global comparison of TTO and DC values and then to study the strengths and weaknesses of various ways of anchoring relative DC values on the QALY scale. If this study generates satisfactory results, we anticipate future studies whereby the design would be optimized in relation to selected anchoring strategies as well as to bias minimization.

Intraclass correlation coefficients (mixed model, average measures) and mean absolute differences were computed to estimate the degree of correspondence between different methods. Except for the DC model (Stata 10 SE), all statistical analyses were performed in SPSS (V. 17.0).

Results

Respondents

Data were elicited in a sample of 444 persons in the general population and 209 students. The general population sample was representative in terms of gender, age, and level of education (table 8.3).

Table 8.3: Characteristics of the two samples.

	Sample (N = 444)	General population norms* (%)	Students (N = 209)
Male, % (N)	48.2 (214)	50.1	30.6 (64)
18-24	3.8 (17)	5.9	79.7 (51)
25-34	7.9 (35)	9	18.8 (12)
35-44	10.8 (48)	11.3	1.5 (1)
45-54	9.7 (43)	10.1	-
55-64	10.4 (46)	8.6	-
65-74	5.6 (25)	5.2	-
Female, % (N)	51.8 (230)	50	69.4 (145)
18-24	4.7 (21)	5.8	82.7 (120)
25-34	9.2 (41)	9	16.5 (24)
35-44	11.5 (51)	11.1	0.8 (1)
45-54	10.4 (46)	9.9	-
55-64	10.1 (45)	8.5	-
65-74	5.9 (26)	5.7	-
Marital status, % (N)			
Single	23.4 (104)	-	68.4 (143)
Married/living together	59.0 (262)	-	16.7 (35)

Table 8.3: Characteristics of the two samples (*continued*).

	Sample (N = 444)	General population norms* (%)	Students (N = 209)
Widowed	3.2 (14)	-	-
Divorced	10.4 (46)	-	1.4 (3)
Missing, other	4.1 (18)	-	13.5 (28)
Educational level, % (N)			
Low	27.0 (120)	26.3	-
Middle	40.1 (178)	42.5	-
High	32.9 (146)	31.3	100.0 (209)
Age, Mean (SD)	45.5 (14.6)	-	22.7 (3.4)
EQ-5D index, Mean (SD)	0.83 (0.23)	-	0.93 (0.1)

* source: Survey Sampling International, Minicensus data (Netherlands).

Preference data elicited in students using Ranks, VAS, and TTO

The observed mean values for the 17 health states that were obtained in students using Ranks, VAS, and TTO are presented in table 8.4 and – supplemented with the DC values – in

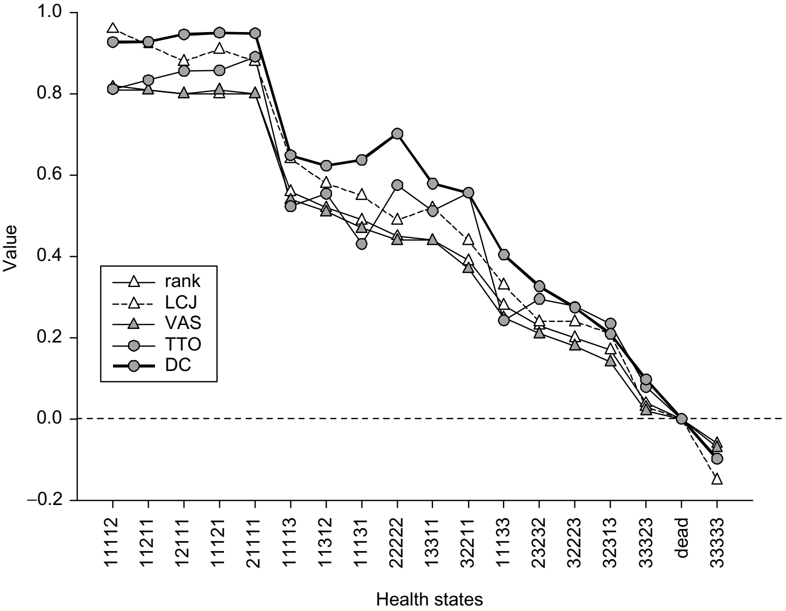


Figure 8.1: Comparison of values elicited from the student sample: Observed rank, Thurstone scaling (LCJ) based on ranks, VAS, and TTO values for the 17 empirically measured EQ-5D health states, and the derived values of the same 17 states based on the DC task (DCE).

Table 8.4: Observed and rescaled Ranks, VAS, and TTO values (means, SD) for the 17 EQ-5D states.

State	Ranks			Thurstone (exploded ranks)	VAS (observed)		VAS (normalized)			TTO (observed)	
	Mean	SD	Rescaled	LCJ	Mean	SD	Mean	SD	Rescaled	Mean	SD
11111	1.01	0.21	1.00	-	98.83	3.35	100.00	0.00	1.00	1.00	-
11112	4.17	1.97	0.81	0.96	82.72	12.87	81.97	14.20	0.82	0.81	0.28
11211	4.20	1.78	0.81	0.92	82.07	11.46	80.99	14.53	0.81	0.83	0.25
12111	4.25	2.27	0.80	0.88	81.03	15.79	79.61	21.14	0.80	0.86	0.25
11121	4.34	1.92	0.80	0.91	81.54	13.69	80.80	14.31	0.81	0.86	0.22
21111	4.36	2.48	0.80	0.88	80.99	14.41	79.87	15.93	0.80	0.89	0.18
11113	8.25	2.99	0.56	0.64	57.50	21.18	53.98	23.79	0.54	0.52	0.44
11312	8.97	2.19	0.52	0.58	55.12	16.16	51.18	20.21	0.51	0.55	0.34
11131	9.40	2.83	0.49	0.55	51.20	19.38	46.73	22.18	0.47	0.43	0.45
22222	9.98	1.92	0.45	0.49	48.47	14.59	43.70	18.63	0.44	0.58	0.36
13311	10.27	2.76	0.44	0.52	48.26	18.14	43.71	22.84	0.44	0.51	0.38
32211	11.05	2.81	0.39	0.44	42.78	18.41	37.33	24.99	0.37	0.56	0.40
11133	12.85	2.94	0.28	0.33	31.54	18.90	24.85	23.52	0.25	0.24	0.49
23232	13.62	2.21	0.23	0.24	27.42	14.09	20.47	18.60	0.21	0.29	0.43
32223	14.21	1.94	0.20	0.24	24.75	13.35	17.49	18.42	0.18	0.28	0.44
32313	14.66	2.01	0.17	0.21	21.98	13.50	13.89	22.86	0.14	0.23	0.46
33323	16.78	1.50	0.04	0.03	10.44	9.69	1.47	17.57	0.02	0.08	0.49
dead	17.43	1.94	0.00	0.00	7.59	11.38	0.00	0.00	0.00	-	-
33333	18.37	0.82	−0.06	−0.15	3.68	5.87	−6.62	21.11	−0.07	−0.10	0.48

figure 8.1. All methods yielded a negative value for state 33333 (rank: −0.06; LCJ: −0.15; VAS: −0.07; TTO: −0.1). Compared to the Dutch TTO-based valuation algorithm, the student sample gave on average slightly higher values for the health states (not presented). The intraclass correlations between the four value sets were high (>0.96). Yet the absolute values differed across the methods, in particular between VAS and LCJ, between TTO and VAS, and between TTO and LCJ. VAS values tended to be lower than TTO values. However, the mean ranks were similar to the VAS values. This similarity may be caused by the relation between the judgmental tasks of the ranking and VAS: the rank-ordered health states were valued using VAS in a specific order. Application of LCJ to rank data resulted in values that were higher than VAS and TTO values.

Comparing DC and TTO

Table 8.5 presents the parameter estimates obtained for DC, DC_{dead}, and TTO. We only

Table 8.5: Parameter estimates for the models based on data derived by DCE and TTO.

	General population			Students								
	DCE			DCE			DCE$_{\text{dead}}$			TTO		
	Coef	SE	sign	Coef	SE	sign	Coef	SE	sign	coef	SE	sign
Constant*	N/A			N/A			N/A			−0.103	0.02	0.000
MO2	−0.267	0.07	0.000	−0.364	0.08	0.000	0.297	0.07	0.000	−0.012	0.02	0.603
MO3	−1.430	0.08	0.000	−1.430	0.09	0.000	1.169	0.07	0.000	−0.091	0.03	0.001
SC2	−0.536	0.06	0.000	−0.382	0.07	0.000	0.296	0.06	0.000	−0.055	0.02	0.012
SC3	−1.092	0.08	0.000	−0.830	0.08	0.000	0.691	0.07	0.000	−0.079	0.03	0.002
UA2	−0.303	0.07	0.000	−0.515	0.08	0.000	0.410	0.07	0.000	−0.054	0.02	0.021
UA3	−0.887	0.08	0.000	−1.337	0.09	0.000	1.062	0.07	0.000	−0.169	0.03	0.000
PD2	−0.143	0.06	0.026	−0.354	0.07	0.000	0.335	0.06	0.000	−0.087	0.02	0,000
PD3	−1.330	0.08	0.000	−1.751	0.09	0.000	1.521	0.07	0.000	−0.297	0.02	0,000
AD2	−0.470	0.07	0.000	−0.516	0.08	0.000	0.424	0.07	0.000	−0.069	0.02	0.001
AD3	−1.499	0.08	0.000	−1.667	0.08	0.000	1.351	0.07	0.000	−0.231	0.02	0.000
N3	−0.599	0.12	0.000	−0.844	0.14	0.000	0.918	0.13	0.000	−0.128	0.02	0.000
Dead dummy	N/A			N/A			6.066	0.16	0.000	N/A		
model fits	pseudo R^2		0.25	pseudo R^2		0.29	pseudo R^2		0.46	R^2		0.35

* The constant is not always estimated and has different meanings when it is, due to difference in the scale on which coefficients are estimated. “DCE” coefficients indicate the impact of a “one-unit” difference in the independent variables which are 10 dummies representing jumps from the no-problem level on an EQ domain to the level of some or severe problems. The parameters are thus estimated relative to full health, a state on which we superimpose the value of 11111. On the basis of provided parameter estimates the value for EQ-5D state 33333 can be computed on the same arbitrary scale, which is −5.837 for the general population and −6.859 for students. A constant is not estimated in the multinomial (conditional) logit model, because this is a difference model so the constant drops out. “DCEdead” coefficients indicate the impact of a “one-unit” difference in the independent variable on the probability that a condition is considered to be worse than being dead. The coefficients are estimated on a scale with two anchors: dead (on which we superimpose the value of 0) and a top anchor represented by the constant. Here the constant thus represents the value for full health. “TTO” coefficients are measured on an absolute scale, anchored by full health (1.0) and dead (0.0). Here, the constant represents the disutility associated with any deviation from full health in so far as it is not attributable to any of the five domains.

report the models that included the N3 parameter, because these performed slightly better than the models without N3. All coefficients were statistically significant, except for mobility level 2 in the TTO model.

A strong relationship was observed between the values obtained for the 243 EQ-5D states predicted by the DC model based on the general population and student DCE data (figure 8.2). Although more health states seem to be valued negatively by the general population, this is mostly due to the rescaling on the basis of the TTO value for the worst EQ-5D state, ‘33333’. Comparison of figures 8.2 and 8.3 suggests that the parameter estimates

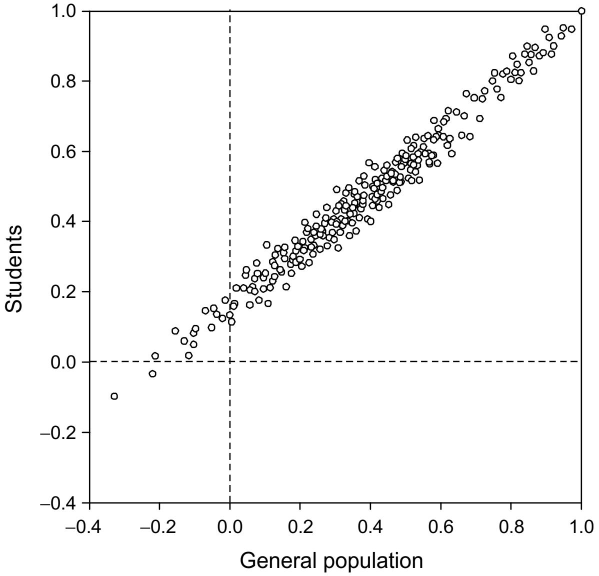


Figure 8.2: DC values for the 243 EQ-5D health states derived from discrete choice judgments by the general population (Dutch) compared with values derived from similar judgments by Dutch students. The student DC values were anchored on the TTO values for 11111 and 33333 from the student TTO task carried out in this study. The general population DC model was anchored on TTO values for 11111 and 33333 from the Dutch valuation study. The differences between the DC values for the students and those for the general population originate from differences in responses to the DC experimental task between the two groups as well as from differences between the TTO values from students and those from the Dutch valuation study, propagating through the DC values via the rescaling procedure.

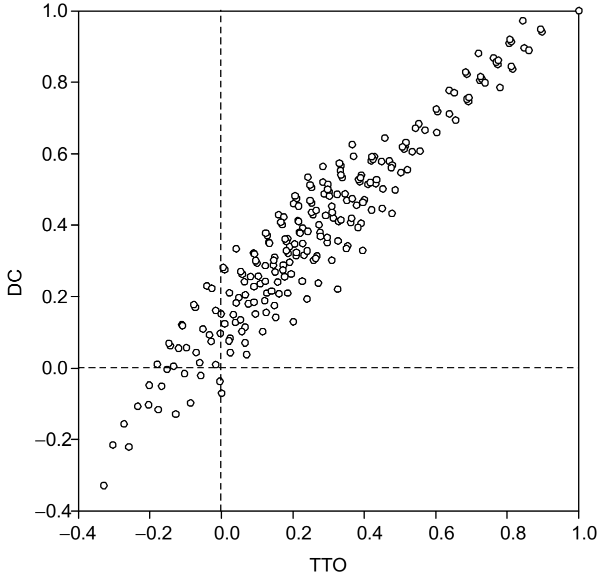


Figure 8.3: Comparison of TTO (Dutch algorithm) values with DC (Dutch general population) values.

obtained using DCE in different samples are closer to each other than the DCE-derived and TTO-derived estimates. Figure 8.3 shows that DC produced higher values than TTO when rescaled on the basis of the TTO values for '33333' and '11111'. The intraclass correlation between the TTO and DC values was 0.93 in the general population and 0.96 among the students. The mean absolute difference between the student TTO and student DC was 0.060 (SD = 0.039).

Absolute values of health states derived by different methods may be different, though in many applications of health-state values the main focus is on marginal differences (e.g., comparisons before and after a medical intervention). Therefore, marginal difference scores for all combinations of the 243 derived EQ-5D states were computed (29,403 combinations) for the TTO and the DC separately (figure 8.4). This analysis shows again that overall DC values are higher than TTO values and also that marginal differences between TTO and DC values for individual pairs of states can be as large as around 0.20.

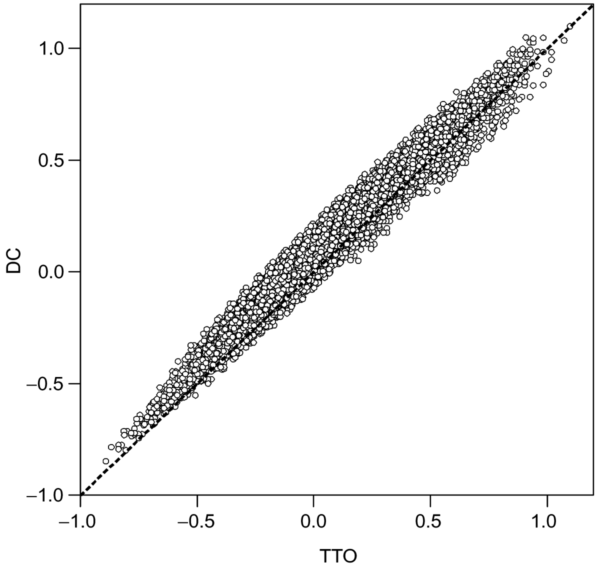


Figure 8.4: Marginal difference scores between the derived values of the 243 EQ-5D states (29,403 combinations) for the TTO (Dutch students) and the DC (Dutch students).

Anchoring DC values on 'dead'

Students considered a health state to be worse than dead in about 10% of the cases. The DC model parameter estimates derived from the DC_{dead} data are presented in table 8.5.

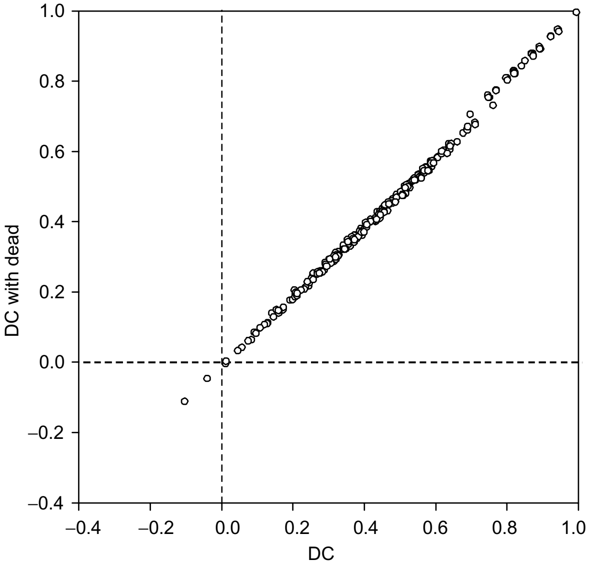


Figure 8.5: The DC values (Dutch students) derived from discrete choices between pairs of EQ-5D health states compared with the DC values (Dutch students) derived from discrete choices of separate EQ-5D health states plus being dead.

The values produced by the two different models (DC vs. DC_{dead}) are congruent (figure 8.5); the intraclass correlation between the two value sets was 0.98, while the mean absolute difference between the values was 0.019 (SD 0.009). The DC_{dead} values were slightly lower than values derived from the DCE involving pairwise comparison of EQ-5D states, except for mild health states. Therefore, the difference between the DC_{dead} values and the TTO values was slightly smaller than the difference between the DC values and the TTO values.

Interaction terms in the DC models

The analysis of the DC models expanded with first-order interaction terms showed that ten of the 40 interaction terms were statistically significant. However, three main effects (mobility level 2, pain level 2, depression/anxiety level 2) were no longer statistically significant when compared to the main effect model. The increase in the amount of explained variance (pseudo R^2) due to the inclusion of the interaction terms was marginal (main effect: 0.266; main effects + interactions: 0.277).

Conclusions and discussion

We have presented a systematic comparison of Ranks and VAS, TTO, and DC (DCE-derived) values for EQ-5D health states in order to investigate whether or not modelling DCE data produces health-state values that are comparable to other conventional valuation techniques, TTO in particular. DC values broadly replicated the pattern found in TTO responses. This observation applies to both samples (general population, students) and to both strategies that were applied to anchor the DC values on the full-health (=1) – dead (=0) scale. Whether or not this degree of congruence will also be found in a sample of the general population remains to be seen. Besides similarities, there were also systematic differences. DC values were consistently higher than TTO values, which were in turn higher than VAS values. Values derived from rank data were higher when analysed using LCJ than when using mean ranks. Instead of the classic Case V model used here, more general Thurstonian models with unrestricted covariance structures may be more appropriate [27]. The results suggest a systematic difference across the methods, with DC values being the highest of all.

The fact that differences were found between DC modelling and TTO is in line with the findings of several other studies where DC models have been applied in the analysis of rank or DCE data. Salomon (2003) compared rank-based models and TTO for EQ-5D using data from the UK general population survey. He found that the rank-based models produced slightly higher values [16]. Ratcliffe et al. [11] compared TTO and DC modelling for a disease-specific outcome measure. DCE-derived values seemed higher than TTO values. A more complex relation was found between rank and TTO data, with better convergence for mild states. McCabe et al. (2006) compared values derived from rank data with standard gamble values for SF-6D and HUI health states; the rank data produced higher values [13]. It thus seems that TTO and DC models are largely measuring the same latent construct (quality of a health state), but the techniques do not produce identical results.

The main difficulty we met in applying DC models is that these models generate values on an arbitrary scale, not on the metric of the quality (of life) component of the QALY scale. We have explored the possibility of anchoring the values derived from DCE data on the QALY scale directly by using ‘dead’ as a choice option. This strategy yielded values that were comparable to those derived from the DCE where two EQ-5D states were compared to each other and anchored on the basis of some TTO values. Although this is a promising result with regard to the possibility of using DC models and their associated DCEs as a stand-alone valuation technique, further research is warranted to see if the DCE_{dead} approach *always* performs at least as well as a DC model that is anchored on TTO. We found substantial differences between the TTO values derived from students and those for the general population. Whether or not the comparability of two anchoring strategies will hold in a general population sample depends on their responses to choices between EQ-5D states and being dead, which were not included in this experiment.

If combined use of DC modelling and TTO is considered for health-state valuation, the strategy for linking DC and TTO data may need to be further explored. Our anchoring strategy involved using the observed TTO value for EQ-5D state 33333 and the set value of 1.0 for EQ-5D state 11111. Relative to these anchor points, we found that DC models produced higher values than TTO, but we do not know if these anchor points were well chosen. In future valuation studies combining DC and TTO, we might use a larger number of observed TTO values and then apply statistical routines to adjust the parameters of the DC model to fit the TTO dataset. By so doing, we might avoid some of the bias related to problems with the application of TTO for the valuation of states worse than being dead and see if that improves the comparability of TTO and DC values. This strategy might also circumvent the problem that the value difference between the best EQ-5D state (11111) and the other states cannot be reliably estimated in the DC model. Perfect health will always be chosen over other health states, which results in an infinite value difference. The computed relative distance between state 11111 and the other states will therefore not be based on empirical responses but modelled on basis of assumptions. Using '11111' as the anchor point may therefore have contributed to systematic differences between TTO- and DCE-derived values. In the paper, we have not made a comparative evaluation of the DCE task and information obtained from ranking of the 17 states, nor have we given a detailed theoretical elaboration of their common basis. Our use of rank responses was limited to checking the convergence between the valuations obtained using different methods (Rank, VAS, TTO, and DC) including a very modest application of the classical Thurstone model.

The modern measurement of DC models builds upon the early work and basic principle of Thurstone's 'LCJ'. In fact, the class of choice- and rank-based scaling models with its lengthy history (1927 to the present) is one of the few areas in the social and behavioural sciences that has a strong underlying theory. In this respect it may be interesting to explore the possibility of extending or combining DC models with other closely related (fundamental) measurement models, e.g. Rasch models and item response theory models [28,29]. This might be an important area for future research.

To conclude, we believe that a strategy based on TTO data supplemented by health-state values derived from DC modelling may be a feasible and accurate option. Although there are small differences in results from the two conceptually different valuation methods, there seems to be a clear systematic relation that would make conversion from one method to the other feasible and defendable.

References

1. Drummond MF, Sculpher MJ, Torrance GW, et al. Methods for the economic evaluation of health care programmes (3rd ed.) Oxford: Oxford University Press, 2005.
2. Krabbe PFM, Stalmeier PFM, Lamers LM, Busschbach van JJ. Testing the interval-level measurement property of multi-item visual analogue scales. Qual Life Res. 2006; 15(10): 1651-61.
3. Green C, Brazier J, Deverill M. Valuing health-related quality of life. A review of health state valuation techniques. PharmacoEconomics. 2000; 17: 151-65.
4. Salomon JA, Murray CJL. A multi-method approach to measuring health-state valuations. Health Econ. 2009; 15: 215-8.
5. McFadden D. Economic choices. Am Econ Rev. 2001; 91: 351-78.
6. Louviere JJ, Hensher DA, Swait JD. Stated choice methods. Analysis and application. Cambridge: Cambridge University Press, 2000.
7. Ryan M. Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation. Soc Sci Med. 1999; 48: 535-46.
8. Ryan M, Netten A, Skatun D, et al. Using discrete choice experiments to estimate a preference-based measure of outcome: an application to social care for older people. J Health Econ. 2006; 25: 927-44.
9. Farrar S, Ryan M, Ross D, et al. Using discrete choice modelling in priority setting: an application to clinical service developments. Soc Sci Med. 2000; 50: 63-75.
10. Hakim Z, Pathak DS. Modelling the Euroqol data: a comparison of discrete choice conjoint and conditional preference modelling. Health Econ. 1999; 8: 103-16.
11. Ratcliffe J, Brazier J, Tsuchiya A, et al. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ. 2009; 18: 1261-76.
12. McKenzie L, Cairns J, Osman L. Symptom-based outcome measures for asthma: the use of discrete choice methods to assess patient preferences. Health Policy. 2001; 57: 193-204.
13. McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. J Health Econ. 2006; 25: 418-31.
14. Szeinbach SL, Barnes JH, McGhan WF, et al. Using conjoint analysis to evaluate health state preferences. Drug Inf J. 1999; 33: 849-58.
15. Flynn TN, Louviere JJ, Marley AA, et al. Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. Popul Health Metr. 2008; 6: 6.
16. Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metr. 2003; 1: 12.
17. Szende A, Oppe M, Devlin N. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Dordrecht, The Netherlands: Springer, 2007.
18. Lamers LM, McDonnell J, Stalmeier PFM, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006; 15: 1121-32.

19. Dolan P. Modeling valuations for the EuroQol health states. Med Care. 1997; 35: 1095-108.

20. Ferrini S, Scarpa R. Designs with a-priori information for nonmarket valuation with choice-experiments: a Monte Carlo study. J Environ Econ Manage. 2007; 53: 342-63.

21. Rose JM, Bliemer MCJ. The design of stated choice experiments: The state of practice and future challenges. Working paper ITS-WP-04-09. The University of Sydney: Institute of Transport and Logistics Studies, 2004.

22. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 2005; 43: 203-20.

23. Thurstone LL. A law of comparative judgments. Psychol Rev. 1927; 34: 273-86.

24. Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care. 2008; 46.

25. Train KE. Discrete choice methods with simulation. Cambridge, UK: Cambridge University Press, 2003.

26. Hensher DA, Greene WH. The Mixed Logit model: The state of practice. Transportation. 2003; 30: 133–76.

27. Maydeu-Olivares A, Böckenholt U. Structural equation modeling of paired-comparison and ranking data. Psychol Methods. 2005; 10: 285-304.

28. Rasch G. Probabilistic models for some intelligence and attainment tests. Expanded edition with foreword and afterword by B.D. Wright. Chicago: University of Chicago Press, 1980.

29. Streiner DL, Norman GR. Item response theory. In: Health Measurement Scales: A Practical Guide to Their Development and Use (4th ed.). New York: Oxford University Press, 2008.

Appendix: Bayesian efficient design algorithm

The Bayesian efficient design specifications were programmed in Microsoft Excel. The Excel code uses nested (quasi) Monte Carlo simulations to compute the design with the lowest D error (i.e., the highest D efficiency). Our design was computed using 1000 Halton draws for the inner loop (i.e., varying the values for the priors) and 2000 simulations for the outer loop (i.e., calculating the most efficient design for each of the 1000 sets of priors). Halton sequencing is a quasi Monte Carlo technique whereby, instead of drawing randomly from a distribution, draws are taken "smartly" to ensure that there are no gaps in the sampled distribution. Therefore, it needs fewer draws to adequately reflect the original distribution.

```
Start outer loop
1) a set of 60 pairs of states is randomly generated.
Start inner loop
    2) A set of priors is drawn from which the utilities are calculated for each of the 120
       states contained within the set of 60 pairs.
    3) The Fisher information matrix, its inverse, the asymptotic variance covariance
       matrix (AVC matrix), and the determinant of the AVC matrix (i.e., the D error) are
       calculated.
    4) Steps 2 and 3 are repeated.
    End inner loop
6) The overall D error is calculated (i.e., the combined D error from the inner loops).
7) Steps 1 to 6 are repeated storing the design with the lowest D error.
End outer loop
```

Chapter 9

General discussion

In this thesis a number of studies are described all related to economic evaluations. As described in the introduction the mathematical techniques used in these studies originate from different disciplines but are all applicable in the development of the economic models for the assessment of cost-effectiveness of health care interventions. The outline of this chapter is as follows: first the results and implications of the methodological studies presented in this thesis are discussed. This will be followed by a broader discussion on the general role of mathematics in problem solving, started in chapter 1.

The first two studies included in this thesis (presented in chapters 2 and 3) are not methodological studies, but applied studies. They provide an illustration of the type of applications common in health economic research. For this reason they are part of the introduction, and the results from those two studies will not be addressed in this discussion, which focusses on methodological questions.

Sources of uncertainty in CUA outcomes

The first methodological study addressed the impact of four different methods of meta-analysis on the outcomes (i.e. cost-effectiveness) of a probabilistic Markov model for Chronic Obstructive Pulmonary Disease. We've shown that the choice for a particular meta-analysis technique has an impact on the uncertainty of the outcomes of an economic model. In particular we found that a Bayesian random effects model results in more uncertainty and a lower acceptability curve than fixed effect models and frequentist models.

This leads to the question of which model's estimates are closest to the "truth". Do frequentist and fixed effects models underestimate the uncertainty, or does the Bayesian random effects overestimate the uncertainty? No final decision can be made based on this study alone, because our study only included data from a few *samples* (which is typical for this type of study) and not from the entire *population*. The fact that pooled samples can generally be assumed to be representative for the population does not mean that this is also the case in this particular problem of assessing uncertainty in meta-analyses. In other words, the only way to objectively assess which type of meta-analysis model most accurately reflects the "truth" is if we know this "truth". A follow-up study was therefore initiated where we start by simulating a patient population of 50,000 patients. Next we draw samples from that population and combined the results using the different meta-analysis techniques. Finally we assess which meta-analysis technique produces estimates that are closest to the population values.

In the models used in economic evaluations it is common to include not only the mean values of the model parameters, but also the uncertainty of the model parameters in a probabilistic sensitivity analysis. In the case of the quality of life estimates (i.e. utilities) this uncertainty is based on the differences in health status between patients. However, the utilities themselves are obtained in a valuation study using a sample from the general population. Because of this there is also uncertainty associated with the estimation of the utilities. This uncertainty is typically ignored. It was shown by the within state variance compared to the between state variance, that the majority of this uncertainty does not originate from a lack of fit of the utility model (i.e. low between state variance), but from differences in the opinions of respondents in how good or bad a health state is (i.e. high within state variance).

We showed that varying the number of states observed and number of observations per state results in different estimates of the uncertainty. The question is whether or not the amount of uncertainty present in the observed or modelled TTO values is negligible compared to the heterogeneity due to differences in patient health status in the clinical trial. In other words, whether the uncertainty of the utility of a Markov state is dominated by the estimation uncertainty of the utilities, or by the patient heterogeneity with respect to the descriptive system. When 30 states and 300 observations per state are used for TTO based data, the expected uncertainty around the predicted utility values will be around 0.01, which is smaller than the typical patient variability in health status. This implies that in such a case the uncertainty related to the exact value of the utilities can safely be left out of a probabilistic sensitivity analyses (PSA).

However, if mapping between a disease specific instrument and a utility instrument is applied to obtain the utilities, then an additional level of uncertainty is introduced. This additional uncertainty can become so large that it is no longer negligible compared to the patient variability in health status and should be included in the PSA of an economic model. This will depend strongly on the level of agreement of the descriptive systems of the instruments that are used in the mapping procedure. As we've shown in the case of total hip replacement, there are marked differences in content between the disease specific questionnaire, the Oxford Hip Score (OHS), and the generic questionnaire, the EQ-5D. This, together with the fact that the data on which to base the mapping did not cover the entire range of the QALY scale, lead us to conclude that using a mapping model to predict individual patients utility values based on the OHS was not feasible.

New approaches in modelling health states

In many fields of physics geometry is the default type of mathematics used (e.g. vibration and wave mechanics). This is done because the mathematics is easier to solve when expressed. We've shown that this is also the case when using an approach based on polar coordinates (i.e. a geometrical approach) to model ratios. The distinct advantage over using an approached based on algebra is the removal of the singular behaviour that occurs when the denominator is (close to or) equal to zero, while the model structure itself is the same (i.e. the same utility model $U = \sum \beta_{ij} * x_{ij} + \varepsilon$ is estimated in both cases)

The pilot study assessing the feasibility of using DC to model EQ-5D utilities showed promising results. It was found that DC values broadly replicated the pattern found in TTO responses, although the DC values were consistently slightly higher than TTO values. The main difficulty in applying DC models was that these models generated values on an arbitrary scale, not on the metric of the quality (of life) component of the QALY scale. This means that DC-based values need to be anchored on the utility scale, where full health has a value of 1 and death has a value of 0.

However, the type of information captured in a DC task is different from that in captured in a TTO task. With DC respondents are asked to indicate which of two EQ-5D states they think is better, whereas in a TTO task, respondents are asked how much length of life they are willing to give up to avoid being in a particular health state. These are fundamentally different questions albeit on the same topic and both have their merits and problems.

TTO has as advantage that you directly ask values making the responses easy to interpret. A disadvantage is that it is more difficult to accurately model the data due to the high within state variance. An advantage of DC is that it is relatively easy to model due to limited within variance. However, only rank data is collected and not actual values leading to anchoring problems. Therefore both methods have opposing strengths and weaknesses, so they might be good as complementary sources of information rather than competing ones.

The role of mathematics in problem solving

The results from this thesis show that a variety of problems in health economics can be solved using mathematical frameworks and techniques that were developed outside this field. The origin and application of the mathematical techniques doesn't matter, as long as you are consistent in its use. For example, if you use a geometric approach, you'll have to define your problem in geometric terms, solve the geometry and interpret the results in geometric terms.

More in general: a real world question is answered by translating it to mathematics, solving the mathematics and translating the mathematics back into real world answers (see figure 9.1). Since solving the mathematics is mostly straightforward computation nowadays,

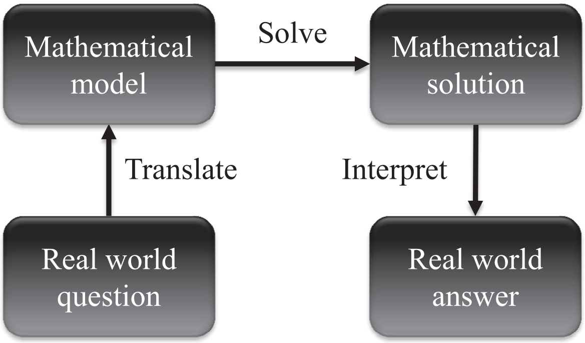


Figure 9.1: Problem solving using mathematics.

the most challenging steps are the translation into mathematics and the interpretation of the mathematical results. The researcher is completely free in their choice of mathematical model, given that it adequately describes the real problem. This choice is broader than the choice of specific (statistical) models or techniques. It also encompasses the choice of the basic type of mathematics that is best suited to solve the real world problem whether that is statistics or geometry, number theory or vector calculus.

Since all types of mathematics have been constructed based on logic and deductive reasoning, all of them are inherently "true" (i.e. a mathematical proof is inherently true). Therefore the mathematical results (if properly obtained) are also inherently "true". The question therefore is not so much which type of mathematics is correct, but which type of mathematics addresses the problem most appropriately and makes it easiest to solve. In other words, the key is in the translation of the real world problem into mathematics, not in solving the mathematics itself. One of the most important mistakes made is the choice of a mathematical model that does not adequately describe the real world problem. This not only goes for the type of mathematics that is used, but also for the type of statistical technique or measure that is used. An example to illustrate this is the use of Cronbach's alpha as a measure of reliability and internal consistency of a questionnaire. Cronbach's alpha is a measure that can be found in most basic psychometric textbooks and is part of the standard psychometric measures. It assesses the amount of agreement between the items that make up a questionnaire.

From a psychometrics perspective (or more precisely, from the perspective of Item Response Theory IRT), questionnaires are constructed to measure *a single underlying construct*. From that perspective, the more agreement between the items of a questionnaire the better. This is because the items are assumed to form a logical consistent representation of the underlying construct that the questionnaire reflects. It is of course very tempting to generalise this to all questionnaires and therefore also to use Cronbach's alpha as a measure of reliability for EQ-5D.

However, EQ-5D was constructed with economic (utility) theory in mind rather than psychometric theory. Because of these different perspectives, the theoretical basis underlying the EQ-5D and the typical IRT based questionnaires differs. When EQ-5D was constructed the idea was that even though the questionnaire was designed to assess a single concept, namely quality of life, it reflects not one but five different constructs (i.e. the 5 EQ-5D dimensions). This fundamental difference between the two perspectives has implications for the appropriateness of the statistical techniques that are used.

Because Cronbach's alpha measures the level of agreement between items, the validity of its use depends on the combination of number of constructs and number of items. If a number of items are supposed to measure the same underlying construct, it is a reasonable assumption that the answers have to be in agreement, and therefore that Cronbach's alpha is a good measure of reliability. However with instruments like EQ-5D, there are five items, but each measures a different construct[1]. Therefore the items should *not* be in agreement but should be independent. Thus, the level of agreement between items as measured by Cronbach's alpha does not reflect the reliability of EQ-5D even though it can be easily calculated.

In conclusion, I would like to remark that borrowing mathematical techniques that were developed in other disciplines can be very useful and rather straightforward to implement from a mathematical point of view, but one needs to think carefully about whether it is conceptually appropriate to do so.

[1] This can be shown by factor analysis. However, using factor analysis in such a way can be considered inappropriate from an IRT point of view.

Summary

Introduction

Health economics (HE) is a multi-disciplinary field. This is especially apparent in economic evaluations (EE) such as cost-utility analyses (CUA) which have become an integral part in the management of health care systems in many western countries. The mathematical techniques employed to obtain and describe the information needed for EE originate from three distinct mathematical disciplines associated with economics, psychology and medicine: econometrics, psychometrics and (bio)statistics. This dissertation shows how ideas and approaches from different disciplines can be applied in solving health economic problems.

The studies described in this thesis show how more specialised techniques and approaches can be used outside the field where they were originally developed. In particular they are used in the investigation of sources of uncertainty in CUA and in the measurement and valuation of health related quality of life. The techniques used include Monte Carlo simulation, Factor Analysis and Discrete Choice Modelling. The approaches include frequentist and Bayesian statistics and a regression model was based on geometry (i.e. directional statistics).

The first two studies included in this thesis provide an illustration of common concepts in health economic research: Health Related Quality of Life (HR-QoL) and economic evaluations. The next three studies are related to sources of uncertainty in the outcomes of cost utility analyses (CUA). The final two studies introduce two new approaches in the modelling of health state valuations.

Age dependency of self-reported health in Europe

In order to investigate the relationship between self assessed health related quality of life and age as measured by E-5D, we analysed population data from 10 European countries. We estimated several regression models where the EQ-VAS values were the dependent variable and found that a simple linear model outperformed more complex models. The linear model resulted in a constant = 93.12 and age parameter = −0.34 with adjusted $R^2 = 0.106$. Using the mean observed EQ-VAS scores per year of age instead of the raw data, (i.e. all the within variance is removed), results in a dramatic increase in R^2. The value of R^2 increases from 10.7% to 93.9% while the regressions coefficients stay the same. We found that differences between countries were larger than the difference between men and women.

Modeling of the proportion of reported problems has been carried out using logistic regression. Again the values of R^2 are low (ranging from 23% to 1%). Distinct differences were found between the relation with age of the physical domains mobility, self-care and usual activities, and the domains pain/discomfort and anxiety/depression. The physical domains start with few reported problems (<5%) at age 18, that increase with age at an increasing rate. Pain/discomfort and anxiety/depression start out higher (18% and 19%) and also increase with age but at a constant rate, which is higher for pain/discomfort than for anxiety/depression.

Economic evaluation of Panitumumab in mCRC

Colorectal cancer is the fourth most commonly diagnosed cancer. Recently, new drugs have become available to treat patients with metastatic colorectal cancer (mCRC). Panitumumab is one of those new drugs. The objective of this study was to assess the cost-effectiveness of panitumumab as monotherapy in mCRC after failure of other chemotherapy regimens for the purpose of temporary reimbursement in the Netherlands.

A micro simulation Markov model was developed to model the cost-effectiveness of panitumumab. The model contained three disease states (mCRC, progressive disease, and death) with time dependent transition probabilities between these states. Data from a pivotal trial comparing panitumumab plus Best Supportive Care (BSC) versus BSC alone in mCRC were used to estimate the model parameters. The base case model resulted in an estimated mean incremental cost-effectiveness ratio (ICER) of €64,321/QALY (95% CI [52,642; 94,187]). About half the patients in the trial suffered from non-mutated (wild-type) KRAS mCRC, which is indicative for response to panitumumab. Our model results for this subgroup of patients resulted in an estimated mean ICER of €64,541/QALY (95% CI [52,136; 90,706]).

An observational study with a follow-up time of 3 years is ongoing, where the focus is on establishing “real life” effectiveness and resource use instead of trial based efficacy and resource use. The results from that study will ultimately be used to re-estimate the model parameters and run the economic evaluation with real-life data. This will facilitate the decision on whether or not panitumumab will be considered for prolonged reimbursement in the Netherlands.

Impact of methods of data synthesis on CUA outcomes

Cost-effectiveness models should always be amendable to updating once new data on important model parameters become available. However, several methods of synthesizing data exist and the choice of method may affect the cost-effectiveness estimates. The goal of this study was to investigate the impact of the different methods of meta-analysis on final estimates of cost-effectiveness from a probabilistic Markov model for COPD. We compared 4 different methods to synthesize data for the parameters of a cost-effectiveness model for COPD: frequentist and Bayesian fixed-effects (FE) and random-effects (RE) meta-analyses. These methods were applied to obtain new transition probabilities between stable disease states and new event probabilities.

The four methods resulted in different estimates of probabilities and their standard errors (SE). The effects of using different synthesis techniques were most prominent in the estimation of the standard errors. We found up to nine-fold differences in standard errors of the exacerbation probabilities and up to almost three-fold differences in standard errors of the transition probabilities. In our study we found that the frequentist FE model produced the lowest means and SEs, whereas the Bayesian RE model produced the highest. The estimates of differences between treatments in total costs, QALYs and cost-effectiveness acceptability curves (CEAC) also varied depending on the synthesis method. With a Bayesian RE model the CEAC was 15% lower than with a frequentist FE model.

These results show that the choice of synthesis technique can affect resulting model parameters considerably, which can in turn affect estimates of cost-effectiveness and the uncertainty around them.

Statistical uncertainty in TTO derived utility values

The utility values for health states that are used in cost utility analyses are generally presented and used as point estimates, implying that they are perfectly known. However, since utilities are based on empirical valuation studies, they can't be perfectly known. We aimed to quantify the uncertainty surrounding utilities and to assess the association between the number of health states and respondents included in the valuation studies and the uncertainty of the utilities.

We analysed the uncertainty from two of the EQ-5D valuation studies: the Dutch TTO study and the UK MVH study. We used ANOVA to assess the appropriateness of the two valuation models and the uncertainty of the estimated utility for each health state. The impact of the number of respondents and health states on the uncertainty was investigated using Monte Carlo simulation. We simulated studies that included between 14 and 42 health states and sample sizes ranging from n = 25 to n = 1000. The standard errors of the utilities in the Dutch study ranged from 0.044 to 0.010 for the different health states, or 0.028 on average. In the UK study they ranged from 0.013 to 0.006 (0.009 on average). Our Monte

Carlo results showed that the standard errors ranged from 0.095 when 17 states are valued by 25 respondents to 0.009 when 42 health states are valued by 1000 respondents.

Based on our results we suggest that in order to obtain a standard error of the utilities below 0.01 on average for EQ-5D TTO valuation studies, one would require around 30 health states to be valued by about 300 respondents each.

Comparison of EQ-5D and Oxford Hip Score

Both disease-specific and generic patient reported outcome measures (PROMs) provide information about the health status of patients. Generally, disease-specific measures provide more clinical information than generic measures but do not provide a utility weight. The aim of this study was to assess the comparability of the information captured by a disease-specific measure, the Oxford Hip Score (OHS) and a generic measure, the EQ-5D, and the viability of mapping between them to obtain utilities for the OHS. Data for 439 NHS patients in England before and six months after undergoing total hip replacement were analysed. The information provided by the OHS and EQ-5D was assessed using principal component analysis and analysis of the correlation matrix. The predictive performance of four mapping models was based on the mean absolute error.

The results of the exploratory and confirmatory principal component analyses showed that the OHS data can be associated with three constructs relating to pain, mobility and self care. Furthermore it was shown that the "anxiety / depression" domain of the EQ-5D was not associated with any of the 3 constructs of the OHS, and had a maximum correlation of 0.3 to any of the OHS items. The differences between the OHS and the EQ-5D do not impede the merits of either instrument when used for their own purposes. However, even though estimating a mapping model via regression is quite straightforward, the conceptual differences between the two instruments restrict the applicability of the obtained mapping models. The differences in the underlying constructs between the two instruments hampers the accurate predictions of individual EQ-5D patient scores based on their OHS scores.

A geometric approach to health state modelling

The value of a health state is typically described relative to the value of an optimal state, specifically as a ratio ranging from unity (equal to optimal health) to minus infinity. Incorporating potentially infinite values in the calculation of a mean value is a challenging issue in the econometrics of health valuation. In this study, we apply a geometric approach using directional statistics. Unlike ratio statistics, directional statistics are based on polar coordinates (angle, radius). In the case of time trade-off (TTO) values, the range of angles is bounded between 45 degrees (unity) and minus 90 degrees (i.e., minus infinity); therefore, mean angles are well behaved and negate the impetus behind arbitrary data manipulations. Using TTO responses from the seminal Measurement and Valuation of Health (MVH) study, we estimate 243 EQ-5D health state values by minimizing circular variance with and without radial weights.

For states with published values greater than zero (i.e., better-than-death), the radially weighted estimates are nearly identical to the published values (Mean Absolute Difference 0.07; Lin's rho 0.94). For worse-than-death states, the estimates are substantially lower than the published values (Mean Absolute Difference 0.186; Lin's rho 0.576). For the worst EQ-5D state (33333), the published value (using arbitrary transformation of the data) is –0.59 and the directional estimate (using untransformed data) is –1.11.

By taking a geometric approach, we circumvent problems inherent to ratio statistics and the systematic bias introduced by arbitrary data manipulations.

Discrete choice modelling of health states

Probability models have been developed to establish the relative merit of subjective phenomena by means of specific judgmental tasks involving discrete choices. The attractiveness of these discrete choice (DC) models, is that they are embedded in a strong theoretical measurement framework and are based on relatively simple judgmental tasks. The aim of our study was to determine whether DCE derived values are comparable to those obtained using other valuation techniques, in particular the time trade-off (TTO). 209 students completed several tasks in which we collected DC, rank, VAS and TTO responses. DC data were also collected in a general population sample (N = 444). The DC experiment was designed using a Bayesian approach and involved 60 choices between two health states and comparison of all health states to death. The DC data were analysed using a conditional logit model. To relate DC derived values to the QALY scale, we applied and compared three different anchoring approaches. Although modelled DC data broadly replicated the pattern found in TTO responses, the DC dataset consistently produced higher values. The three methods for anchoring DC derived values onto the QALY scale produced similar results. On the basis of the high level of comparability between DC-derived values and TTO values, future valuation studies based on a combination of these two techniques may be considered.

Conclusions and discussion

In this thesis a number of studies are described all related to economic evaluations. As described in the introduction the mathematical techniques used in these studies originate from different disciplines but are all applicable in the development of the economic models for the assessment of cost-effectiveness of health care interventions.

Sources of uncertainty in CUA outcomes

The first methodological study addressed the impact of four different methods of meta-analysis on the outcomes (i.e. cost-effectiveness) of a probabilistic Markov model for Chronic Obstructive Pulmonary Disease. We've shown that the choice for a particular meta-analysis technique has an impact on the uncertainty of the outcomes of an economic model. In particular we found that a Bayesian random effects model results in more uncertainty and a lower acceptability curve than fixed effect models and frequentist models.

In the models used in economic evaluations it is common to include not only the mean values of the model parameters, but also the uncertainty of the model parameters in a probabilistic sensitivity analysis. When 30 states and 300 observations per state are used for TTO based data of the EQ-5D, the expected uncertainty around the predicted utility values will be around 0.01, which is smaller than the typical patient variability in health status. This implies that in such a case the uncertainty related to the exact value of the utilities can safely be left out of a probabilistic sensitivity analyses.

However, if mapping between a disease specific instrument and a utility instrument is applied to obtain the utilities, then an additional level of uncertainty is introduced. This additional uncertainty can become so large that it is no longer negligible compared to the patient variability in health status and should be included in the PSA of an economic model. As we've shown in the case of total hip replacement, there are marked differences in content between the disease specific OHS, and the generic EQ-5D. This led us to conclude that using a mapping model to accurately predict individual patients utility values based on the OHS was not feasible.

New approaches in modelling health states

We've shown that when estimating ratio's such as TTO values, a geometrical approach has advantages over an algebraic approach because of the removal of the singular behaviour that occurs when the denominator is (close to or) equal to zero, while the model structure itself is the same (i.e. the same utility model $U = \sum \beta_{ij} * x_{ij} + \varepsilon$ is estimated in both cases).

The pilot study assessing the feasibility of using DC to model EQ-5D utilities showed promising results. It was found that DC values broadly replicated the pattern found in TTO responses, although the DC values were consistently slightly higher than TTO values. The main difficulty in applying DC models was that these models generated values on an arbitrary scale, not on the metric of the quality (of life) component of the QALY scale. This means that DC-based values need to be anchored on the utility scale, where full health has a value of 1 and death has a value of 0.

Conclusion

Borrowing mathematical techniques that were developed in other disciplines can be very useful and rather straightforward to implement from a mathematical point of view, but one needs to think carefully about whether it is conceptually appropriate to do so.

Samenvatting

Inleiding

Gezondheidseconomie (GE) is een multidisciplinair veld. Dit is vooral zichtbaar in economische evaluaties (EE) zoals kosten-utiliteit analyses (KUA) die een integraal onderdeel zijn geworden in het beheer van de gezondheidszorg in veel westerse landen. De wiskundige technieken die worden gebruikt bij het verkrijgen en beschrijven van de informatie die nodig is voor EEs zijn voornamelijk afkomstig uit drie verschillende wiskundige disciplines verbonden met economie, psychologie en geneeskunde: econometrie, psychometrie en (bio)statistiek. Dit proefschrift laat zien hoe ideeën en benaderingen uit verschillende disciplines kunnen worden toegepast bij het oplossen van gezondheids-economische problemen.

De studies beschreven in dit proefschrift laten zien hoe meer gespecialiseerde technieken en benaderingen kunnen worden gebruikt buiten het gebied waar ze oorspronkelijk werden ontwikkeld. In het bijzonder worden zij gebruikt in het onderzoek naar bronnen van onzekerheid in KUA en in de meting en waardering van gezondheidsgerelateerde kwaliteit van leven. De gebruikte technieken zijn Monte Carlo simulatie, Factor Analyse en Discrete Keuze Modellering. De benaderingen omvatten frequentistische en Bayesiaanse statistiek en een regressiemodel gebaseerd op de geometrie (dwz directionele statistiek).

De eerste twee studies in dit proefschrift geven een illustratie van veel gebruikte concepten in de gezondheidszorg economisch onderzoek: gezondheidsgerelateerde kwaliteit van leven (KvL) en economische evaluaties. De volgende drie studies hebben betrekking op bronnen van onzekerheid in de uitkomsten van een KUA. De laatste twee studies introduceren twee nieuwe benaderingen in het modellen van waarderingen van gezondheidstoestanden.

Leeftijdsafhankelijkheid van zelf-gerapporteerde gezondheid in Europa

Om de relatie tussen zelf-beoordeelde KvL en leeftijd gemeten met EQ-5D te onderzoeken, hebben we bevolkinggegevens geanalyseerd uit 10 Europese landen. We schatten verschillende regressiemodellen waar de EQ-VAS waarden de afhankelijke variabele waren en vonden dat een eenvoudig lineair model beter functioneerde dan meer complexe modellen. Het lineaire model resulteerde in een constante = 93.12 en leeftijdsparameter = –0.34 met adjusted $R^2 = 0.106$. Een model met de gemiddelde waargenomen EQ-VAS scores per leeftijd in plaats van de onbewerkte data (dwz. de within variantie is verwijderd), resulteerde in een dramatische toename van R^2. De waarde van R^2 stijgt van 10.7% tot 93.9%, terwijl de regressies coëfficiënten hetzelfde bleven. We vonden dat het verschil tussen de landen groter was dan het verschil tussen mannen en vrouwen.

Modellering van het aantal gerapporteerde problemen is uitgevoerd met logistische regressie. Weer de waren de waarden van R^2 laag (variërend van 23% tot 1%). Duidelijke verschillen werden gevonden tussen de relatie met de leeftijd van de fysieke domeinen mobiliteit, zelfzorg en dagelijkse activiteiten, en de domeinen pijn/ongemak en angst/depressie. De fysieke domeinen beginnen met weinig gemelde problemen (<5%) op de leeftijd van 18 jaar, die vervolgens toenemen met de leeftijd in toenemende mate. Pijn/ongemak en angst/depressie beginnen hoger (18% en 19%), maar nemen toe met leeftijd met een constante snelheid die hoger is voor pijn/ongemak dan angst/depressie.

Economische evaluatie van panitumumab bij uitgezaaide darmkanker

Colorectale kanker is de vierde meest gediagnosticeerde vorm van kanker. Onlangs zijn er nieuwe medicijnen beschikbaar gekomen om patiënten met uitgezaaide colorectaalkanker (mCRC) te behandelen. Panitumumab is een van die nieuwe geneesmiddelen. Het doel van deze studie was om de kosteneffectiviteit van panitumumab te beoordelen als monotherapie bij mCRC na het falen van andere chemotherapie met als doel het in aanmerking komen voor tijdelijke vergoeding in Nederland.

Een micro-simulatie Markov-model is ontwikkeld om de kosteneffectiviteit van panitumumab te modelleren. Het model bevat drie ziektetoestanden (mCRC, progressieve ziekte en overlijden) met tijdsafhankelijke overgangskansen tussen deze toestanden. Gegevens van een klinische trial waarin panitumumab plus Best Supportive Care (BSC) vergeleken werd met BSC alleen werden gebruikt om de modelparameters te schatten. Het model resulteerde in een geschatte gemiddelde incrementele kosteneffectiviteit ratio (ICER) van € 64,321 / QALY (95% CI [52,642; 94,187]). Ongeveer de helft van de patiënten in de studie heeft een niet-gemuteerd (wild-type) KRAS gen, hetgeen indicatief is voor reactie op panitumumab. Onze resultaten voor deze subgroep van patiënten resulteerde in een geschatte gemiddelde ICER van € 64,541 / QALY (95% CI [52,136; 90,706]).

De uiteindelijke beslissing over het al dan niet in aanmerking komen voor langdurige vergoeding van panitumumab in Nederland zal worden gebaseerd op data van een observationele studie. In deze observationele studie worden de kosten en effecten bepaald van panitumumab in de dagelijkse klinische praktijk, in plaats van in de “laboratorium setting” van een gerandomiseerde klinische trial. Deze data zullen gebruikt worden om opnieuw de parameters van het Markov model te schatten, wat uiteindelijk resulteert in een economische evaluatie voor panitumumab in de dagelijkse klinische praktijk.

Impact van de methode van datasynthese op de resultaten van kosten-utiliteits analyses

Kosteneffectiviteitsmodellen moet altijd aangepast kunnen worden als er nieuwe gegevens over belangrijke modelparameters beschikbaar komen. Echter, er bestaan verschillende methoden voor het combineren van data en de gekozen methode kan de kosteneffectiviteit schattingen beinvloeden. Het doel van deze studie was om het effect van de verschillende methoden van meta-analyse te onderzoeken op de uiteindelijke schattingen van de kosten-effectiviteit van een probabilistisch Markov model voor Chronic Obstructive Pulmonary Disease (COPD). We vergeleken vier verschillende methoden om gegevens voor de parameters van een kosten-effectiviteit model voor COPD te synthetiseren: frequentistische en Bayesiaanse fixed-effects (FE) en random-effects (RE) meta-analyses. Deze methoden werden toegepast om nieuwe overgangskansen tussen de ziekte toestanden uit het model te verkrijgen en om nieuwe schattingen te verkrijgen voor de kansen op een exacerbatie en de zwaarte van de exacerbaties.

De vier methoden hebben geleid tot verschillende schattingen van de kansen en hun standaardfouten (SE). De effecten van het gebruik van verschillende synthese technieken waren het meest prominent in de schatting van de standaardfouten. We vonden verschillen in de SE's van de exacerbatiekansen van een factor 9. De SE's voor de overgangskansen verschilden tot bijna een factor 3. In onze studie vonden we dat het frequentistische FE-model resulteerde in de laagste gemiddelen en SE's , terwijl het Bayesiaanse RE-model resulteerde in de hoogste. De schattingen van verschillen tussen de behandelingen in de totale kosten, QALYs en acceptability curves (CEAC) varieerde ook afhankelijk van de gekozen methode voor datasynthese. Met een Bayesiaanse RE model was de CEAC 15% lager dan met een frequentistische FE model.

Deze resultaten tonen aan dat de keuze van de synthese methode de resulterende model-parameters aanzienlijk kan beïnvloeden, die op hun beurt weer invloed hebben op de schattingen van kosteneffectiviteit en de bijbehorende onzekerheid.

Statistische onzekerheid in utiliteiten gemeten met TTO

De utiliteiten voor gezondheidstoestanden die gebruikt worden in kosten utiliteit analyses worden veelal als puntschattingen gebruikt, waardoor geimpliceerd wordt dat deze perfect bekend zijn. Aangezien utiliteiten gebaseerd zijn op empirische waarderingsstudies, kunnen ze echter niet volledig bekend zijn. We wilden de onzekerheid rond utiliteiten kwantificeren en de relatie bekijken tussen het aantal gezondheidstoestanden en respondenten opgenomen in de waarderingstudies.

We analyseerden de onzekerheid van twee van de EQ-5D waarderingstudies: de Nederlandse time trade-off (TTO) studie en de Engelse MVH studie. We gebruikten ANOVA om de toepasbaarheid van de twee utiliteits modellen en de onzekerheid van de geschatte utiliteit voor elke gezondheidstoestand te beoordelen. De impact van het aantal respondenten en gezondheidstoestanden op de onzekerheid werd onderzocht met behulp van Monte Carlo simulatie. We simuleerde studies die tussen de 14 en 42 gezondheidstoestanden bevatten met een steekproefomvang variërend van n = 25 tot n = 1000. De standaardfouten van de utiliteiten in de Nederlandse studie varieerde 0.044 tot 0.010 voor de verschillende gezondheidstoestanden, of 0.028 gemiddeld. In de Engelse studie varieerden die van 0.013 tot 0.006 (0.009 gemiddeld). Onze Monte Carlo resultaten toonden aan dat de gemiddelde standaard fouten varieerden van 0.095 bij 17 toestanden, elk gewaardeerd door 25 respondenten tot 0.009 bij 42 gezondheidstoestanden elk gewaardeerd door 1000 respondenten.

Op basis van onze resultaten stellen wij voor dat, om een gemiddelde standaardfout van de utiliteiten te verkrijgen lager dan 0.01 voor EQ-5D TTO waarderingsstudies, men ongeveer 30 gezondheidstoestanden moet laten waarderen elk door ongeveer 300 respondenten.

Vergelijking van de EQ-5D en de Oxford Hip Score

Zowel ziekte-specifieke als generieke patiënt-gerapporteerde-uitkomstmaten (PROM's) geven informatie over de gezondheidstoestand van de patiënt. In het algemeen geven ziekte-specifieke instrumenten meer klinische informatie dan generieke instrumenten, maar geven zij geen utiliteiten. Het doel van deze studie was om de vergelijkbaarheid van een ziekte-specifiek instrument, de Oxford Hip Score (OHS), en een generiek instrument, de EQ-5D, te beoordelen, en om de mogelijkheid tot het verkrijgen van utiliteiten voor de OHS via een mapping te ondrzoeken. Gegevens van 439 NHS patiënten in Engeland (voor en zes maanden na een heupoperatie) werden geanalyseerd. De informatie verzameld met de OHS en EQ-5D werd beoordeeld met behulp van principale componenten analyse en de analyse van de correlatie matrix.

Uit de resultaten van de explorerende en de bevestigende principale componenten analyse bleek dat de OHS kan worden geassocieerd met drie constructen: pijn, mobiliteit en zelfzorg. Verder werd aangetoond dat het "angst/depressie" domein van de EQ-5D niet was geassocieerd met een van de drie constructen van de OHS. Het had een maximale correlatie van 0.3 met de items van de OHS. De verschillen tussen de OHS en de EQ-5D vormen geen belemmering voor de toepasbaarheid van beide instrumenten voor hun eigen doeleinden. Echter, hoewel het schatten van een mapping model via regressie vrij eenvoudig is, beperken de conceptuele verschillen tussen de twee instrumenten de toepasbaarheid van de verkregen mapping modellen. De verschillen in de onderliggende constructen tussen de twee instrumenten belemmert de nauwkeurige voorspellingen van de individuele EQ-5D patiënt scores op basis van hun OHS scores.

Een geometrische benadering voor het modelleren van gezondheidstoestanden

De waarde van een gezondheidstoestand is meestal ten opzichte van de waarde van een optimale toestand beschreven, specifiek als een ratio die loopt van één (gelijk aan optimale gezondheid) tot min oneindig. Het opnemen van potentieel oneindige waarden in de berekening van een gemiddelde waarde is een uitdagend probleem in de econometrie van gezondheidswaardering. In deze studie hanteerden wij een geometrische benadering met behulp van directionele statistiek. In tegenstelling tot conventionele statistiek, is directionele statistiek gebaseerd op poolcoördinaten (hoek, radius). In het geval van time trade-off waarden wordt het bereik van de hoeken begrensd tussen 45 graden (één) en min 90 graden (dat wil zeggen, min oneindig). Dit betekent dat de gemiddelden van hoeken zich goed gedragen waardoor willekeurige data manipulaties niet nodig zijn. Met behulp van TTO antwoorden van de Engelse MVH studie, schatten wij de utiliteiten voor de 243 EQ-5D gezondheidstoestand door het minimaliseren van de circular variance met en zonder radiale gewichten.

Voor toestanden met waarden groter dan nul (dat wil zeggen, beter dan de dood), zijn de radiaal gewogen schattingen vrijwel identiek aan de gepubliceerde waarden (gemiddelde absolute verschil 0,07; Lin's rho 0,94). Voor toestanden slechter dan dood zijn de schattingen zijn aanzienlijk lager dan de gepubliceerde waarden (gemiddelde absolute verschil 0.186; Lin's rho 0.576). Voor de ernstigste EQ-5D toestand (33333) is de gepubliceerde waarde (met behulp van willekeurige transformatie van de gegevens) −0.59 en de directionele schatting (met niet-getransformeerde gegevens) −1,11.

Door het gerbuiken van een geometrische benadering omzeilen we de problemen die inherent zijn aan de statistiek van ratios en de daarbij horende systematische bias als gevolg van willekeurige data manipulaties.

Discrete keuze modellering van gezondheidstoestanden

Waarschijnlijkheidsmodellen worden gebruikt om de relatieve waarden van subjectieve verschijnselen vast te stellen door middel van specifieke beoordelingstaken op basis van discrete keuzes. De aantrekkelijkheid van deze discrete keuze (DC) modellen is dat ze zijn ingebed in een sterk theoretische raamwerk en gebruik maken van relatief eenvoudige beoordelingstaken. Het doel van deze studie was om te bepalen of waarden verkregen met een DC experiment vergelijkbaar zijn met die verkregen met andere waarderingstechnieken, zoals de TTO. 209 studenten voltooide diverse taken waarin DC, rangordes, VAS en TTO antwoorden werden verzameld. DC data werden ook verzameld in een steekproef uit de algemene bevolking (N = 444). Het DC experiment werd ontworpen met behulp van een Bayesiaanse benadering en bestond uit 60 keuzes tussen twee gezondheidstoestanden en de vergelijking van alle gezondheidstoestanden met "dood". De DC data werden geanalyseerd met een conditional logit model. Om DC waarden te plaatsen op de QALY schaal hebben we drie verschillende benaderingen voor verankering vergeleken. Hoewel het DC model in grote lijnen het patroon van de TTO repliceerde, produceerde de DC dataset consistent hogere waarden. De drie methoden voor het verankeren van de DC waarden op de QALY schaal leidden tot vergelijkbare resultaten. Op basis van de hoge mate van vergelijkbaarheid tussen DC afgeleide waarden en TTO waarden kan worden overwogen om toekomstige waarderingsstudies te baseren op een combinatie van deze twee technieken.

Conclusies en discussie

In dit proefschrift werd een aantal studies beschreven die allemaal gerelateerd zijn aan economische evaluaties. Zoals beschreven in de inleiding, zijn de wiskundige die zijn technieken gebruikt in deze studies afkomstig uit verschillende disciplines, maar allemaal toepasbaar bij het ontwikkelen van de economische modellen voor de beoordeling van kosteneffectiviteit van interventies in de gezondheidszorg.

Bronnen van onzekerheid in CUA uitkomsten

In de eerste methodologische studie werd ingegaan op de impact van de vier verschillende methoden van meta-analyse op de resultaten (dwz kosteneffectiviteit) van een probabilistische Markov model voor COPD. We hebben aangetoond dat de keuze voor een bepaalde meta-analyse techniek een impact heeft op de onzekerheid van de uitkomsten van een economisch model. In het bijzonder vonden wij dat een Bayesiaanse random-effects model resulteert in meer onzekerheid en een lagere acceptability curve dan fixed-effects modellen en frequentistische modellen.

In de modellen gebruikt voor economische evaluaties is het gebruikelijk om niet alleen de gemiddelde waarden van de modelparameters, maar ook de onzekerheid van de modelparameters in een probabilistische gevoeligheidsanalyse mee te modelleren. Als 30 gezondheidstoestanden en 300 waarnemingen per toestand worden gebruikt in een EQ-5D TTO studie, wordt de verwachte onzekerheid over de voorspelde utiliteitswaarden 0.01, kleiner dan de typische variabiliteit in gezondheidstoestand tussen patienten in een klinische studie. Dit betekent dat in dat geval de onzekerheid met betrekking tot de exacte waarde van de utiliteiten betrekkelijk veilig kan worden weggelaten uit een probabilistische gevoeligheidsanalyses (PSA).

Echter, indien mapping tussen een ziekte specifiek instrument en een generiek instrument wordt toegepast om utiliteiten te verkrijgen, wordt een extra niveau van onzekerheid rond de utiliteiten geïntroduceerd. Deze extra onzekerheid kan zo groot zijn dat het niet langer verwaarloosbaar is ten opzichte van de variabiliteit in gezondheidstoestand tussen patiënten en moet worden meegenomen in de PSA van een economisch model. Zoals we hebben laten zien in het geval van heup vervanging, zijn er duidelijke verschillen in inhoud tussen de ziekte specifieke OHS, en de generieke EQ-5D. Dit leidde ons tot de conclusie dat het gebruik van een mapping model gebaseerd op de OHS om nauwkeurig de utiliteiten voor individuele patiënten te bepalen niet haalbaar was.

Nieuwe benaderingen in het modelleren van utiliteiten

We hebben aangetoond dat bij het schatten van ratio's, zoals TTO waarden, een geometrische benadering voordelen heeft boven een algebraïsche benadering, door het verwijderen van het singuliere gedrag wanneer de noemer richting de nul gaat, terwijl de modelstructuur zelf hetzelfde is (dwz het zelfde utiliteitsmodel $U = \Sigma\ \beta ij * xij + \varepsilon$ wordt geschat in beide gevallen).

De pilot studie naar de haalbaarheid van het gebruik van DC om EQ-5D utiliteiten te modelleren toonde veelbelovende resultaten. Het bleek dat, in het algemeen, DC het patroon van TTO repliceerde, hoewel de DC waarden steeds iets hoger waren dan TTO waarden. Het grootste probleem bij de toepassing van DC modellen was dat deze modellen waarden genereren op een willekeurige schaal en niet op de kwaliteit (van leven) component van de QALY schaal. Dit betekent dat het DC model moeten worden verankerd op de utiliteitsschaal, waarbij volledige gezondheid een waarde heeft van 1 en dood een waarde heeft van 0.

Conclusie

Het lenen wiskundige technieken die ontwikkeld zijn in andere disciplines kan erg handig zijn en (vanuit wiskundig perspectief) relatief eenvoudig te implementeren, maar men moet er goed over na denken of het conceptueel juist is om dit te doen.

Dankwoord

Frank, ik weet nog goed dat ik half December 2001 bij je langskwam bij het cGBR voor een introductiegesprek (ik wist überhaupt niet dat er zoiets als gezondheidseconomie bestond), en dat ik een uur later al bij je aan het werk was. In de jaren erna heb je me altijd met veel geduld en kennis bijgestaan om de stap van beta-wetenschap naar sociale-wetenschap succesvol te kunnen maken. Uiteindelijk is het je zelfs nog gelukt om me iets van commercieel gevoel bij te brengen, wie had dat ooit gedacht! Ik ben je zeer dankbaar voor de afgelopen 12 jaar. Zonder jou was dit proefschrift nooit totstandgekomen.

Carin, ik ben je heel dankbaar dat je me binnenhaalde bij het iMTA, dat je mijn promotor wilde zijn en dat je zoveel geduld met me hebt gehad bij de totstandkoming van dit proefschrift. De eerste keer dat ik met een opzet voor mijn proefschrift bij je langs kwam had ik de de focus gelegd op value of information analyses. Het uiteindelijke resultaat gaat over iets compleet anders, wat laat zien hoeveel vrijheid je me altijd hebt gegeven om mijn ideeen uit te werken en mijn eigen lijn te kiezen, waarvoor mijn dank.

De leden van de promotiecomissie dank ik hartelijk voor het beoordelen van mijn proefschrift en het opponeren bij de verdediging daarvan. Ook gaat mijn dank uit naar de co-auteurs van de artikelen waaraan ik meewerkte binnen en buiten dit proefschrift.

Zonder een goede werksfeer met fijne collega's en stimulerende gesprekken zou ik dit proefschrift nooit hebben kunnen voltooien. Ik wil mijn collega's van het iMTA en de EuroQol Executive Office hier heel hartelijk voor bedanken, alsook de leden van de EuroQol Group en niet te vergeten Ben Tomee, zonder wiens coaching de hobbels op de weg bergen zouden zijn gebleven.

En dan de paranimfen: Siok, ik kan me geen betere kamergenoot bedenken dan jij. Ik ben je erg dankbaar dat je altijd voor me klaarstaat ook nu we niet meer bij hetzelfde instituut werken. Milica, je bent m'n grote zus en de beste die een broertje zich kan wensen. Ik ben heel blij dat je als paranimf deze dag samen met me kan beleven.

Als laatste wil ik graag mijn ouders bedanken: Siem, je stond en staat altijd voor me klaar om over mijn ideeen (en frustaties) wat betreft psychometrie, econometrie, statistiek en in het algemeen over de verschillen in aanpak bij natuur- en sterrenkunde en sociale wetenschap te sparren. De kennis die ik van jou heb opgepikt had ik nooit uit een boek kunnen halen. Ik vind het geweldig dat je co-auteur bent van een artikel in mijn proefschrift en ik hoop dat het niet bij dit ene artikel blijft. Wette, ookal ben je geen co-auteur heb ook jij je stempel op dit proefschrift gedrukt. Door je cynisme verdwenen mijn frustaties als sneeuw voor de zon en kon ik er altijd de humor van inzien en er mijn voordeel mee doen. Paps, mams (en grote zus), bedankt voor alle onvoorwaardelijke steun en liefde die jullie me mijn hele leven hebben gegeven.

Curriculum vitae

After obtaining his masters degree in astrophysics in 2001 at Utrecht University, Mark started to work in health economics as a researcher for the Centre for Health Policy and Law of Erasmus University Rotterdam. During this time he carried out various projects on End Stage Renal Disease and on Quality of Life measurement (the latter on behalf of the EuroQol Group Executive Office). In 2005 he joined the institute for Medical Technology Assessment of Erasmus University Rotterdam. The focus of his work at iMTA was on quantitative research, in particular on probabilistic Markov modelling, meta-analyses techniques and mathematics. While at iMTA about 40% of his research projects were carried out on behalf of the EuroQol Group Executive Office. His work for the EuroQol Group centres on elicitation and modelling techniques to obtain utility values with EQ-5D. In particular he was involved in the management and development of the valuation protocol and software for the valuation of the EQ-5D-5L. Since November 2012 Mark has a full time position as senior researcher at the EuroQol Group Executive Office.

Selected Publications

Oppe M, de Charro F Th. Population norms and their uses. In: Szende A, Williams A (Eds). Measuring self-reported population health: An international perspective based on EQ-5D. SpringMed Publishing 2004.

Oppe M, Weijnen TJG, de Charro FT. Development of a questionnaire to assess the quality of care in Dutch dialysis centers from the patient's perspective. Expert Rev Pharmacoecon Outcomes Res 2005;5:255-265.

Oppe M, Barendregt W, Treur MJ. Statistical report 2007. The Netherlands: Renal Registry Renine; 2008.

Szende A, Oppe M, Devlin N, eds. EQ-5D value sets: inventory, comparative review and user guide. Dordrecht: Springer; 2007.

Oostenbrink JB, Al MJ, Oppe M, Rutten-van Mölken MPMH. Expected value of perfect information: an empirical example of reducing decision uncertainty by conducting additional research. Value Health 2008;11(7):1080-1090.

Tan SS, Oppe M, Zoet-Nugteren SK, Niezen RA, Kofflard MJ, Ten Cate FJ, Roijen LH. A microcosting study of diagnostic tests for the detection of coronary artery disease in the Netherlands. Eur J Radiol. 2009;72(1):98-103.

Craig BM, Oppe M. From a different angle: a novel approach to health valuation. Soc Sci Med. 2010 Jan;70(2):169-74.

Stolk EA, Oppe M, Scalone L, Krabbe PF. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. Value Health. 2010;13(8):1005-13.

Oppe M, Devlin N, Black N. Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. Value Health. 2011;14(6):884-91.

Heijink R, van Baal P, Oppe M, Koolman X, Westert G. Decomposing cross-country differences in quality adjusted life expectancy: the impact of value sets. Popul Health Metr. 2011;9(1):17.

Oppe M, Al M, Rutten-van Mölken M. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics. 2011;29(3):239-50.

Attema AE, Versteegh MM, Oppe M, Brouwer WB, Stolk EA. Lead time TTO: leading to better health state valuations? Health Econ. 2013;22(4):376-92.

Xie F, Pullenayegum E, Gaebel K, Oppe M, Krabbe PF. Eliciting preferences to the EQ-5D-5L health states: discrete choice experiment or multiprofile case of best-worst scaling? Eur J Health Econ. 2013;