Quality of life in economic evaluations of health

Matthijs Versteegh

Quality Of Life In Economic Evaluations Of Health

ISBN: 978-94-6108-626-6

Layout and printed by: Gildeprint - Enschede
Cover design: Sjaak van der Vooren - www.sjaakvandervooren.nl

Quality Of Life In Economic Evaluations Of Health

Kwaliteit van leven in economische evaluaties van gezondheid

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

Donderdag 10 april 2014 om 15:30 uur

door
Matthijs Michaël Versteegh
geboren te Utrecht

Promotiecommissie

Promotor:
Prof.dr. W.B.F. Brouwer

Overige leden:
Prof.dr. J.J. van Busschbach
Prof.dr. C.D. Dirksen
Prof.dr. C.A. Uyl-de Groot

Copromotor:
Dr. E.A. Stolk

Table of Contents

1

Introduction

Health care expenditures have increased rapidly over the last decades in the Netherlands, in absolute terms and as percentage of gross domestic product. Curbing the rising health care costs has proven to be a very sensitive and complex societal issue. An important driver of rising costs is the availability and use of new and expensive medical technologies, causing a greater number of patients to be treated with more expensive interventions (1). Typically, these interventions do not only increase expenditures but also contribute to societal health and well-being. Since health care expenditures are high on the political agenda, policy makers are interested in the relative effectiveness and efficiency of new medical interventions: do they achieve larger health effects than other treatments, and if so, at what additional costs? Economic evaluations address this question. In economic evaluations, costs and effects of alternative medical interventions are compared, to see whether the new intervention offers good 'value for money' (2). Researchers that wish to apply economic evaluations to health care interventions face several methodological challenges. This thesis deals with one particular challenge: the measurement of the benefit of health care interventions in terms of quality of life.

Performing economic evaluations of medical interventions requires some form of quantification of the (health) effects, and this poses an important challenge. The three most common types of economic evaluation: cost-benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA) differ in how they measure and value health effects. In CBA health effects are quantified in monetary terms, for instance using willingness to pay techniques (3). A downside of CBA is that there may be "a strong resistance, both within the medical community and amongst the general public towards attaching a monetary value to the health (and life) of a person" (4). Moreover, expressing the value of health gains in monetary terms using willingness to pay techniques is not without methodological challenges (5). In CEA, health effects are quantified in natural (clinical) units, such as life years gained, hip fractures avoided or percentage blood pressure lowered. A downside of CEA is that only the efficiency of treatments which meaningfully use the same outcome measure can be compared. Due to the shortcomings of CBA and CEA, the third type of economic evaluation, the so-called cost-utility analysis (CUA), gained popularity. In CUA, health effects are quantified in terms of Quality Adjusted Life Years (QALYs). The QALY is a measure of effect, which combines length of life (the number of years a patient survives) and the health-related quality of life in those years. Health-related quality of life (HRQoL) captures those elements of broader quality of life, utility or wellbeing that relate to the health domain (6).

Obviously, for CUA to be a legitimate policy tool, its constituent elements need to be measured in a methodologically sound manner. Methodological issues of CUAs relate to many different aspects of this type of economic evaluations, ranging from perspective chosen (7, 8), costing methodology (9), handling uncertainty (10), discounting (11) to measuring and valuing health

effects in terms of QALYs (12). In this thesis, the focus is on the latter issue with the overarching research question: *How can the measurement and valuation of health related quality of life for QALY computation in economic evaluations be improved?*

Before specifying the exact focus of this thesis and the research questions it addresses, first the goal and methodology of economic evaluations in the health care sector will be introduced to provide the context in which this research was performed.

1.1 Resource allocation with CUA

Economic evaluations aim to support welfare improving decisions (13). In their traditional form, i.e., in a classical CBA, policy change is considered welfare improving if the benefits of the intervention exceed the costs of that intervention, compared to some alternative. Applied to health care interventions, economic evaluations generally aim to support decisions on the collective reimbursement of new interventions. A new intervention may replace an old one, and this is welfare improving (i.e. eligible for collective reimbursement) if the additional benefits due to the new intervention outweigh the additional costs of that intervention (compared to the old one). More formally, the decision rule for adopting an intervention becomes:

$$\Delta B - \Delta C > 0 \qquad (1)$$

In equation (1) ΔB denotes the difference in benefits between the old and the new intervention and ΔC denotes the difference in costs. The decision rule thus simply describes that the incremental benefits (ΔB) should exceed the incremental costs (ΔC). In a CBA, the benefits are normally expressed in terms of money, making a direct comparison between costs and benefits possible. In a CUA, such a direct monetization of health effects does not occur and thus CUA requires an alternative quantification of benefit. Indeed, without a sensible (and uniform) quantification of the effect of health care interventions, any societal intention to optimize the effect of health care interventions would be meaningless[1]. As written by Dolan (2001):

> "If governments are going to deploy the resources at their disposal where they will be of greatest benefit, then information on the benefits associated with alternative allocations is required. This raises questions about how benefit is to be defined and measured (…)"(14).

Like CBAs, CUAs of medical interventions are aimed at aiding decision makers in selecting those health care interventions for collective reimbursement which yield most benefit to society, relative to their costs. This is done by separating benefits into quantities (health effects

[1] This is an adaptation of Friedman & Savage (1952): "The ethical precept that society "should" promote the "welfare" of individuals is meaningless until "welfare" is given content." (Quotation marks in original).

expressed in terms of QALYs) and the monetary value of these QALYs:

$$\Delta B = v\ \Delta Q \tag{2}$$

where v is the monetary value of one unit of health (here expressed as QALYs), while ΔQ are the incremental QALY gains. Combining equations (1) and (2) results in the decision rule:

$$v\ \Delta Q - \Delta C > 0 \tag{1'}$$

which can be rewritten as:

$$\frac{\Delta C}{\Delta Q} < v \tag{1''}$$

Hence, the common outcome of a CUA, the incremental cost-effectiveness ratio (ICER), expressing the incremental costs per QALY gained, should be lower than the value per QALY*v*. In other words, the costs per QALY should not exceed what society is willing to pay to gain a QALY. The merit of such an approach is that a range of medical interventions can be compared in terms of costs and effects, since QALY gains are measurable in many different diseases. Ultimately, such an approach intends to inform decision making in such a way that resources can be allocated to those interventions that offer good value for money, i.e., are welfare improving.

Traditional CUA has been described as "the primary method for evaluating health policy under a utilitarian ethic" (15). As can be inferred from the decision rule introduced above, choosing health programs with the most favorable ICERs results in maximization of health benefits (QALYs) from a given budget. The utilitarian principle of maximizing total utility with a given budget has been criticized, however, for reasons of equity, since maximization of QALYs does not consider the distribution of these QALYs over a population (16-19). To illustrate, under a utilitarian maximization principle, an intervention achieving a benefit of 0.1 QALY per patient in ten relatively healthy patients (so that the sum of the gains is 1 QALY) would receive priority over an equally expensive intervention yielding 0.3 QALYs per patient in 3 patients in a relatively poor health state (so that the sum of the benefits is 0.9). While traditional CUAs may thus support the maximization of health benefits, it has been argued to be incomplete as a tool for prioritization in policy making when equity considerations play a role, such the notion to treat those worse off (i.e. the three patients in poor health) with priority.

In order for CUAs to be more directly useful for policy making, it has been argued that the goal of maximization of total health benefits (efficiency), may need to be balanced with the goal of a fair distribution of health and health care (equity); sometimes referred to as the efficiency – equity

tradeoff (20). To account for equity principles, it has been suggested to prioritize treatments for worse-off individuals (21), where 'worse off' can be and has been defined in different ways (22). First steps in more systematically taking into account these equity considerations are currently undertaken, both in the Netherlands (23) and in the UK (24).

One way to allow equity considerations to enter the decision rules underlying CUA, is to allow different QALYs to have different (social) values:

$$\frac{\Delta C}{\Delta Q_i} < v_i \qquad (3)$$

where subscript *i* refers to a specific 'equity class', defined in some meaningful fashion. ΔQ_i therefore now reflects the QALY gain in a group in equity class *i* (e.g. more or less severely ill or younger or older patients). Since societies willingness to pay for a QALY may be different for these different equity classes, v now is equity class specific: v_i. In The Netherlands, such an approach has been proposed in using the outcomes of CUAs in decision making, in an attempt to balance equity and efficiency (23). While there is support for this approach, researchers face several methodological challenges in its operationalization. As a consequence, methodological tools are still under development.

1.2 QALYs

One of the major challenges in performing economic evaluations of health care interventions is to obtain a suitable and valid measure for the benefit of medical interventions. In CUA, the QALY is considered to be such a measure, capturing two central effects of medical treatment: increased survival and improved HRQoL. For QALY computation, patients' HRQol is expressed in an index value (also labeled 'utility', hence the term cost-utility analysis[2]), which commonly is standardized to take the value of 1 in case of perfect health and that of 0 in case of the state dead. Other health states have some value between 0 and 1, lower values indicating worse HRQoL. For states considered to be 'worse than dead' the index value may become negative (26, 27). These index values are commonly used as 'tariffs': an equal value is attached to a health state whomever is in that state (20). This requires some way of uniformly describing health states, which is discussed further below, and a way of valuing these (uniformly described) health states. Obtaining these health state valuations is not straightforward, as will also be highlighted below.

Once health states have quality of values attached to them, and the time people (on average) spent in different health states is known, it is possible to perform QALY calculations. The QALYs

[2] Throughout this thesis, the terms 'quality of life value', 'utility', 'preference value' and 'health state value' are used interchangeably, although it is acknowledged that there differences between these terms that could be meaningful in particular areas of research.

computed for two treatment options (e.g. a new intervention and the currently used alternative), can be compared and the difference in the number of QALYs gained in both treatment options can be derived (28). Figure 1.1 illustrates this for two hypothetical treatments (A and B), where the area under the curves shows that treatment A yields 2.3 more QALYs than treatment B (while leading to 3 years of prolonged survival).

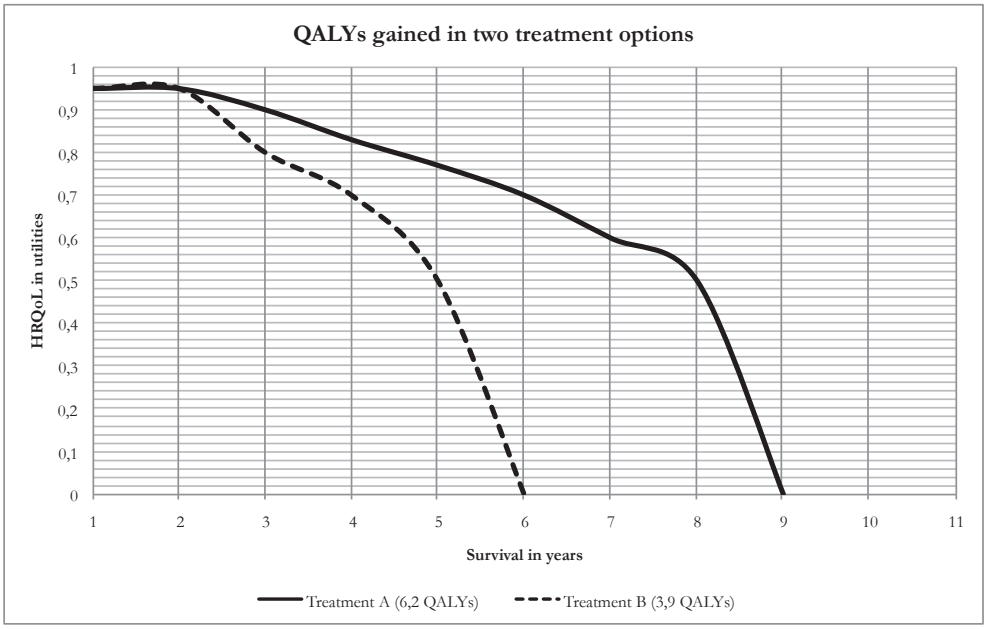

Figure 1.1 Quality and length of life following two hypothetical treatments

1.3 Calculating QALYs

As shown in Figure 1.1, the total benefit in terms of QALYs consists of information on longevity and the quality of life value attached to the different health states people are in after a treatment. While longevity, or survival, is a relatively straightforward concept, the health states people live in during those years as well as the quality of life value attached to them arguably are more challenging to obtain.

A common method for measuring the health state people are in is through standardized questionnaires. In such questionnaires respondents can indicate how they score on different, predefined health domains (e.g., from "I am free of pain" to "I am in severe pain"). A well-known instrument in this field is the EQ-5D (29), which measures health using five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) and three levels

of functioning (no problems, some problems, severe problems). Using such instruments, a health state *h* is determined by the combination of answers on the questionnaire. The EQ-5D, for instance, can distinguish 243 different health states.

Knowing the health states of patients and the time spent in those states is not sufficient to calculate QALYs. The next step is to combine this information with quality of life values for those particular health states. Several often used measurement instruments, such as the EQ-5D, the SF-6D (30) or the HUI III(31), have quality of life values readily available (in a so called 'tariff') for all the health states they are able to describe. This allows straightforwardly attaching quality of life values to all health states captured with the instrument. The quality of life values in these tariffs were commonly obtained in the general public through structured choice tasks. The general public indicated their strength of preference for different health states, as explained further below, from which HRQoL values were derived. Hence, instruments such as the EQ-5D, are often labeled as 'preference-based instruments', because the quality of life values attached to the health states reflect the preferences of people for such states. In sum, there are three distinct phases in QALY calculations with preference-based instruments: describing the health state of a patient with a questionnaire, assigning a preference-based value to that health state, and subsequently multiplying the preference-based value of the health state with the duration of the health state.

Example QALY computation with a preference based instrument

A preference based instrument is used to measure the health states patients are in and has quality of life tariffs available to transform observed health states into quality of life values. Multiplying the health states with the time spent in these health states gives the total QALY profiles. Taking perfect health or the situation without a disease as a comparator, such QALY profiles indicate how much HRQoL is lost due to some illness. In economic evaluations, the QALY gain (or loss) of a new intervention compared to an old intervention is calculated by comparing the QALY profile of patients receiving the new intervention with that of patients receiving the old intervention. An example of computing QALY scores using a preference-based questionnaire is provided in Box 1.

Box 1: Generating average QALY values during 1 year for two individuals

The Health Utilities Index (HUI) mark-III is a preference-based HRQoL questionnaire consisting of 8 dimensions of health, which patients complete themselves. This example simply describes two hypothetical patients who identified their health problems on HUI-III, to indicate the process of health description (measurement) and the transformation into a quality of life value (valuation). Patient A has a sprained ankle following a soccer accident and patient B has a mood disorder. The preference scores for these patients quantify the values attached to the respective health states. Patient A will be in the specified health state during one month, while the condition of B lasts four months. Both are in perfect health the remainder of the year. We will compute the average QALY score for one year for these two patients.

HUI-III items*	Soccer injury	Major depression
Vision	Able to see well	Able to see well
Hearing	Able to hear what is said without a hearing aid	Able to hear what is said without a hearing aid
Speech	Able to be understood completely when speaking	Able to be understood completely when speaking
Ambulation	Full use of hands and 10 fingers	Full use of hands and 10 fingers
Dexterity	Able to walk only short distances with walking equipments	Able to walk around the neighborhood without difficulty
Emotion	Happy and interested in life	So unhappy that life is not worthwhile
Cognition	Able to remember most things	Able to remember most things
Pain	Moderate pain that prevents a few activities	Free of pain and discomfort
Quality of life value of health state	0.56**	0.26**
Yearly average QALY value	1 month duration followed by 11 months perfect health (value 1): 0,96	4 months duration followed by 8 months perfect health (value 1): 0,75

* Item text is shortened and adapted from Feeny et al. (2002) (31)
** This is the quality of life value (with 1 denoting perfect health and 0 denoting 'dead') of the described state.

1.4 Issues in QALY measurement

Despite the widespread use of the QALY metric, important questions remain regarding how to best obtain quality of life estimates, which cover both the description and the valuation of health states. There is little scientific consensus on these issues. Three questions and areas of debate, clearly formulated by Dolan in the 32nd chapter of the *Handbook of Health Economics* (14) are: (i) who should value health, (ii) how is health to be valued and (iii) how should health be described? A fourth, more practical issue, which is gaining attention, is the estimation

of QALY profiles when quality of life has not been measured directly in a study. Commonly, clinical studies measure the effect of interventions with common clinical measures, rather than quality of life measures. Consequently, quality of life is not directly measured and, therefore, directly computing QALYs is impossible. If one, ex post, one still wishes to calculate the QALY gain due to a treatment, the question of how quality of life values can be estimated using other information on patient health arises. These four questions are central in this thesis and they will be highlighted below.

The description of health

The description of health is concerned with identifying which 'dimensions' of health (such as mobility, pain, sensory function and mood) may be affected by an intervention and are considered part of HRQoL. These dimensions can be grouped in a questionnaire and appropriate levels can be assigned to these dimensions that express how well patients perform on those dimensions (e.g. "I am never cheerful" vs. "I am always cheerful").

Dimensions of health can be formulated in rather generic terms, or in more specific terms which may be relevant to specific health conditions (e.g., 'loss of grip strength' may be more important to arthritis than to diabetes). The most popular preference-based instruments (EQ-5D, Health Utilities Index and SF-6D) describe health in generic terms[3]. Therefore, these instruments are applicable to a wide set of conditions and allow direct comparisons across disease areas. However, it has been argued that such generic HRQoL instruments may come at the expense of being insensitive to some disease specific improvements (32, 33) as the description of health is relatively 'crude'. If, for example, a new cancer treatment reduces nausea but not stomach aches compared to the old treatment, patients may experience difficulty in report this 'subtle' yet beneficial effect if asked to generally indicate 'the level of pain/discomfort experienced' (as is the case in EQ5D and HUI-III). To overcome the potential problem of insensitivity of generic instruments, condition-specific (preference-based) instruments have been developed in an attempt to improve health state measurement and, consequently, QALY calculations. These may be more sensitive to the health problems experienced due to some specific disease, but as a consequence reduce comparability across diseases. In this thesis, both generic and condition-specific preference-based measures were studied.

The valuation of health

In order to generate QALY calculations, health state descriptions must be transferred into quantitative HRQoL values with 1 (perfect health) and 0 (dead) as anchors, as indicated above.

[3] It is worth noting that the most common preference-based instruments do not cover the same dimensions of health. For example, the Health Utility Index has an item about 'vision' while EQ-5D and SF-6D do not. See Brazier {{234 Brazier,J. 2007; }} (chapter 8) for an in depth discussion.

In the QALY metric, 'a health state that is more desirable is more valuable', i.e., receives a higher quality of life value (12). The relative desirability of health states is based on preferences for health states, elicited in health state valuation exercises. Because the quality of life values will be used for mathematical operations (it will be multiplied with the duration of a health state), these values need to be expressed on an interval scale. Different types of health state valuation methods exist that meet this requirement, but two popular methods are the Time Trade-Off method (TTO) and the Standard Gamble (SG). In the first method people are asked to indicate their indifference between living a fixed period (e.g. 10 years) in an impaired health state or living shorter in perfect health. In the second, people are asked to indicate their indifference between a certain health state H and a gamble in which one outcome is better than H (e.g. perfect health) and the other is worse (e.g. immediate death). Although both methods are used, empirical evidence suggests they result in (systematically) different preference values (e.g. (34, 35)). Hence, the choice for a particular preference elicitation method is important as it can affect the outcome of CUA. In this thesis, the focus was on the most popular method, the TTO, which has been used to elicit much used tariffs for the EQ-5D, for example in the UK and the Netherlands (29).

The TTO was first developed by Torrance and others (36). In TTO, quality of life values are assigned to health states by asking respondents to indicate their indifference between living shorter in good health or living longer in poor health. A typical exercise will ask respondents to imagine living in an impaired health state for T (often 10) years. Then, the respondent is asked to indicate how many years (x) in perfect health s/he considers to be equivalent to T years in the impaired health state. Obviously, x should be equal to or less than T. Once the point of indifference is given, the value of the health state is simply calculated as x/T. For example, if an individual considers living 10 years with low back pain to be equivalent to living 9 years in full health (i.e. they would trade-off 1 healthy life year to avoid low back pain), the health state value of low back pain is often calculated as 9/10=0.9. Note that this assumes all life years to receive equal weight in decision making, regardless of their timing, which is commonly not the case, due to discounting, see for example the following reference: (37). While the TTO method is popular, it is certainly not without important methodological problems (38-41). One of the issues is that a TTO exercise can also be used to obtain negative quality of life values, that is, for health states considered 'worse than dead'. However, this was commonly done in a separate procedure, not fully identical to the exercise used to obtain positive quality of life values, hence potentially leading to problems of comparability (39, 42, 43). In this thesis, the TTO method is extensively discussed and alternative specifications of the TTO are tested. The topic of discounting is also addressed.

Who should value health states?

An important issue in the context of health state valuations is 'who should value health states?' Different options can be considered: experts like doctors who see many different patients, patients who actually experience a described health state or the general public after having had the health state explained to them. One may consider it a logical choice to ask patients in specific health states to value their own health. While this indeed has been advocated (44), it is relatively uncommon to do so. Often used tariffs (i.e. the set of quality of life values for all health states), such as those for the EQ-5D, were based on health state preferences obtained in the general public. Several explanations for this choice have been put forward (see Chapter 2 of this thesis for a fuller discussion). One important explanation for the use of general public preferences appears to be that patients may adjust to their bad health, hence may value the health states they experience (much) higher than the general public would. Using these high values can result in reduced cost-effectiveness when considering a cure for the underlying ailment. Hence, using public preferences could be seen as 'protecting' patients against undesirable consequences of their adaptation. Whether this reasoning and other reasons for using general public preferences to obtain quality of life values for health states are fully convincing and completely justify current practice, is an important issue which will be addressed in this thesis.

Predicting quality of life values

Although economic evaluations are increasingly common in the context of decisions regarding reimbursement of health interventions, clinical research into the effect of treatments is generally focused on other outcomes than preference-based quality of life. After such clinical research has been performed, there may well be an interest to perform an economic evaluation of the treatment. However, this requires QALY computations, for which the required data then may be lacking. One pragmatic solution to this problem is to *estimate* these quality of life values, for instance using regression techniques, in the literature often referred to as 'mapping' (45). Typically, this approach involves using another dataset in which the statistical relationship between patient characteristics (i.e. scores on clinical variables such as grip strength or answers to other questionnaires) and quality of life values of a preference-based instrument can be established. The resulting regression equation can subsequently be used to predict what the quality of life values of a preference-based instrument would have been, had they been measured. In this thesis, the advantages and disadvantages of such mapping procedures are studied.

A strong practical consensus

Despite the open questions above and ongoing debates in the literature, there appears to be a relatively strong practical consensus regarding how to obtain and use QALY estimates. In some instances, this practical consensus is represented by existing guidelines which attempt to guard quality and uniformity in terms of methodology of the measurement of effect for

CUA. Regarding the four open questions introduced in section 1.4, the *Guide to the methods of technology assessment* (46) of the National Institute for Clinical Excellence (NICE) of the United Kingdom states the following: 'to quantify the effects of technologies, the EQ-5D (…) is preferred'; '[Quality of life values] should be based on public preferences using a choice-based method' and 'When EQ-5D data are not available , methods can be used to estimate EQ-5D utility data by mapping (…)' (see sections 5.4.2, 5.4.4, 5.4.5, 5.4.6 and 5.4.7 in the NICE guide (46)). The Dutch Board of Health Insurance (CVZ) also mentions the EQ-5D as the preferred outcome measure (47). The guidelines of NICE and CVZ reflect a strong practical consensus for using generic preference-based measures, for the use of the TTO (since EQ-5D has a tariff based on TTO), for the general public to assign quality of life values to health states and (if preference-based values are not available in a given patient population) for the application of mapping techniques to estimate quality of life values when they have not been measured. While the rationale for *having* measurement guidelines may be quite clear –there is a real world need for standardized data- the *content* of guidelines may be contestable on both theoretical and empirical grounds.

1.5 Research questions and outline of this thesis

This introduction started out with the notion that economic evaluations (of health interventions) aim to support welfare improving decisions and that this goal requires that 'content' is given to the concept of 'welfare'. In CUA, a policy option is considered to be welfare improving when the incremental cost effectiveness ratio (equation 3) is smaller than an appropriate threshold, reflecting the societal willingness to pay for a QALY. Computing QALYs, however, is not straightforward and is associated with several methodological challenges. This thesis investigates some of these methodological challenges, in relation to measurement and valuation of health related quality of life. Below, the research questions that are addressed in this thesis are formulated. These questions revisit some of the key issues in measuring the quality adjustment part of the QALY. The answers aim to improve the validity of the QALY as a measure of benefit in CUAs and, by extension, to improve policy decisions based on CUA.

Research question 1
Who's preference values should determine the value of health states?
Although many important text books indicate that the question 'who should value health' is essentially unanswered and the topic of normative debate (2, 4), a common practice in, for example, The Netherlands and the United Kingdom, is to have a representative sample of the general public value health, a practice which is endorsed by the influential Panel on Cost Effectiveness in Health and Medicine (48). In Sweden, on the other hand, preferences for health states are elicited from patients. It has been empirically shown that preference values of

the general public and patients are different (see for example references (49-52)), with patients generally attaching higher values to health states than the general public. A clear cut description of the issue is formulated by Milton Weinstein and colleagues:

> 'Value is equated with preference or desirability. A critical question is: desirable to whom?' (12).

In **chapter 2** it is argued that excluding patient preferences reflects insensitivity to the advantages of (also) using patient preferences and that more inclusive approaches in decision making could be explored.

Research question 2)
How can the TTO exercise be improved for the measurement of preferences for health states worse than dead?
It has been shown that the same health state may receive rather different values depending on the specification of different parts of the TTO task (42, 43). It is well-established that the exact specification of the preference elicitation task may influence the observed preferences and, consequently, final QALY calculations and cost-utility ratios. Unsurprisingly therefore, the optimal specification of preference elicitation tasks has become an important research topic. One specific point of attention in the dominant TTO procedure is that different specifications of the TTO task are used when valuing health states worse than dead than when valuing health states better than dead. New specifications of the TTO exercise, called lead time TTO and lag time TTO (42, 53) have been developed to overcome this problem. These novel methodologies of eliciting preferences for health states are discussed and tested.

In **chapter 3** it is shown that these new specifications of the TTO exercise can indeed overcome some of the problems related with conventional TTO exercises. However, it is also shown that behavioral characteristics have a large influence on the value that respondents place on a health state, leading to new methodological concerns. **Chapter 4** describes research into the timing of hypothetical ill health (i.e. now or in the future) in the new specifications of the TTO. In lag time TTO, the impaired health states 'begin' immediately and are followed by a period of 'lag time' in perfect health. In lead time TTO, a period of perfect health is subsequently followed by the impaired health state under valuation. Hence, the timing of the health state under valuation is different in both approaches. It is shown in chapter 4 that individuals, who value hypothetical health states, prefer being ill in the future rather than being ill immediately. As a consequence, two identical health states occurring at different moments in time, receive a different health state value, with the one occurring later receiving a higher health state value, solely due to the timing of the health state in the valuation exercise.

Research Question 3)
Are generic preference-based measures preferable to condition-specific preference-based measures?
The use of generic rather than disease specific descriptions of health is generally advocated, as it enables comparisons across different disease areas. However, it has been argued for several diseases that generic measures are not always (sufficiently) sensitive to disease specific improvements. To address this issue, new condition specific preference based questionnaires have emerged for, for example, asthma (54), urinary incontinence (55), erectile dysfunction (56) and epilepsy (57). Chapter 5 reflects on this development, alongside reporting the results of large study in which three new condition-specific questionnaires were developed.

In **chapter 5** it is argued that condition-specific preference-based questionnaires have some advantages compared to the generic preference-based EQ-5D. Relative advantages identified were a better sensitivity to the measurement of mild impairments and, in some cases, a better ability to discriminate between two groups of patients that have the same illness but different types of symptoms. Chapter 5 presents three new condition-specific preference-based measures which future researchers can test and possibly use in CUAs of interventions that target cancer and multiple sclerosis.

Research question 4)
Is it possible to estimate health state values using ordinary least squares regression?
Besides the methodological and theoretical issues described in research questions 1, 2 and 3, health economists frequently encounter more practical issues in the measurement of benefit. A key practical issue is that HRQoL has not always been measured in a way suitable for QALY calculations (i.e. preference-based) (58). Therefore, this thesis discusses the estimation of preference based HRQoL values in patient groups where quality of life has not been measured with preference-based instruments.

In **chapter 6** ordinary least squares regression is used to estimate the statistical relation between health state values of the EQ-5D and other more commonly gathered information on patient health. This statistical relation, captured in a 'mapping function', can then be used to estimate EQ-5D health state values in patient samples that have collected the commonly gathered data, but not the EQ-5D. Four algorithms are presented and tested in chapter 6. These can be used to estimate health state preference-based values in cancer, arthritis and multiple sclerosis trials based, on other patient characteristics.

A problem with 'mapping functions' based on ordinary least squares regression is that they tend to overestimate health state values of patients that are in very poor health. In **chapter 7** this problem is further explored, specifically in its relation to health state values of EQ-5D. A stepped

approach, with a separate 'mapping function' for patients in poor health, seems to reduce the problem of overprediction, as is shown in chapter 7.

Finally, chapter 8 discusses the merits and limitations of the research performed for this thesis, and sketch some avenues for future research. Note that chapters 2-7 are based on publications in (or intended for) international peer reviewed journals and can thus be read independently.

2

The royal road or the middle way? Patient and general public preferences for health outcomes

With Werner Brouwer

ABSTRACT

In economic evaluations of health care interventions, benefits are often expressed in terms of Quality-Adjusted Life-Years (QALYs). The QALY comprises length and quality of life into one measure which allows cross-disease comparability. The quality adjustment of the QALY is based on preferences for health states. An important normative choice is the question whose preferences for states of health we wish to capture. The answer to this question is directly related to the normative question regarding the appropriate maximand in health care decisions. Currently, preferences are commonly derived from the general public, rather than from actual patients. This choice, which can have large consequences on final outcomes of economic evaluations, appears to increasingly be a topic of debate. This paper clarifies and furthers the discussion regarding the appropriate source of preferences for health state valuations.

2.1 INTRODUCTION

Preference values for health states are generally interpreted as the 'health related quality of life' in a health state. There is compelling evidence that preference values for health states of patients differ from those of the general public and generally, patients consider their own health state to be more preferable than the general public would judge it to be (49-52). As a consequence, spending one year in some impaired condition is likely to be worth more from the perspective of a patient than from the perspective of the general public, as patients generally place higher values on impaired health states. In the Netherlands and the United Kingdom, regulatory bodies prescribe preferences for health states used in the context of economic evaluations to be obtained from the general public. In Sweden, however, patient preferences are preferred. The differences between general public and patient preferences, and the international differences in guidelines in this context, stress the need for an open and thorough discussion of the issue 'whose preferences count'.

In economic evaluations, effects of health care interventions are commonly measured and valued in terms of Quality Adjusted Life Years (QALYs). The QALY combines quality of life (morbidity) and survival (mortality) in a single metric. Preference scores indicate the desirability of health states compared to being in full health and form the quality adjustment of the QALY. The differences between the QALY score prior to treatment and that of the QALY score post treatment (or that between an intervention and a control group) determines the size of the gain (or loss) in quality and/or length of life. Preferences elicited from the general public (referred to here as 'general public preferences') are often considered to be most appropriate for use in health care decision making. However, there has always been considerable doubt that public preferences "tell the whole story" (59) and recently the issue of the appropriate source of health state valuations is again gaining wider attention (60-65).

The discrepancy between preferences for health states obtained either from patients or the general public is often explained by the umbrella concept 'adaptation' which may entail many different processes that "mitigate the impact of … circumstances" (66). According to Sprangers & Schwartz (67) people want to feel as good as possible about themselves, which results in adaptation to less than favorable conditions. On average, patient preferences indicate a smaller impact of health impairments than expected by the general public (49, 51, 68), although there are studies reporting opposite effects (69-71), as well as studies that only report a difference when preferences are measured with a time trade-off task, rather than with a discrete choice task (72). Any difference found between patient and general public preferences indicates that patient preferences provide other or additional information, which is not captured with general public preferences and vice versa. When only one source of information (i.e. either patient preferences

general public preferences) is prescribed in health care decision making, there is a risk of losing information that is potentially relevant to the decision making context, and therefore such a choice requires justification.

It is not straightforward to hypothesize precisely how using either patient or general public preferences affects the outcome of economic evaluations. In some instances using patient preferences may result in more favorable cost-effectiveness ratios than using public preferences (73), but this need not to be the case (74). The effect of using patient or public preferences also depends on whether an intervention is purely quality of life improving or rather life prolonging, as we will highlight below.

In this paper, we aim to clarify and further the discussion regarding the appropriate source for health state valuations in the context of economic evaluations. To that end, and given the dominance of the general public as source of health state valuation, we will revisit some main arguments for using general public preferences put forward in the debate. We will conclude that the current justification for the use of general public preferences for health states is, at best, incomplete, and suggest an alternative way forward.

Background

A sound discussion about the appropriateness of patient and general public preferences requires clarity (which is now sometimes lacking) about main concepts important in the debate. Therefore, we first clarify some important concepts.

Patients and the general public

Patients are those individuals who are currently experiencing impaired health. Patient preferences generally refer to patients valuing their own experienced health state. General public preferences generally refer to the valuation of health states, described in some way, by the general public. A 'general public' sample will consist of both patients and non-patients, and thus cannot be considered a non-patient or 'perfectly healthy' sample (14). In the Dutch general public sample used for the valuation of health states, described with the EQ-5D descriptive system, 21.1 percent of the population indicated to be in 'fair/poor' health (75). Moreover, it is important to note that a representative sample of the general public is unlikely to contain large groups of individuals experiencing the health states commonly valued in health state valuations. Thus, the general public is asked to value a state of health they are mostly not experiencing at the moment of valuation, and which they typically never have experienced at all (although some might have).

Ex-ante and Ex-post information

In the context of health state valuations, ex-post preferences are obtained after/during an individual experiences illness, and ex-ante preferences are obtained before the experience of illness. According to proponents of the ex-post approach, public decisions, such as health care resource allocations, ought to be based on the "tastes of individuals in those states of nature that are realized" (76). In the current example, these are patients who are experiencing a specific health state under valuation. The ex-post position is contrasted with the ex-ante position, which suggests that social decisions, like insurance decisions, ought to be based on the tastes of individuals about states of the world (here health states) that are not yet realized or currently experienced. In this case, that would be the general public. In 1890 the famous economist Alfred Marshall described the distinction in terms of desires and satisfactions in his influential 'Principle of Economics' as following:

> "It cannot be too much insisted that to measure directly, or *per se*, either desires or the satisfaction which results from their fulfillment is impossible, if not inconceivable. If we could, we should have two accounts to make up, one of desires, and the other of realized satisfaction. And the two might differ considerably. For … some of those desires … are impulsive; many result from the force of habit; … and many are based on expectations that are never fulfilled." (Italics in original. Marshall quoted in Joan Robinson (77)).

In the context of health state valuations, however, the demarcation between ex-ante and ex-post valuations, or desires and satisfactions, is not entirely correct for two reasons. First, the general public also contains patients, and thus does not fully align with an ex ante position. Indeed, a purely ex-ante position would require a sample of only healthy persons or, at least, a sample with no one in the state under valuation. Second, as further explained below, depending on the valuation method used, preferences elicited from patients can also be considered to be 'ex-ante', since the elicitation process itself also includes non-experienced health states.

Preferences, decision utility and experienced utility

Preference is a general term that is frequently used indistinctively to describe all sorts of valuations and orderings of alternatives. It is possible, however, to distinguish between different types of preferences. The preference one holds for states of the world before they actually occur can be labeled as decision utility which can be meaningfully differentiated from the experienced utility once one is in that state of the world (78). Preferences for health states, whether for own health in patients or for hypothetical health states in the general public, are typically elicited in choice based valuation studies. The choices individuals make in such stated choice experiments indicate the order and strength of preferences for certain aspects of health. According to Kahneman & Sugden (78), any preference elicitation, whether from patients or the general public, is (partly)

ex-ante information as the elicitation process itself involves choices between health states of which at least *one* state is hypothetical, given that one can only be in one state of the world (i.e. health state) at a time. In a Time Trade-Off exercise, for example, respondents are asked to choose between living in an impaired health state for a period of time *t* or living in full health for a period of time *x* where $x<t$. From the perspective of the (majority of the) general public, the impaired health state will be hypothetical while perfect health is the experienced state, but from the perspective of the patient full health will be hypothetical while the impaired state (when valuing own health) is experienced. Therefore, from both perspectives the stated choice design produces information (partly) based on ex ante preferences (decision utility) and partly based on ex post preferences (experienced utility). While common health state valuations may mix decision and experienced utility, the obtained valuations do differ considerably when elicited in the general public rather than in patients. This difference may have important consequences for the size of health gains, which, in economic evaluations, is based on the valuation of health.

Valuation of health and effect measurement

The valuation of health refers to the process of assigning a preference value or a 'utility number' to a certain state of health. The measurement of health benefits due to a treatment refers to pre- and post-therapy differences (or those between a treatment and a control group) in the health states of patients, which can be subsequently quantified by attaching preference values to both health states and calculating the difference. When public preferences are used, effect measurement is a process distinct from the valuation of health. Then, commonly, effect is measured by letting patients indicate their state of health on a descriptive system (questionnaire). Each health state in the descriptive system has a preference score attached to it, separately elicited from the general public. The magnitude of effect of a treatment is then defined by comparing the preference values attached to the health states of patients pre-therapy (or intervention group) to the preference scores attached the health states they were in post-therapy (or control group). When using the patient perspective, preferences are generally (but not necessarily) elicited for own health. In that case effect-measurement and the valuation of health can be performed in one and the same process.

It is important to note that measuring changes in health states using a descriptive system (like EQ-5D or HUI-III) is commonly performed in the affected patients. As we will discuss later, adaptation may not only influence the value attached to health states, but may also influence the way in which descriptive systems are interpreted and completed. At least, it is important to distinguish between the valuation phase and the measurement phase, whenever relevant.

Above, the main concepts that are relevant to the debate about "who should value health" were clarified. Now we proceed with the debate itself, which we categorized in three key subjects: the societal perspective, adaptation (with some related arguments) and the insurance principle.

2.2 REVISITING THE ARGUMENTS

Without claiming to be exhaustive, we discuss some of the main arguments put forward in favor of using use of public preferences below. We will argue that none of these arguments, alone or jointly, provides a convincing argument for solely using of general public preferences.

The societal perspective

The influential Panel of Cost-Effectiveness in Health and Medicine (48), stated that "for purposes of resource allocation, the relevant preferences are those of the general public" (p.111). The Panel links this claim to the argument that cost effectiveness analyses ought to adopt a societal perspective:

> "One way to see the desirability of the societal perspective … is to imagine for a moment that we are looking at the world before we are born, or at least before we encounter any serious health problems, and to ask what kind of world we would like it to be. In that "ex-ante" position we would not yet know what sort of health problems we were destined to develop … we might reasonably prefer a system in which decisions about health interventions reflected the seriousness of the problem and the ability of the intervention to do something about it, without reference to the specific individuals with the problem or to particular budgets or special interests". (p.6 & 7).

A central element of this quote refers to the Rawlsian veil of ignorance. Adopting such a perspective, according to the Panel, would lead us to favor a system in which decisions about health interventions are made without favoring special interests. The implicit assertion is that patients are a 'special interest' group while the general public, behind 'the veil of ignorance' is not. However, the 'veil of ignorance' in itself is not an argument for the use of public preferences for health states. Indeed, one can doubt whether the general public reflects Rawls' 'original position' where "no one knows his place in society, his class position or social status … and the like" (79). In that sense, the original position is a truly ex ante position, where every health state is still hypothetical, including perfect health. Obviously, the reference to the situation "… at least before we encounter any serious health problems…" is a rather specific interpretation already. In that context, however, one could state that a specific part of the general public, i.e. healthy individuals, would be the relevant source of information.

Whether one, behind the veil of ignorance, would not want to know how health states are actually experienced, seems unclear. Thus, if the societal perspective refers back to the Rawlsian original position, it seems that the assertion that a sample from the general public would be the 'logical' source of preferences is perhaps a bit too straightforward, since it may equally well

be that behind the veil of ignorance, we might reasonably prefer that the seriousness of the problem and the ability of an intervention to do something about it is based on values of those who have first hand experience with the seriousness and effectiveness, i.e. patients.

According to the panel (48), the quality adjustment part of the QALY has to be based on decision utilities (rather than experienced utilities), and community samples of the general public are the appropriate source of preferences to make comparisons across interventions and populations (p.99). Public preferences are said to be "a logical extension of the societal perspective" (p.99). The societal perspective normally relates to the scope of costs and benefits considered in economic evaluations, like Gold et al. (1996) define:

> "When a CEA is conducted from the societal perspective, the analyst considers everyone affected by the intervention and counts all significant health outcomes and costs that flow from it, regardless of who experiences the outcomes or costs" (48) (p 6).

Following the societal perspective, costs and effects beyond those directly related to the patients are considered of importance, such as productivity costs (9) and measuring health effects in informal caregivers (80). It is not entirely clear, however, why adopting this broad perspective which incorporates everyone affected by the intervention should suggest that the general public is the most appropriate source of preferences. At most, it suggests that preferences of, and effects in, others than patients matter too.

Adaptation: hypothetical health states

The use of preferences obtained from general public "might be reflecting the objective fact that the range of capabilities for people having certain conditions and disabilities is lessened compared to the normal range" (48) (p.99). In contrast, patients may adapt to their condition and therefore have shifted preferences. The general public, on the other hand, is generally healthy and may judge the loss of capabilities from the viewpoint of someone who is in full health. This may result in a 'better', or at least uniform, representation of the 'distance' between being in full health and having the health impairment. For patients the reference point they are reasoning from may have shifted and perfect health may be too hypothetical (4), which may be considered problematic when wishing to come to universally applicable health state valuations. On the other hand, for the general public impaired health may be too hypothetical, causing the general public to overestimate the impact of health impairments yielding rather low, and perhaps relatively often negative, that is 'worse than dead', preference scores.

The differences between patient and public preferences for health states seem to indicate that the general public may not adequately forecast experienced utility of being in a health state, and are, in that sense, 'wrong'. Fitzpatrick writes:

> "individuals appear to make value judgements about the desirability or undesirability of hypothetical health states by focusing on the transition from their own current state to the imagined hypothetical states" (Fitzpatrick, quoted in Sharma et al, 2003 (66)).

The focus on transitions may lead to beliefs about the impact of disease by the general public that is not reflected in experiences of patients (61). Indeed, when uninformed, there may be a general misunderstanding of what it is like to live with a disability (81), at least when this disability is presented to the general public in the form of a health state in a valuation task. Indeed, there may be several complex mechanisms involved that cause differences between health state values derived from patients and the general public. For now, however, we are most interested in how these differences in valuations affect economic evaluations.

Adaptation: underestimation of effect

In the words of Cohen, when explaining the critique of Amartya Sen on the informational base of utility:

> "The fact that a person has learned to live with adversity, and to smile courageously in the face of it, should not nullify his claim to compensation" (82).

Indeed, one of the perhaps most influential reasons for favoring the general public (or rather healthy respondents), as source for health state valuations is that the "gains possible from the intervention become larger when the perspective is that of the general population" (48) (p.102). Adaptation may affect the size of the effect of successful treatments, as explained below, with important implications for resource allocation in health care.

Example 1: curing a patient

Let's say that the general public values some health state at 0.4, and patients value this health state higher due to adaptation, for example 0.6. This means that curing that patient (to 1 for example) yields a gain of 0.6 according to the general public but of 0.4 according to patients (table 1). Using the values of patients in such cases therefore lowers the size of the benefit (and cost-effectiveness) of treatments. Since the lower benefit in case of patient preferences is due to the process of adaptation, it is questionable, in line with the quote of Cohen above, whether this process should lead to lower claims to health care. Using general public valuations, therefore, can be advantageous for patients in that sense. This is in line with the finding that patient preferences

for their own health state are 'unusually' high. The practical implications of this 'unusually' high value, compared to general public values, are indeed relevant in many situations.

Table 2.1 Curing patients

	General public preferences	Patient preferences
Pre treatment	0.4	0.6
Post treatment	1	1

Example 2: prolonging life

When we extrapolate the assumptions from the previous example (i.e. that patients place higher values on health states) to life prolonging treatments, we find an opposite effect. Indeed, for treatments that prolong life, patient preferences may yield *more favorable* cost-effectiveness ratios. When an intervention prolongs life, a higher quality of life weight in those gained life years will result in higher health gain in, for example, cancer patients (50). Using patient preferences, any additional life year lived receives 'more weight' than with general public preferences, as exemplified in table 2. Multiplying additional life years lived with the values from table 2 indicates that more QALYs are gained when using patient preferences. In practice, this means that a cost-effectiveness ratio based on general public preferences may be higher, simply because the general public considers the impaired life years gained less valuable than those whose life is actually prolonged, be it in an impaired condition. Obviously, this effect is larger when more life years are gained.

Table 2.2 Prolonging life

	General public preferences	Patient preferences
Post therapy	0.4	0.6

Example 3: Adaptation opportunities

Theoretically speaking, adaptation may also increase the possible gains from a curative intervention (contrary to what was argued in example 1). This depends on the nature of the adaptation process. Imagine a very poor health state, valued rather low by general public and patients, 0.3 and 0.4 respectively. Treatment can bring these patients to a better, but still impaired, health state, to which patients may adapt more easily, for example, going from a hospital bed to a wheelchair. The new health state (the wheelchair) may allow a patient to have a relatively large 'adaptation opportunity' compared to the old health state. In other words, the magnitude of adaptation can be asymmetrical in size pre and post therapy, due to the nature of the condition pre and post therapy. At present, it appears largely unknown when and with which magnitude adaptation occurs in impaired health. This (theoretical) example shows that using patient preferences which include adaptation does not always result in less favorable incremental

cost-effectiveness ratios and indicates that the 'adaptation' argument for using general public preferences is perhaps less straightforward than sometimes assumed.

Table 2.3 Asymmetrical adaptation

	General public preferences	Patient preferences
Pre treatment	0.3	0.4
Post treatment	0.6	0.8

In sum, using general public rather than patient preferences can be advantageous as well as disadvantageous for patients, depending on the specific circumstances. Identifying these circumstances is of crucial importance for a better understanding of the real effects of (ignoring) adaptation in cost-utility analyses in public decision-making.

Adaptation: an aside on descriptive systems

In current practice the outcome of interventions is assessed using a preference-based descriptive system (i.e. a questionnaire with an algorithm attached to it). Although issues with adaptation often lead health economists to favor public preferences (83), the current two-stage assessment system with a questionnaire and a separate algorithm to compute preferences of the general public does not necessarily rule out adaptation fully. Adaptation may also be 'picked up' by the descriptive system of a questionnaire (4). The elderly, for example, might report that they do not 'have problems' with their mobility or usual activities as they consider their somewhat declined mobility or shifted usual activities to be normal *for their age and peer group*. The same elderly individual will probably not state that s/he is able to walk 5 km without resting, suggesting that the phrasing of the question itself might also induce sensitivity to adaptation despite the presence of a value set elicited from the general public. Thus, next to the valuation phase, descriptive systems may be sensitive to adaptation. As a result, using public preferences to value health states of a descriptive system may not be a complete solution to the issue of adaptation.

The insurance principle

Some have argued that the "insurance principle", (84) is an argument for using public preferences to value generic health states:

> "But whose preferences should be used to determine the *value of a treatment* for purposes of societal allocation rules? Patients who rely on others to pay their medical bills … cannot expect that these others will pay for everything they (the patients) might wish to receive. Permitting patients unlimited access to care based on post-illness preferences would too often result in the provision of marginally beneficial care. The lack of any associated marginal financial cost to the patient often makes any potentially beneficial treatment desirable or

> 'worth trying'. For this reason, the importance and priority of treatments should be based on the average pre-illness preferences of the entire beneficiary population …" (84).

The insurance principle, therefore, argues that having insurance is likely to alter sensitivity to the cost of treatments and that, in absence of opportunity costs to the individual patient (given insurance), experiencing illness may result in inelastic demand for treatments. Consequently, patients may desire marginally beneficial care despite high costs of care, as they are insured against those costs and share the burden of additional costs of treatments with the entire insured population. When patient preferences would be used to determine the content of insurance packages, the price attribute of treatments would possibly be disregarded by the patients. The insurance principle points out that, as a consequence, there would be no sensible constraint on the provision of marginally beneficial care, and the insurance premium, paid both the patient and the entire insured population, would soar.

This, of course, may be completely correct, but is not very relevant to the question at hand here, as it does not apply to the current use of preferences in cost-utility analysis. In cost-utility assessments, preferences are elicited for health states. These preferences are subsequently used to indicate the effect of treatment, separately from both the costs and the monetary valuation of health changes. In a later stage, mostly without direct patient involvement, the marginal societal costs of treatments are divided by marginal preference based effects, relative to some comparator, and compared to some relevant societal willingness to pay threshold for health gains. Given how preferences are commonly used in economic evaluations, the insensitivity to costs of treatments induced by insurance and in some cases the decreased marginal utility of money (e.g. in life threatening situations) is irrelevant for the question whose values count in valuing health states. Patients would only be called upon to value health states and to indicate in which health state they are in different stages of the illness and treatment. This information is then used to calculate the effect a treatment, but the desirability of *a treatment itself* is not, in any direct sense, asked. Using patient preferences in cost-utility analyses would therefore provide a logical limit to the provision of care, just like public preferences.

Another interpretation of the insurance principle, "found routinely in other areas of insurance" (84), may be considered more compelling. An individual chooses an insurance package, in an ex ante situation, under uncertainty about a future state of the world, such as the likelihood of an adverse event like an earthquake that may damage property, or, when applied to health insurance, the likelihood of needing treatment. Since the individual chooses a specific health insurance package *ex ante* (since in many contexts burning houses are hard to insure), the logical thing would be to also use ex ante preferences in deciding on the package. Moreover, since in many Western countries all individuals are (obliged to be) insured, it is also logical to use the

preferences existing in a representative sample of the general public. However, two points can be made here indicating that patient preferences can still be relevant. First of all, while individuals and collectives often have to decide ex ante on insurance packages, this does not mean that they would always opt for the use of own ex ante preferences regarding health states. It is well conceivable that informed citizens would opt to use ex post preferences that arguably better reflect the actual impact of illness on well-being, if this were possible. Second, in a collectively financed health care system, the ex ante position holds for most insureds and payers, but not for all. One may well claim that the system is constructed (using solidarity principles) in such a way as to maximally contribute to the avoidance and relief of *actual* losses of well-being, rather than expected losses and such losses can be observed in those individuals at which the health care system is primarily targeted, i.e. patients.

If the insurance principle is taken to mean that those who pay for health care (through insurance premiums) should be able to participate in policy decision concerning health care, it has to be acknowledged that such involvement may take many forms and does not preclude the use of ex post patient preferences for health states. Hence, while providing important arguments in the debate regarding whose values (should) count, the insurance principle, regardless of the specific interpretation, does not seem to provide a decisive argument in favor of ex ante, public preferences *for the valuation of health states*.

2.3 WAY FORWARD

Although the properties of both the public and the patient preference perspectives have been extensively discussed in the literature, there does not seem to be any conclusive or theoretically sound justification for disregarding either perspective. Historically, the disadvantages of patient preferences and advantages of general public preferences appear to have been emphasized, resulting in (or perhaps even following from) a large practical consensus to use general public preferences. However, it seems difficult to claim that patient preferences, which are arguably better informed concerning the actually experienced loss of health related quality of life in a health condition, do not carry important information for societal decision making. Indeed, public preferences provide a reliable and systematic standardization for the computation of QALYs, but do not adequately forecast the experience of physical and mental impairment due to shifting preferences and, potentially, a lack of experience with impairments.

Reflecting on the different arguments put forward in the literature, it seems that both public and patient preferences hold information on health related quality of life that is potentially relevant for societal decision making. Consequently, neither public nor patient preferences can be the 'royal road' to the evaluation of health care interventions in themselves. One may

wonder whether the two positions are necessarily mutually exclusive. To date, the debate seems to have centered around the question which of the two perspectives would be most appropriate. It seems unclear, however, why such a dichotomy would be necessary or, in fact, useful. It has been suggested to provide the general public with "more information on the size and the nature of the adaptation experienced by patients over time" (4) in valuing health states. One may feel, however, that such a mix of both worlds may also result in a loss of information, since exactly the discrepancy between QALYs derived from general public or patients is relevant information to decision-makers. This helps decision makers in recognizing the consequences of applying either perspective. To quote Sen:

> "Euclid is supposed to have told Ptolemy: "there is no 'royal road' to geometry". It is not clear that there is any royal road to evaluation of economic or social policies either. A variety of considerations that call for attention are involved, and evaluations have to be done with sensitivity to these concerns." (85) (p.85)

For the evaluation of health care interventions it may be important to acknowledge that while the prospect of some impaired health states may seem extremely undesirable to the general public, it may be experienced as less undesirable once experienced by actual patients. These two considerations, both before and after health impairments, seem relevant in their own right in decision making.

Indeed, it seems difficult to argue why a societal decision maker, faced with a choice between two programs which are fully identical in all respects but the ex post patient health state valuation would not, and even should not, use this information to allocate the money there where less adaptation is possible. This can be highlighted with two examples. First, when decision makers have to prioritize one of two programs (table 4a), which are equal in all aspects but the burden of illness from the perspective of the general public, would they not logically prioritize program A over program B, as in this program the burden of illness from the perspective of the general public is higher (i.e. a lower quality of life)?

Table 2.4a Prioritization based on differences in burden of illness across programs

	General public preferences	Patient preferences
Program A	0.5	0.7
Program B	0.6	0.7

Second, when decision makers have to prioritize one of two programs (table 4b), which are equal in all aspects, except that in program B patients are more prone to adaptation than in program B, would they not logically prioritize program A over program B?

Table 2.4b Prioritization based on differences in burden of illness across programs

	General public preferences	Patient preferences
Program A	0.5	0.6
Program B	0.5	0.8

The fact that patients may adapt to certain health impairments (to different extends) need not be seen as disqualifying patient preferences, but rather stress the importance of including them in societal decision making. Still, some have argued that adaptation (and the use of preferences affected by adaptation) results in undervaluation of the benefit of curative treatments. It seems that this 'protecting the patient against own adaptation' argument has received most attention so far, even though it is only a part of the full story as explained above. Often disregarded, but nonetheless of importance, is that the advantage of using public preferences -the potential insensitivity to adaptation- may currently not even be achieved as the descriptive systems to which societal preferences are applied are sensitive to adaptation.

Perhaps the way forward for the evaluation of health care interventions is to acknowledge the plurality of relevant perspectives rather than advocating only one 'royal road'. Indeed, there does not seem to be any reason why health economists could not inform decision makers with cost utility analyses based on two kinds of outcomes: QALYs calculated with patient preferences and QALYs calculated with societal preferences.

2.4 CONCLUSION

Preferences for health states are commonly derived from the general public, rather than from actual patients. This choice appears to increasingly be a topic of debate. We argued that both viewpoints are reasonable and compelling, and that, therefore, the challenge is to investigate possibilities to intelligently combine the different sources of information.

3

Time to tweak the TTO: results from a comparison of alternative specifications of TTO

With Arthur Attema, Mark Oppe, Nancy Devlin and Elly Stolk.
European Journal of Health Economics (2013)

ABSTRACT

This article examines the effect that different specifications of the time trade-off (TTO) valuation task may have on values for EQ-5D-5L health states. The new variants of the TTO, namely lead-time TTO and lag-time TTO, along with the classic approach to TTO were compared using two durations for the health states (15 years and 20 years). The study tested whether these methods yield comparable health-state values. TTO tasks were administered online. It was found that lag-time TTO produced lower values than lead-time TTO and that the difference was larger in the longer time frame. Classic TTO values most resembled those of the lag-time TTO in a 20-year time frame in terms of mean absolute difference. The relative importance of different domains of health was systematically affected by the duration of the health state. In the tasks with a 10-year health-state duration, anxiety/depression had the largest negative impact on health-state values; in the tasks with a five-year duration, the pain/discomfort domain had the largest negative impact.

3.1 INTRODUCTION

Attempts to improve the measurement of health-state values have led to several methodological innovations in valuation techniques such as the time trade-off (TTO), which are used to determine the desirability of a hypothetical state of health. Novel specifications of the classic approach to TTO have been developed to make the measurement of health states considered 'worse than death' (WTD) more accurate (36). Lead-time TTO and lag-time TTO are in theory equally capable of addressing issues in the valuation of WTD health states. However, there is little evidence on how these methods compare. To help fill that gap, the classic approach to TTO (here referred to as 'classic TTO') and two novel methods (lead-time TTO and lag-time TTO) have been compared in an online study.

In the TTO, a value can be assigned to a health state by letting respondents trade off length of life against quality of life. The resulting value is generally taken to reflect the health-related quality of life per period for the duration of the health state. Classic TTO applies two different procedures for the valuation of health states that are considered better than death and those considered WTD. Therefore, TTO values for health states better than death and WTD may not lie on the same utility scale (27). Furthermore, sacrificing one additional year in the WTD procedure lowers the value of a health state more than when one year is sacrificed in the better than death procedure (43). While values for health states better than death are restricted to points between 0 and 1, the values measured with the procedure for WTD can become very low (42, 86). Therefore, an arbitrary transformation of those values is subsequently needed to avoid distortion of the mean value.

The two alternative specifications of TTO do not have the above-mentioned limitations of the classic version. In lead-time TTO, first proposed by Robinson and Spencer (39) and extensively discussed elsewhere, the impaired health state 'begins' after a period of healthy years (the lead-time) (42). In lag-time TTO, healthy life years *follow* the impaired health state rather than *preceding* it (87). Probably the most important application of either lead-time TTO or lag-time TTO is in the valuation of health state descriptive systems, such as EQ-5D.

In lead-time TTO, the health state under valuation is further away in the future than in lag-time TTO, where the health state 'begins' immediately. It could be hypothesized, therefore, that lead-time values for the same health state will be higher than lag-time values if respondents have positive time preferences, as frequently observed (88, 89), although there are also reports of negative time preferences for TTO (90). Alternatively, it could be hypothesized that lag-time TTO results in higher values, since the lag-time of full health after a given health state might be interpreted as having been cured, which, arguably, influences the perception of the severity

of the health state. Conceptually, lag-time TTO might be more 'plausible' for mild states and curative treatments, since it is based on the premise that poor health is followed by good health. Lead-time TTO may be more plausible for very severe health states and preventive treatments since it poses that the health state starts in the future and is followed by death (91).

In this study, respondents participated in an online experiment where they engaged in either lead-time TTO, lag-time TTO, or classic TTO. The purpose was to see how the health-state values produced by each of the TTOs compare. It also investigated how both the type of TTO and the duration of a health state would affect the values for each of the EQ-5D domains of health. Values generated by the online mode of administration were compared to values estimated on the basis of a face-to-face TTO. Scores on the respondents' engagement with and understanding of the task were used to explain potential differences.

3.2 METHODS

Respondents

A sample of respondents was drawn from members of a commercial panel. Only persons between 18 and 65 years of age were approached to participate in the online experiment. Stratification to represent the Dutch population was based on gender, education, and age. Respondents were not given a financial reward for participating.

Health-state selection and description

Health states were based on the Dutch version of the five-level EQ-5D (EQ-5D-5L) (92). This instrument consists of five domains of health: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The instrument has five answer categories for each domain, generating 3125 (5^5) health states. Out of the total of 3125 possible health states, 100 were selected in light of a previously developed D-optimal design (93).

Study design

All respondents performed a combination of tasks. First, they filled out a background questionnaire. They also indicated how they perceived their own health on the EQ-5D-5L instrument and the EQ-5D visual analogue scale. Scores on the latter ranged from 0 to 100, where 0 stood for the worst imaginable health and 100 the best imaginable. Then the respondents had to choose which of two EQ-5D-5L health states they considered best in a paired comparison task. Upon completing these preliminary tasks, the respondents were randomized over five different specifications of TTO: lead-time TTO with a duration of 15 years and of 20 years; lag-time TTO with a duration of 15 years and of 20 years; and classic TTO with a duration of 10 years. Within these five specifications, respondents were randomized over 10 blocks containing

10 EQ-5D-5L health states, and each state was presented in random order. The study ended with a short feasibility questionnaire.

The TTO tasks

In classic TTO, health-state values are elicited by asking respondents if they would prefer living x years in a period of full health to living t years in impaired health where $x < t$. If respondents accept living a shorter period t in full health, they are essentially willing to trade length of life for quality of life. The health-state value is then given by x/t, at the point of indifference. When the respondents would rather trade off all healthy life years than have to live in a particular health state for period t, they indicate that this health state is worse than death (WTD), at least when the duration of that health state is equal to period t. Respondents then enter a different task to measure their negative preference values (since $x < 0$). In this WTD task, they are asked to choose between immediate death and a life of duration t, with x years in full health preceded by t-x years in the imperfect health state. The value for the health state following this WTD task is generally $-x/(t\text{-}x)$. In lead (or lag) time TTO, they were also asked if they would prefer living x years in full health compared to living t years in impaired health preceded (or followed) by l years in full health. An indifferent point was estimated by repeating this question for different values of x. The value of the health state is then given by $(x\text{-}l)/t$, where x is the estimated indifference value. When $x<l$, the formula results in a negative value, implying that these are WTD health-states.

The TTO tasks were preceded by an animated instructional video. It explained how to trade off life years by giving an example with a hypothetical EQ-5D state, whereby an animated figure of a 'doctor' pointed out the various elements of the task. The video was designed to highlight the characteristics of the different TTO tasks. Thus, the examples shown in each animation preceding the real TTO task were identical in characteristics and lay-out to the real TTO task that followed, with the exception that the health state they presented was not used in the study.

The classic TTO is a two-part task. The visual design and the health-state value equations for health states better than death are different from those for WTD health states. The other four TTO tasks have a uniform visual representation and health state value equations for better than death and WTD valuations. In all tasks, respondents are asked to choose between a fixed period in Life A and a variable period t in Life B. The value of x depends on the respondents' previous choice for either Life A or B and follows the fixed iteration procedure described below.

Iteration procedure

The first two 'steps' of the fixed iteration procedure were similar for all five TTO tasks. At the first iteration, respondents were asked to choose between two scenarios: Life A, which contained

the health state and, depending on the task, a lead-time or lag-time in full health; and Life B, which was set at the maximum of all years in full health (health-state value=1, or *x*=10, 15, or 20, depending on total time frame). At the second iteration, the health-state value of Life B was 0 (or *x*=0 for the classic TTO and *x*=10 for the other variants). If respondents preferred Life A at value=0, they would indicate that the health state is WTD. If they preferred Life B, they would indicate that the health state is better than death. After this 'sorting question', the iteration procedure continued with a choice between Life B and Life A where the value of B was set at *x* for value=0.5 or -0.5. Conditional on choosing Life A or B, the remaining iterations represented value increments or decrements of 0.1 or 0.05 with the corresponding values of *x* in Life B.

Health-state value equations

The equations applied for the lead-time TTO in a 20-year time frame are (without discounting):

1) $$10U_{FH} + 10U_{HS_i} = xU_{FH}$$

where U_{FH} is the value (utility) of full health, U_{HSi} the value of the health state *i,* and *x* the number of years in full health at which the respondent indicated being indifferent in the TTO task. Solving for U_{HSi} gives:

2) $$U_{HS_i} = \frac{x-10}{10}$$

For a respondent who considers *x*=13 years in full health equal to 10 years in full health followed by 10 years in health state *i,* the value for *i* is: U_{HSi} = (13-10)/10 = 0.3. In the same vein, the equation for lag-time TTO is:

3) $$10U_{HS_i} + 10U_{FH} = xU_{FH}$$

Equation 3 can also be solved for U_{HSi}, which again results in equation 2.

The most relevant details of the TTO specifications included in this study are described in Appendix I to enable easy comparison with other studies performed with a TTO checklist (94).

Analysis

All respondents who completed the online exercise were included in the analyses. To check for consistency in findings the analyses were rerun in a smaller sample without those respondents who: 1) indicated on the feasibility questionnaire that they did not understand the task; 2) did not differentiate between any of the 10 health states; or 3) had used only three or fewer iterations for all health states.

Comparison of health-state values

Mean lead-time TTO and lag-time TTO values were compared for all 100 health states. The different minimum health-state values set for the TTO methods distort comparisons of the mean values between tasks. For example, solving the equations for t=0 (trading in all life years) results in U= -2 for a ratio of lead-time to disease time of 2:1 and U= -1 for a ratio of 1:1. Therefore, comparisons of the mean are only made for tasks with similar attainable health-state values. Convergence of lead-time TTO and lag-time TTO with classic TTO was measured in terms of the mean absolute difference (MAD) to get a feel for the comparability of values despite the different ranges of health-state values.

The relative importance of the domains of EQ-5D in the different specifications of TTO is compared through random effects regression analysis to take account of the panel structure of the data (multiple TTO observations per respondent). Although the sizes of the coefficients are not directly comparable due to different ranges of the dependent variable (the TTO values), the relative importance of the domains within each regression model can still be compared. Independent variables in the regression model were the EQ-5D health domains, applied as continuous variables.

The online mode of administration of the TTO is still in an experimental stage. Also, the health-state values generated by the different tasks cannot be compared to a non-experimental EQ-5D-5L tariff, as the valuation protocol of the EQ-5D-5L was still under development at the time of this study. To get an indication of the convergent validity of the values produced in the online exercise, these values were compared to the estimated EQ-5D-5L values derived from a mapping function (95). These estimates reveal which health-state value is expected for an EQ-5D-5L health state on the grounds of previous valuations for the EQ-5D-3L applied in face-to-face TTO.

Task engagement and response characteristics

Agreement among respondents in the different TTO tasks was ascertained with Levene's test and Brown & Forsythe tests. It was assumed that differences in valuations between respondents, regardless of the cause, would result in greater variance and thus less precise health-state value estimate. Although larger standard deviations may reflect preference heterogeneity rather than poorer task engagement, a valuation method that is identical in all respects but the onset of the health state (i.e., before or after a period of full health) is arguably preferable if there is more agreement among respondents. Variances for classic TTO (with transformed negative values) were only compared to the TTO tasks with a 20-year time frame, as TTO values for these two lie on the same -1 to 1 scale. Accordingly, the variances were not compared to values from the TTO tasks with a 15-year time frame (with a lead-time to disease time ratio of 2:1), which lie on a -2 to 1 scale and, thus, logically have larger variances. Standard deviations, which lend themselves

to a more intuitive interpretation than variances, were plotted for lead-time TTO and lag-time TTO. Other indicators of task engagement were used as well: whether the respondents were willing to trade off any time at all (non-traders); how many iterations the respondents used before reaching their point of indifference; how many respondents 'used up' all tradable time; and how many did not differentiate between health states.

Feasibility

Differences between tasks were compared using four items of a feasibility questionnaire presented after the TTO task. Respondents were asked to indicate their level of agreement with four statements: 1) The instructions that were given made it clear what I needed to do; 2) It was easy to understand the questions I was asked; 3) I found it difficult to decide on the exact point where Life A and B were about the same; 4) I found it easy to tell the difference between the health states I was asked to think about. The answer categories ranged from 1 (completely agree) to 5 (completely disagree). Mode, median, and percentiles of the answers on these questions were compared.

Since health-state values have been shown to be affected by the number of health states valued by a respondent, we repeated our analysis using only the first five valued health states (96). We tested for significance of order effects by regressing the sequence of a health state on the number of iterations using ordinary least squares (OLS) regression, as proposed by Augestad et al. (96). All statistical analyses were run in STATA 11.

3.3 RESULTS

In total 5208 respondents finished all the tasks, with approximately 1000 respondents per task. The resulting dataset was a balanced panel with 10 TTO observations for each respondent. Respondents in the online panel were slightly older than the Dutch population average (*mean*=42.3 (*sd*=14.2) vs. Dutch population mean of 2009 = 40.1). Furthermore, the panel contained more females, with 58.3 percent female and 41.7 percent male, compared to a nearly 50/50 distribution in the Netherlands. Mean self-assessed health on the EQ-5D visual analogue scale (VAS) was 76.7 (*sd*=17.4). Regression analysis indicated that respondents used fewer iterations ($p<.001$) for health states presented later in the sequence; on average, they used 0.4 iterations less than the previous health state for each consecutive one. Therefore, where relevant, results were rerun using only the first five health states.

Comparison of health-state values

Lead-time TTO resulted in systematically higher values than lag-time TTO for the 20-year time frame (on average 0.25 higher) with larger average differences for poorer health states (Figures 3.1a & 3.1b). In the 20-year time frame, none of the lag-time values were higher than the lead-

time values. Results for the 15-year time frame were mixed: on average, lead-time TTO values were 0.13 higher in the 15-year time frame and lower than lag-time TTO values for 18 out of 100 health states (28 out of 100 using the first five health states). In terms of mean absolute deviation (MAD), differences between classic TTO and the other specifications (from least to most) were as follows: lag-time TTO in a 20-year time frame (MAD = 0.07); lead-time TTO in a 15-year time frame (MAD = 0.14); lead-time TTO in a 20-year time frame (MAD = 0.23); and lag-time TTO in a 15-year time frame (MAD = 0.26).

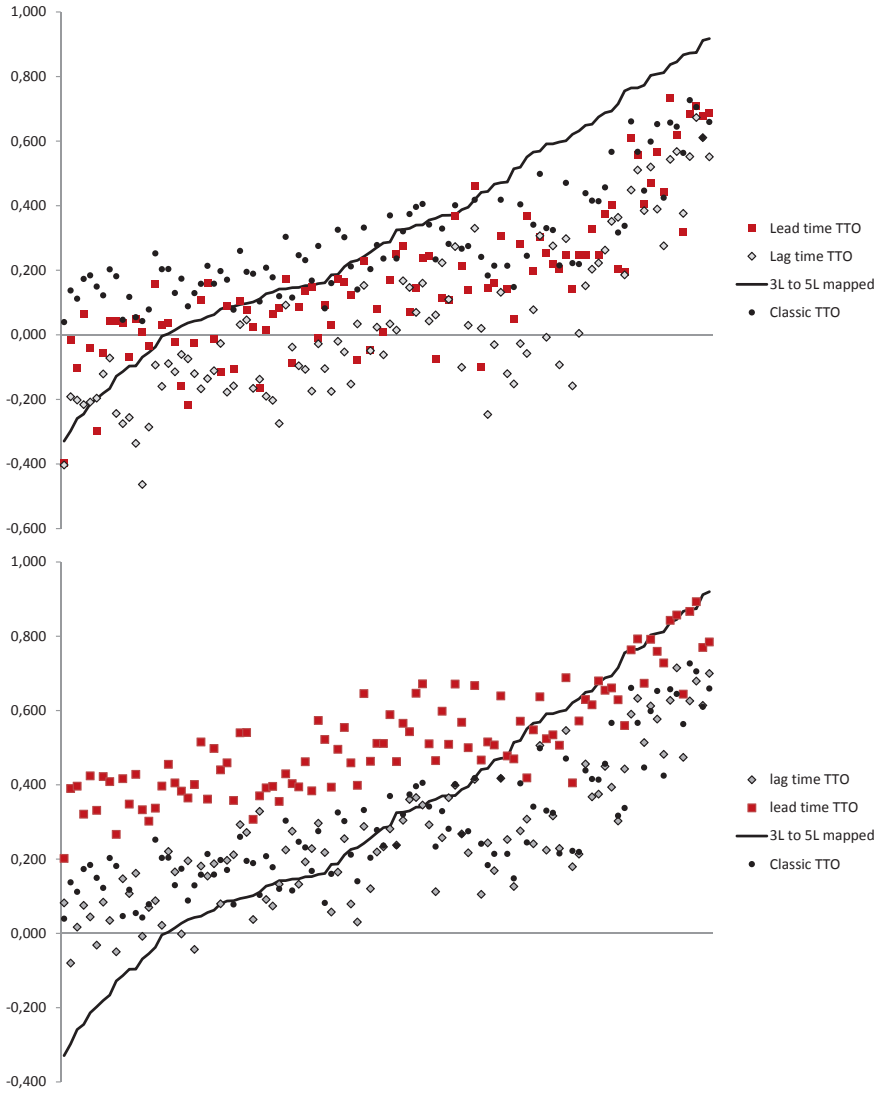

Figure 3.1a and 1b TTO in a15 year time frame (a) and in a 20 year time frame (b) compared to classic TTO and estimated TTO values.

The range of health-state values in the 15-year time frame was 1.13 for lead-time TTO (from -0.4 to 0.73) and 1.14 for lag-time TTO (-0.46 to 0.68). In the 20-year time frame, values were higher than in the 15-year time frame for both variants. The higher health-state value was most likely due to the range of attainable values in the 20-year time frame (the minimum value of the 15-year time frame was -2, compared to -1 in the 20-year time frame. The minimum value of -1 also influenced the observed range of values in the 20-year time frame, which was smaller for both variants: the range was 0.69 for lead-time TTO (0.20 to 0.89) and 0.80 for lag-time TTO (-0.08 to 0.72). The range of values produced by the lead-time TTO and the lag-time TTO was smaller than would be expected in view of the estimated EQ-5D-5L values. Classic TTO, the method used for EQ-5D-3L, also produced a range that was smaller than expected (0.69). The worst health-state value[4] with classic TTO was 0.04 (for state 55555) and the best was 0.73 (for 12111).

The specification of the TTO task influenced the relative importance of the different domains of health (Table 1). The size of the regression coefficients represents the marginal decrement in health-state values caused by scoring one point higher in a particular domain on the five-level descriptive system. The order of their relative importance was not affected by the choice for lead-time TTO or lag-time TTO but by the duration of the health state. In the 20-year time frames, with a disease duration of 10 years, the health domain 'anxiety/depression' was considered worse than 'pain/discomfort'. The inverse was found for the 15-year time frame, which has a disease duration of five years. Similarly, problems in usual activities were considered more problematic than problems with self-care in the 20-year time frame while the inverse was found for the 15-year time frame. The order in the classic TTO was different from the order in the lead-time TTO and lag-time TTO. The regression models using only the first five health states gave orderings that were identical to those found using all 10 health states.

Table 3.1 Relative importance of different domains of health

	Classic TTO		15 year lead-time TTO		20 year lead-time TTO		15 year lag-time TTO		20 year lag-time TTO	
	Coef.	imp.	Coef.	imp.	Coef.	imp.	Coef.	imp.	Coef.	imp.
Mobility	-0.026	3	-0.032	3	-0.026	3	-0.039	3	-0.036	3
Self-care	-0.020	1	-0.027	1	-0.020	2	-0.033	1	-0.028	2
Usual activities	-0.022	2	-0.028	2	-0.020	1	-0.038	2	-0.019	1
Pain/Discomfort	-0.040	4	-0.057	5	-0.031	4	-0.060	5	-0.043	4
Anxiety/Depression	-0.043	5	-0.053	4	-0.040	5	-0.058	4	-0.045	5
Constant	0.731		0.740		0.915		0.692		0.751	
Adjusted R-square	0.12		0.13		0.10		0.12		0.13	

All coefficients $p<0.05$

Imp. = relative importance

[4] Negative values transformed with $-x/(t-x)$

Task engagement and response characteristics

Lag-time TTO showed a larger variance than lead-time TTO for nearly all health states (Figure 3.2). The mean variance of lag-time TTO is higher in both the 15-year time frame ($p<.001$) and the 20-year time frame ($p<.001$). The classic TTO with transformed negative values has a smaller variance than lag-time TTO ($p<.001$) but a larger variance than lead-time TTO ($p<.001$). When including only those respondents who had indicated on the feasibility questionnaire that they thought the task was clear (answer 1 to question 1), that they understood the task (answer 1 to question 2) or had not valued all 10 health states equally, all statistical tests indicated significant differences ($p<.001$). A mean standard deviation of 0.81 (N=1067) was found for the online lead-time TTO in a 15-year time frame. When including only those respondents who were randomized to the LT-TTO in a 15-year time frame and who indicated they thought the task was clear and understood the task, the mean standard deviation increased somewhat to 0.83 (N=359). Using only the first five valued health states increased the mean standard deviation of the lead-time TTO in a 15-year time frame to 0.84 (N=533).

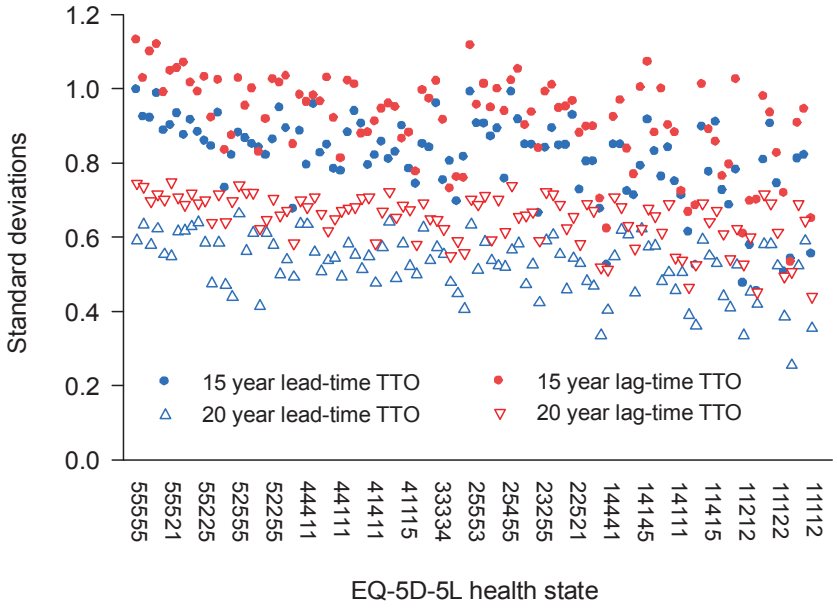

Figure 3.2 Standard deviations of tested TTO methods

The number of non-traders (percentage health-state value = 1) and the distribution of better than death (health-state value > 0) and WTD (health-state value < 0) responses suggest that the lead-time TTO causes respondents to judge health states as being less severe compared to lag-time TTO (Table 2). Interestingly, a large percentage of the respondents valued a state as

equal to being death (health-state value = 0). Also, more than 60% of the respondents used only four iterations or less and about 35% of the sample valued health states at value = 1. The median number of iterations was three for all specifications of TTO.

Table 3.2 Percentage of responses with similar response features (%)

	Health state value = 1	Health state value = 0	Health state value < 0	Lowest value	No differentiation between 10 health states	Respondents using 4 or less iterations
Classic TTO	29.8	21.8	23.9	3.8	11.1	64.6
15 year lead-time TTO	31.4	22.2	25.7	2.3	11.5	65.5
20 year lead-time TTO	39.7	13.2	12.7	2.1	13.4	63.8
15 year lag-time TTO	33.5	17.3	35.7	3.5	10.6	65.5
20 year lag-time TTO	32.8	18.5	29.2	4.2	10.8	64.8

Feasibility

The mode response on statements 1, 2, and 3 of the feasibility questionnaire was 'completely agree' in all five specifications of TTO. For statement 4 ('I found it easy to tell the difference between the health states I was asked to think about'), the mode response was 'neutral', which again was similar for all specifications of TTO. Answer distributions differed for statements 1, 2, and 4 (Kruskall-Wallis test, $p<.001$) but were similar for statement 3 ('I found it difficult to decide on the exact point where Life A and Life B were about the same') (Kruskal-Wallis test, $p=0.43$). For statements 1, 2, and 4, the lead-time TTO in a 20-year time frame was systematically considered slightly more difficult. No clear patterns were discerned between feasibility statements and gender or health of the respondents as measured by EuroQol-VAS.

3.4 DISCUSSION

In this study, classic TTO and novel specifications of the TTO method were compared to explore the impact of the specifications of the task on health-state values. The specifications of the TTO tasks applied in this study systematically affected health-state values and the relative importance of domains of health. In the 20-year time frame, lag-time TTO produced lower values than lead-time TTO, but the results for the 15-year time frame were mixed. Classic TTO values with transformed negative values most resembled those from lag-time TTO in a 20-year time frame. The relative importance of different domains of health was affected by the duration of the impaired health state, but not by the choice for lead-time TTO or lag-time TTO. It appears that respondents considered anxiety/depression to be worse than pain/discomfort only for a duration longer than five years.

Lag-time TTO resulted in lower values than lead-time TTO, and this effect was most pronounced in the 20-year time frame. On average, the effect of time preference (i.e., preferring to be in the best health state immediately) on health-state values is larger than the 'preference for improvement' effect (i.e., that the bad health state will be followed by a good health state). From these findings, it seems that the additive separability assumption of the QALY model (i.e., a health-state value is independent of the health states preceding or following it) does not hold, as health-state values elicited with lag-time TTO are lower than those found with lead-time TTO. We are only aware of one previously published study testing lag-time TTO (91). In that study, lag-time TTO did not produce the same values as lead-time TTO using seven EQ-5D health states. In the present study, which used five years disease time and 10 years lead/lag-time, lead-time TTO values were lower for more severe states than lag-time values. However, in lag-time TTO, more people were willing to trade off time for mild states, though less time on average (i.e., higher mean values) than in lead-time TTO. Thus, the findings were mixed regarding the effect of the specification of TTO on health-state values.

A 1995 study into time preferences and the duration of health states by Dolan and Gudex (90) compared lead-time TTO with lag-time TTO, but without using those exact terms for the TTO specifications. That study had a lead-time TTO and a lag-time TTO with nine years in full health and one year in an impaired health state. For three out of five health states, lead-time median values were lower than lag-time values. Thus, for three out of five health states, respondents considered having the health impairment earlier preferable to having it later (i.e., negative time preferences). Although this finding seemingly contradicts the results presented here, it may well be that individuals obtain more utility from having the health impairment earlier when the duration of the health state is relatively short; that is, they might prefer to get the poor health state 'over with'. This reasoning would be in line with our finding that for the shorter disease duration the difference between lead-time TTO and lag-time TTO is smaller. These results highlight the influence of time preference in TTO tasks, especially when the addition of lead or lag-time increases the considered time horizon. A detailed study into correcting the TTO values from this study for time preferences is currently underway.

The relative importance of different domains was affected by the duration of the health state in the experiment. Although all of the variants tested indicated that the domains 'pain/discomfort' and 'anxiety/depression' caused the largest decrement in health state utilities, the 'anxiety/depression' domain was given more weight for longer durations in all three TTO tasks. If the relative importance of an attribute of a health state depends on its duration, it is unlikely that the specific health-state value decrement can be extrapolated to durations other than the one applied in the TTO task.

Although the instructions for the online TTO were very carefully designed by a team of researchers with experience in TTO, and even though the respondents were given both textual and graphical explanations, the level of task engagement was low in the online setting. Roughly two-thirds of the observations used a maximum of four iterations to achieve the point of indifference. With the iteration procedure applied in this study, this means that two-thirds of the health states were valued at either 1 (one iteration), 0 (two iterations), 0.5/-0.5 (three iterations), or 0.6/-0.6/0.4/-0.4 (four iterations). It is possible that the respondents did not know their preference more precisely than that represented by one of these health-state values. Yet perhaps the level of task engagement could be improved by a different mode of administration. For example, the median number of iterations for classic TTO in a face-to-face interview setting, as reported elsewhere (96), was seven, compared to three in this study. Indeed, TTO data collection via the Internet may produce lower data quality for classic TTO (97), although it has also been argued that it facilitates a good geographical coverage of respondents at a low cost (98). Nonetheless, a comparison of our online study with results from face-to-face interviews does highlight some differences. In a previous Dutch valuation study of EQ-5D-3L, using classic TTO, the value of the worst health state (33333) was -0.39 and that of the second best health state (11211) was 0.897 (75). That range was not reflected in any of the TTO specifications tested here. Excluding participants who claimed not to understand the task, those respondents who did not differentiate between health states or used less than three iterations did not alter this finding. Similarly, the health-state values of the classic TTO, with a transformation for negative values to be bound at -1 as applied in the previous TTO valuation studies of EQ-5D-3L, did not produce negative mean values for any of the health states. Thus, classic TTO also had a rather limited range of values compared to previous EQ-5D valuation studies (26, 75). Unlike Devlin et al. (2012), we did not find notably less non-trading for mild states in lag-time TTO compared to lead-time TTO (91).

Heterogeneity was greatest for lag-time TTO variants, suggesting that respondents' answers differ more in this task than in classic TTO or lead-time TTO, which could be due to several unknown variables. These results seem to indicate that respondents were better able to grasp the lead-time TTO task, leading to less difference in answers. Yet such a conclusion would not fully align with the self-reported feasibility of the task. The latter indicates that lead-time TTO was, on average, considered slightly more difficult than lag-time TTO. The increased variance in the lag-time TTO tasks is thus not solely attributable to understanding of the task.

3.5 CONCLUSION

Lead-time TTO and lag-time TTO yield different health-state values. Differences between lead-time TTO and lag-time TTO seem to be systematic, an observation that requires further study.

Appendix A

Checklist to compare TTO studies

Methodological	
Type of TTO method	Lead-time TTO and lag-time TTO
Incorporated discounting?	No
Total time frame in years	Lead-time TTO: 15 and 20, Lag-time TTO: 15 and 20
Health-state duration	5 years and 10 years
Lead-time length	10 years
Lag-time length	10 years
Ratio of lead/lag-time to health state duration	2:1 and 1:1
Lowest possible value	-1 and -2
Who valued the health state?	General population
Description of health state at which value=1	Full health
Worst health state	55555
What was valued?	EQ5D-5L
Health state selection	D-optimal design
Blocked design?	Yes
Number of health state	100
Health states per respondent	10
Sample size	5208
Valuations per health state	About 100
Procedural	
TTO procedure	Structured iteration
Iteration first question	Value = 1
Iteration second question	Value = 0
Number of attributes	5
Levels per attribute	5
Highest attainable value	1
Warm-up tasks	Discrete choice experiment with similar health states
Visual representation	Side-by-side presentation of alternatives
Interviewer interaction with protocol standardized?	Yes
Mode of administration	Online interviews
Smallest tradable unit	3 months
WTD procedure	Not applicable
Analytical	
Transformation of WTD	Not applicable
Exclusion criteria	Respondent who did not complete the task.

4

When would you rather be ill, now or later?

With Arthur Attema
Health Economics, 2012

ABSTRACT

The Time Tradeoff (TTO) method is frequently used to calculate the quality adjustment of the Quality Adjusted Life Year, and is therefore an important element in the calculation of the benefits of medical interventions. New specifications of TTO, known as 'lead time' TTO and 'lag time' TTO, have been developed to overcome methodological issues of the 'classic' TTO. In the lead time TTO, ill-health is explicitly placed in the future, after a period of good health, while in lag time TTO a health state starts immediately and is followed by a 'lag time' of good health. In this study, we take advantage of these timing properties of lead and lag time TTO. In particular, we use data from a previous study that employed lead and lag time TTO to estimate their implied discounting parameters. We show that individuals prefer being ill later, rather than now, with larger per-period discount rates for longer durations of the health states.

4.1 INTRODUCTION

It is uncertain whether values derived from preference elicitation tasks partly reflect, not just the value of a health state, but also the preferences individuals have for health impairments to occur now or in the future. Time preferences reflect the value given to the timing of an event (99). Any preference for timing, regardless whether it reflects a preference for events occurring sooner or later, has large consequences for the valuation of heath states and by extension for the assessment of the benefit of medical interventions with Quality Adjusted Life Years (QALYs). The quality adjustment of the QALY represents preferences for health states, which can be elicited with the popular Time Tradeoff (TTO) method. Recently, a new specification of this method, called lead time TTO, has been developed, which explicitly places health states in the future, after a so called 'lead-time' of good health (for a complete introduction into the methodology see (39, 42, 43). If individuals derive greater utility from good health now and poor health later, the valuation of a life profile where a health state occurs later in life (lead time TTO) is likely to yield higher utility than a life profile where the same health state, with the same duration, starts in the present, rather than in the future, and is followed, rather than preceded, by good health (lag time TTO).

This paper links two streams of research, one of valuation and one of time preference. Regarding the former, Robinson and Spencer (39) designed their lead time TTO to overcome a methodological flaw of the 'classic' TTO procedure which uses two different procedures for the valuation of health states 'better than dead' and 'worse than dead' and was, therefore, an important innovation in the methodology of valuing health. Indeed, this approach has changed the way researchers presently think of valuing health states better or worse than death. The time preference stream had already been investigated earlier. In 1995, Dolan and Gudex (100) published an article aiming to disentangle time preference from duration effects in TTO. Their experimental approach was an application of lead and lag time TTO, although they did not denote it as such, and their purpose was not to overcome the problem of TTO regarding health states worse than death, which was the main motivation of the study by Robinson and Spencer (2006). While the application of lead time TTO as a valuation method has been picked up by other authors (e.g. (42, 43)), we are unaware of other studies applying lead and lag time TTO for the computation of discount rates.

A first direct measurement of the discounting function for health benefits under certainty was undertaken by Cairns (1992) (101). The method used for this measurement involves the increase of days in ill-health that a respondent is willing to accept in order to obtain a delay of the onset of this spell of ill-health (delay of illness method [DOIM]). Then, one has to specify a particular parametric shape of the discounting function and assume that there is no discounting

within the period of ill-health, which allows one to analytically solve for the discounting parameter. The Direct Method (102) is comparable to the DOIM, but needs no parametric assumptions. Furthermore, it does not have to assume there is no discounting during the period of ill-health, which causes discontinuities in the discounting function. Olsen (1994) (88) proposed to measure discounting using two different horizons in the classic TTO. In particular, this approach predicts lower TTO scores for longer durations because individuals are thought to more easily give up life years that occur farther in the future. However, in addition to having to assume a particular parametric shape of the discounting function, this method is not able to capture discounting for the power function.

In addition to the use of two classic TTOs, one may also consider using one lead and one lag time TTO to elicit time preferences, as was applied by Dolan and Gudex (1995). An advantage of this approach is that it is able to also capture power discounting. Here, this approach is applied and used to present empirical support for the hypothesis that individuals prefer being ill later, rather than now, at least for the observed illness durations. We do so by measuring time preferences using a study in which both lead and lag time TTO were applied (103).

Discounting in TTO

Within the assumptions of the generalized QALY model, TTO scores represent the value of a health state V(Q) by the amount of years, *T-x*, an individual is willing to trade off to regain full health (FH). Thus, for lead time TTO in a 20 year timeframe, with 10 years in FH and 10 years in the impaired health state (α), assuming no discounting, the utility equation is:

$$10V(FH)+10V(\alpha)=xV(FH), \tag{1}$$

which can be solved for V(α), giving:

$$V(\alpha)=\frac{xV(FH)-10}{10}. \tag{2}$$

However, if we assume individuals have a preference for timing, life years will be weighted for time preferences according to the function W(t), resulting in equation 3:

$$W(10)V(FH)+W(10)V(\alpha)=W(x)V(FH). \tag{3}$$

The utility equation for lag time TTO is identical, be it that V(FH) and V(α) are placed in reversed order.

A crucial issue is the identification of the shape of the discount function W(t), or, in other words, to measure how individuals value timing. The discount function can adopt different parametric shapes. Two popular parametric families are the exponential family (implying constant discounting) and the power family (implying hyperbolic discounting, i.e., decreasing discount rates over time). The exponential family can take the following form[5]:

$$W(t) = be^{-rt} + c, \tag{4}$$

where r is the discount rate and t the amount of years. Because W(t) is unique up to scale and location, we can freely fix b and c. For convenience, we set these values such that W(0)=0 and W(20)=1. This holds for $b = -1/(1-e^{-20r})$ and $c = 1/(1-e^{-20r})$. The power function, instead, can be expressed as[6]:

$$W(t) = bt^{s} + c, \tag{5}$$

with the power indicating the degree of hyperbolic discounting. For this function, we obtain W(0)=0 and W(20)=1 for the parameter values $b = (1/20)^{s}$ and $c = 0$.

By substituting one of the discount functions given in equations 4 and 5 into equation 3, we get the discounted utility functions for lead time TTO (and the same can be done for lag time TTO). Then, the value of r or s can be varied until V(α) is the same for lead and lag time TTO. See appendix 1 (online) for the complete derivation of the discounted utility functions.

4.2 METHOD

The linear QALY models predicts equal values for two health profiles which are identical in all aspects but the onset of the ill-health period. In the study by Dolan and Gudex (1995), lead and lag time TTO profiles were presented to respondents, which were identical except for the onset of disease. Given that the linear QALY model predicts equal outcomes for those profiles, the "relative preferences over [the two]… scenarios can be seen as tradeoffs between outcomes occurring at different points in time and thus from these responses each respondent's time preference rate for health could be estimated"(100). Of course, other factors than time preferences may cause differences between lead and lag time TTO, such as loss aversion, because good health is attained after a period of illness in lag time TTO; whereas, in lead time TTO it is lost. Dolan and Gudex tested several TTO specifications, for example a TTO with a total duration of 10 years, with 9 years lead [lag] time and 1 year disease time. We will denote this approach the 'onset of disease method' (ODM).

[5] And W(T)=b*t+c for r=0.

[6] And W(T)=b*ln(t)+c for r=0.

Dataset

We used data from another study, which applied lead and lag time TTO as valuation methods in an online sample of 6222 respondents, reflecting the Dutch general population[7]. Several TTO methods (see table 1) were applied to 100 Dutch EQ-5D-5L health states. Respondents valued 10 health states each, in a randomized blocked design, of one particular specification of TTO. The mean rank of all health states by TTO method is provided in appendix 2 (online).

The EQ-5D-5L consists of 5 dimensions of health (mobility, self care, usual activities, pain/discomfort and anxiety/depression) and five level answer categories, where level '1' represents absence of problems and level '5' represents extreme problems on that particular health dimension. Health states can be described with numbers for ease of use in reporting. A health state description '11211' signifies a health profile with absence of health impairments in all dimensions, represented by '1', except for slight problems in 'usual activities', represented by '2' in the third digit location.

Table 4.1 TTO specifications in the dataset

	TTO type	Total timeframe	Onset of disease	Duration of disease
a	Lead time TTO	15 years	after 10 years	5 years
b	Lead time TTO	20 years	after 10 years	10 years
c	Lag time TTO	15 years	immediately	5 years
d	Lag time TTO	20 years	immediately	10 years

The study reported that, in the 20 year time frame, lag time TTO values were always lower than lead time TTO values. In the 15 year time frame (with only 5 years disease duration rather than 10 years disease duration in the 20 year time frame), this difference was much smaller and in 18 out of 100 health states lead time values were higher than lag time values (i.e., time preferences were negative).

ODM

The ODM offers an 'implied discount rate', as the difference between the two valuation methods is interpreted as an expression of preferences for timing. We applied the ODM to the mean TTO values for each health state, using both exponential discounting and power discounting. Hence, we generated 100 discount parameter estimates for the 15 year time frame, as well as 100 discount parameter estimates for the 20 year time frame for both the exponential discounting and the power model.

[7] The details of this study and of the TTO procedures are presented in Versteegh *et al.*, 2012.

The mean discount parameter ($\bar{r}$ and $\bar{s}$) of each TTO type (a, b, c or d) was applied to all 100 health states of the relevant TTO type. The fit of $\bar{r}$ and $\bar{s}$ was assessed with the root of the mean squared error (RMSE). To clarify the procedure we provide a short example in table 2.

Table 4.2 Example of ODM for a 15 year time frame

	TTO type	EQ-5D-5L Health state (α)	Mean xV(FH)	Utility value V(α) (TTO value - 10 / 5)	Implied r (at which V(α)-a = V(α)-c using exponential discounting)	Mean r	Corrected utility values for mean r	RMSE of corrected values
a	Lead time TTO	52555	10.3	0.1	0.028	0.024	0.014	0.038
c	Lag time TTO	52555	8.9	-0.2			0.05	
a	Lead time TTO	25551	10.1	0.0	0.02		0.07	
c	Lag time TTO	25551	9.0	-0.2			0.03	

4.3 RESULTS

Without discounting, the difference between lead time TTO values and lag time TTO values, expressed in terms of RMSE, was 0.189 for the 15 year time frame and 0.273 for the 20 year time frame. Mean time preferences were positive, for both exponential and power discounting, suggesting that respondents consider profiles of health in which ill-health starts in the future to be more desirable than profiles of health in which ill-health starts immediately. Both exponential and power discounting indicated more per-period discounting for the longer disease duration. Furthermore, both parametric families resulted in an equal but still sizable RMSE, suggesting that time preferences did not fully explain the differences between lead and lag time TTO, or at least not when the same average implied discount rate is used for all health states.

Exponential discounting

For the disease duration of 5 years (a and c from table 1) we found a mean yearly discount rate of 0.015 (sd = 0.016). For the disease duration of 10 years (b and d) we found a mean yearly discount rate of 0.054 (sd = 0.019). RMSE was 0.13 (compared to 0.189 without correcting for mean discount value) and 0.06 (compared to 0.273 without discounting) for the 5 and 10 year disease durations, respectively. There was no clear increasing or decreasing relationship between discount rates and health state severity. As shown in figure 4.1, time preferences were negative for 18 health states for the 5 year disease duration. The health states did not share common features, such as impairments on specific dimensions of health, to explain this phenomenon.

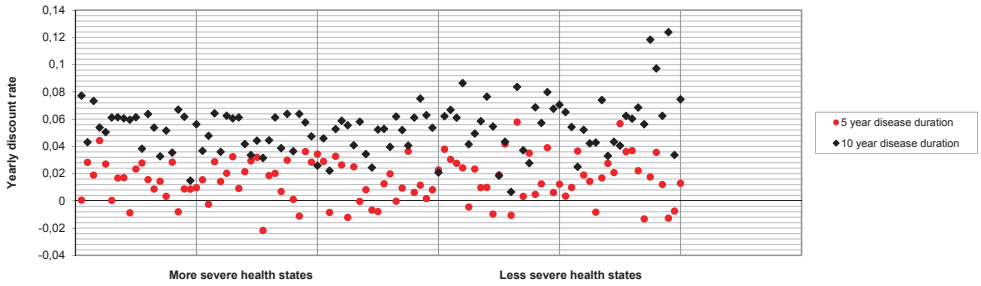

Figure 4.1 Yearly discount rates for all 100 health states

Hyperbolic discounting

We found a mean power coefficient of 0.925 (sd = 0.079) for the 5 year disease duration and a mean power coefficient of 0.697 (sd = 0.089) for the 10 year disease duration. RMSE was 0.129 (compared to 0.189 without correcting for mean discount value) and 0.06 (compared to 0.273 without discounting), respectively. There was no clear relationship between the magnitude of the power coefficients and health state severity. Figure 4.2 shows the hyperbolic discount values for all 100 health states. For the 5 year disease duration, 18 health states were associated with negative time preferences (i.e., powers greater than 1).

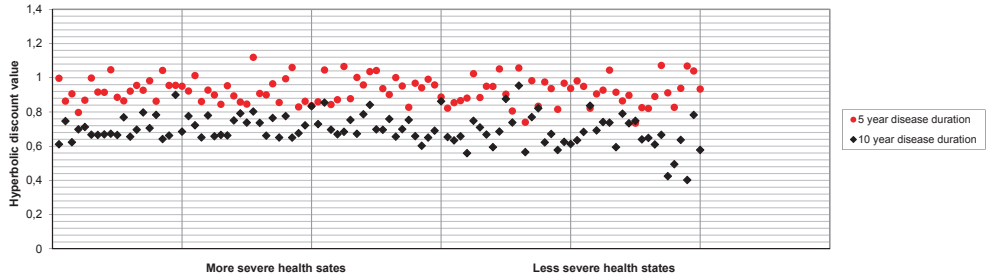

Figure 4.2 Hyperbolic discount values for all 100 health states

4.4 DISCUSSION

On average, individuals displayed positive time preferences for health states, indicating that for the disease durations tested here, respondents preferred ill-health to occur later rather than sooner. These results seemingly contradict the findings of Dolan and Gudex (1995), who found negative discount rates for their disease duration of 1 year, also using the ODM. However, the latter observation may indicate a tendency for lower discount rates when the disease duration is shorter, which is in line with the finding in our study that discounting is higher for a 10 year disease duration than for a 5 year disease duration. In terms of preferences for illness, it

seems that individuals want to get a health state 'over with' if it is short-lasting (negative time preferences (104)), and prefer a delayed onset when duration is longer, at least under certainty. Several attempts have been performed to estimate time preference for health outcomes under certainty. This literature highlights the wide variety of discounting estimates, which are highly influenced by procedural differences. The estimates vary between extremely high discount rates (above 100% per year, (105, 106)) to negative discount rates (100) . Moreover, the type of health state under consideration also seems to affect results. Ganiats *et al.* (2000), for example, (106)found considerable differences between time preference in the case of headaches and chickenpox.

The consistent results found in our own study should thus be considered in the light of the diverse discounting literature which is, in itself, less consistent in findings. Due to the variability in procedures of eliciting discount values, it is difficult to conclude on the exact direction and size of the influence of time preferences on health state valuations, but there seems to be some consistency that they *are influenced* by time preferences.

There are advantages and disadvantages to the use of lead and lag time TTO as a discounting elicitation method. The advantages are that it is a relatively easy task where respondents only have to perform two TTOs, rather than an additional, often cognitively demanding, discounting task. Also, the method for eliciting discounting and valuation of health states is one and the same, which allows a direct estimation of discounted health state values for the health states at hand. One crucial disadvantage, however, is that in the currently applied method the differences in health state values between lead time TTO and lag time TTO are interpreted as differences *due to time preferences*, while there may be other causes for the difference than a preference for timing. For example, a lag time TTO health profile may seem less attractive than a lead time TTO health profile, since death is introduced after a period of good health, which may seem worse than when death occurs after a period of disease, as in lead time TTO. These disadvantages may also explain why the approach of Dolan and Gudex (100) has not been used more widely. This approach, therefore, does not seem to have superior theoretical properties with respect to elicitation of time preference when compared to separate time preference elicitation methods and was used here mainly for reasons of convenience.

Our study was limited by the mode of administration of the TTO study. Versteegh *et al.*, 2013 (103) indicated that the quality data of an online TTO seems to be lower than when interviewer guidance is present, likely because not all individuals properly understand the task, or prefer to complete the task quickly, rather than thoroughly. Conducting this interview in a face-to-face setting would improve data-quality and strengthen our conclusions concerning time preferences. Furthermore, the current design was a between-subject design where respondents

participated in either the lead time TTO or the lag time TTO. As a consequence, we could not compute individual discount rates, but instead performed the discounting analysis on the mean TTO data for the 100 health states. Consequently, the size of the discounting parameters might have differed had we been able to capture individual time preferences and then calculate the mean value of those time preferences.Consider the following example with three individuals. For some health state, Individual 1 returns x=17 in the lead time TTO and x=16 in the lag time TTO, whilst Individual 2 [3] returns x=16 [14] and x=18 [11], respectively. Solving for the exponential discount rate (Eqs. A3-A4) then yields r=0.03 for Individual 1, r=-0.06 for Individual 2, and r=0.05 for Individual 3, for a mean discount rate of r=0.01. However, our approach would consider the mean values of x, 15.67 for lead time TTO and 15 for lag time TTO, giving r=0.015.

Another consequence of the between-subject design was that we could not relate discount values to person characteristics, since a discount parameter was calculated on an aggregate level, using the TTO values of multiple individuals. A within-subject design that allows for possible heterogeneity within the sample and how this heterogeneity impacts upon discount rates would therefore definitely strengthen conclusions.

5

Condition-Specific Preference-Based Measures: Benefit or Burden?

With Annemieke Leunis, Carin Uyl-de Groot & Elly Stolk.
Value in health, 2012

ABSTRACT

Some argue that generic preference-based measures (PBMs) are not sensitive to certain disease specific improvements. To overcome this problem, new condition specific PBMs (CS-PBMs) are being developed, but it is not yet clear how such measures compare to existing generic PBMs.

We generated CS-PBMs from three condition specific questionnaires (HAQ for arthritis, QLQ-C30 for cancer and MSIS-29 for multiple sclerosis). First the questionnaires were reduced in content, then, a Time Trade-off (TTO) study was conducted in the general public (N=402) to obtain weights associated with the dimensions and levels of the new questionnaire. Finally we compared utilities obtained using the CS-PBMs with utilities obtained using EQ-5D in four data sets.

Utility values generated by the CS-PBMs were higher than those of EQ-5D. The HAQ based measure for arthritis proved to be insensitive to comorbidities. The MSIS-29 and QLQ-C30 based measures discriminated comorbidities and side-effect equally well as EQ-5D and were more sensitive than EQ-5D for mild impairments.

The introduction of preference-based measures which are specific to a certain disease may have the merit of sensitivity to disease specific effects of interventions. That gain, however, is traded off to the loss of comparability of utility values and, in some cases, insensitivity to side-effects and comorbidity. The use of a CS-PBM for cost-utility analysis is only warranted under strict conditions.

5.1 INTRODUCTION

A preferred method for generating the quality adjustment required for computation of QALYs is through generic Preference Based Measures (PBMs) such as EQ-5D (26) or the Health Utilities Index (HUI) (31). Some argue that such generic PBMs are not sensitive to certain disease specific improvements. Consequently, the existing PBMs may not always be the best tool to assess the effect of an intervention. To overcome this problem, new condition-specific PBMs have been developed, e.g. for asthma (54) and urinary incontinence (55). Not much is known, however, about how these new instruments compare to generic instruments such as EQ-5D. It is feared that using condition specific Preference Based Measures (CS-PBMs) may lead to the exaggeration of health problems due to a focusing effect, render comparison of utility values impossible, as utilities are derived from different PBMs, and may be insensitive to comorbidities (32, 33). Evidence, however, is scarce. In this study, three CS-PBMs are developed for the purpose of exploring these and other issues, one for arthritis (based on the Health Assessment Questionnaire), one for multiple sclerosis (based on the Multiple Sclerosis Impact Scale 29) and one for cancer (based on the EORTC Quality of Life Questionnaire C30).

A PBM is a questionnaire with a scoring function to weight the responses according to preferences for certain health conditions over others. These preference weights are elicited in studies where respondents are asked to express their preference for a health state, for instance using Time Trade-off, or Standard Gamble. Existing generic PBMs such as EQ-5D and HUI were developed to have a standardized tool to measure health related quality of life for the quality adjustment part of the Quality Adjusted Life Year (QALY). These generic preference-based instruments aim to measure quality of life on a sufficient degree of generality to allow comparisons across conditions. For these instruments the key tradeoff is between generality of the included health dimensions to allow cross-disease comparisons and sensitivity to pick up (relevant) treatment effects (32). The EQ-5D, for example, consists of five items with three levels measuring mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The choice to include only these basic dimensions of health ensures the level of generality that is required for comparison across diseases at the potential cost of losing sensitivity for disease specific complaints. For example, the view is widely held that the EQ-5D is not an appropriate measure to assess quality of life of patients with sensory problems (bad eye sight, or hearing problems), as sensory problems are beyond the scope of health defined by dimensions of the EQ-5D (4). Another perceived problem of EQ-5D is that very mild conditions cannot be adequately assessed using only 3 levels of impairment due to low ceiling sensitivity (107, 108)

The increased use of economic evaluations by health authorities seems to have created a sense of urgency within the health assessment community to deal with the shortcomings of generic

PBMs. In recent years new CS-PBMs have emerged for which the development was motivated by either the absence of generic PBMs in a specific context, or by the judgment that generic PBMs would not be appropriate for a condition. Contrary to generic instruments, a CS-PBM contains dimensions specifically targeted at the affected population. In terms of the tradeoff mentioned above, these instruments are expected to demonstrate superior sensitivity to specific diseases, although this may come at the cost of comparability of utility values across conditions. Due to difference in scope of different instruments, utility values derived from a CS-PBM may not be comparable with those derived from a generic instrument, even though they seem to lie on the same 0 to 1 scale. Although the development of CS-PBMs is valuable for research purpose, for example to better investigate the shortcomings of generic PBMs, there is concern about the application of CS-PBMs in economic evaluations. Unfortunately, empirically founded guidance on how and when to apply CS-PBMs is absent.

There has been little reflection so far on the comparability of the obtained quality of life weights to those obtained from generic PBMs. Specific issues in comparability are described in a recent expert editorial (32). First, CS-PBMs may cause an exaggeration of health problems (reflected by low utility values) due to focusing effects. When the health states in a preference elicitation study consist of a set of disease-related items, rather than general items of HRQoL, the context of the valuation is narrower, possibly leading to lower utility values. The logic behind this hypothesis is that narrow focused items may seem less important when presented in a wider context of general health (e.g. having a cold may seem less severe when presented alongside problems with mobility), but may seem quite problematic when presented separately. This may result in a downward bias on preference values when compared to generic PBMs. Second, a CS-PBM might have difficulty capturing co-morbidities as the focus is on disease related items. This may result in an upward bias on utility values. Furthermore, developing a CSPBM is not a clear cut exercise. Researchers face many decisions, like the reduction of items in a questionnaire, the selection of health states (how many and which?) that have to be valued to develop a PBM (4), on the valuation method (e.g. Time Trade-off or Standard Gamble?), and on the modeling approach. How these decisions are dealt with may differ per study which decreases comparability.

The primary aim of this paper is to provide empirical evidence about the comparability of CS-PBMs and generic PBMs. To do so, three CS-PBMs were developed from existing questionnaires. The values generated by these CS-PBMs were then compared to EQ-5D values for the same patient samples. By providing empirical evidence we hope to provide a better understanding of the effects of using CS-PBMs and contribute to development of guidance for their use. This is important, as it can be expected that in the nearby future more cost utility analyses will contain utilities based on conditions specific measures.

5.2 METHOD

Questionnaires for CS-PBM development

The CS-PBMs were developed from the Health Assessment Questionnaire (HAQ, (109)), the Multiple Sclerosis Impact Scale 29 (MSIS-29 (110)) and the Quality of Life Questionnaire for Cancer (QLQ-C30 (111)). These instruments were selected based on expert advice and commonality of use within clinical settings. The HAQ is a widely used questionnaire in rheumatology to measure functional abilities using 20 items with four levels spread across 8 domains (dressing, rising, eating, walking, hygiene, reach, grip and usual activities). The scale has been shown to be unidimensional (112). The MSIS-29 measures the impact of multiple sclerosis on a physical and psychological dimension. Dimensionality of the subscales has been confirmed using Rasch analysis (113). The QLQ-C30 (version 2) is a cancer specific questionnaire consisting of 30 items. These items cover five functional scales, nine symptom scales and a global health status scale .These questionnaires were chosen as they differ in scope and because EQ-5D data was available for the purpose of comparing results. For MSIS-29 it has been shown that the physical scale is better capable of discriminating among sub categories of the clinically assessed Expanded Disability Status Scale (EDSS) than EQ-5D (114). There was no evidence known to us on a lack of responsiveness of EQ-5D or the superiority of the condition specific measure HAQ and QLQ-C30 in arthritis or cancer.

Reducing the content of the questionnaires

Developing a PBM from an existing questionnaire does not lead to an entirely new instrument but attaches weights to some of the items of the existing questionnaire. Such an approach generally requires a method to reduce the questionnaire content as only a limited number of items can be valued in a preference elicitation study (4). Typically only a fraction of the total amount of all theoretically possible health states is valued. Through modeling techniques the values for the remainder of the health states are estimated.

The optimal number of items in a health state was considered to be in the order of 5 to 9 items, as more items may cause difficulties for respondents in the valuation study (4). The HAQ, MSIS-29 and QLQ-C30 respectively contain 20, 29 and 30 items, so reduction of content was required. Relevant and well-functioning items from the questionnaires were selected using the following criteria proposed by others (115): *i)* the item had to fit the Rasch model, *ii)* the item had to meet basic psychometric criteria and *iii)* the selected items had to be approved by a clinical expert. Four datasets were available for these analyses: the Rotterdam Early Arthritis CoHort (REACH) for the HAQ (N=738), the Multiple Sclerosis Risk Sharing Scheme Monitoring Study (N=1295) for the MSIS-29 and the HOVON 24 (pooled N=716) and HOVON 25 (pooled N=789) trials for the QLQ-C30. The dataset characteristics are described in detail

in Versteegh et al. (2011) (114). A set of a priori criteria were used to determine which items were suitable for the health state description (115, 116). As it was expected that neither of these criteria could be sufficient on its own, the three criteria were employed 'side by side' (i.e. no hierarchical order).

Criterion 1: fit to the Rasch model

Rasch analysis was used to test the psychometric validity of a scale and to identify well-functioning items. The Rasch model assumes that the probability of scoring level λ on item *i* is a logistic function of the relative distance between the item location (how much disability it represents) and the respondent location (how disabled the patient is) (117).

The main performance criteria within the Rasch model were whether the item: *i*) has ordered thresholds (having more of the latent trait θ results in endorsing a higher level answer category (118)); *ii*) fits the Rasch model (fit residual <2.5 and non-significant bonferroni adjusted probability); *iii*) combined scale fits the Rasch model (described by a nonsignificant item-trait interaction Chi-square probability (118)) and *iv*) shows no differential item functioning (DIF). After each single scale amendment the analysis was rerun for the remaining items. Rasch analysis was performed on the dimensional structure originally suggested by the questionnaires.

Criterion 2: psychometric properties

Psychometric criteria were laid alongside the Rasch results to come to a final selection of items amenable for valuation. The functioning of the items was tested by investigating the loading of items on factors identified by factor analysis; missing data; internal consistency of items with its scale score; distribution of the responses on an item including floor and ceiling effects, regression coefficients between a general health indicator and an item. Psychometric analyses were applied to the full datasets.

Criterion 3: expert opinion

The selected items from the questionnaires were presented to experts in the respective fields. Experts from the Erasmus Medical Centre and the VU Amsterdam Medical Centre were consulted to gain insight in important aspects of the disease and to evaluate the result of the previous selection process.

Health state selection

Even after data reduction the selected set can still generate an enormous amount of possible health states, therefore, a fractional factorial design was favored over a full factorial design.

The QLQ design was a level-balanced design, meaning that all levels of each item occurred with the same frequency. Within the balanced design health states covered the entire spectrum of severity, measured by averaging the item levels of a health state. For the MSIS-29 and the HAQ, items and levels were selected with an orthogonal main effects plan (OMEP) as is applied in other studies (30, 119) to ensure zero statistical correlation between the attributes. The set was complemented with a selection of the most observed health states (4 or more observations) over the severity range of the questionnaire. TTO values estimated with additive main-effect models (one based on the OMEP states and one based on the OMEP and the most observed states) were compared to the observed TTO values of the most occurring states using standard predictive performance measures like mean absolute error (MAE) to see if the addition of these states led to improved prediction of the most frequently occurring states.

The final design was blocked. In such a design respondents value a number of health states which belong to the same 'block'. The mean severity of the combination of items in a block was similar and measured through summing the level scores of the items in a block.

Health state valuation with Time trade-off method

The preferences of a sample of the general public were elicited through a Time trade-off exercise (TTO) for each of the selected health states of the questionnaires. To optimize comparability to generic PBMs the CS-PBM health states were valued with the same TTO protocol, the same computer assisted personal interview tool, the same procedure to measure states 'worse than dead' and the same rescore procedure of negative values (negative TTO scores were rescaled to have a range between -1 and 0 with (-t/-x-t) as was adopted in the Dutch EQ-5D valuation study (75). Unlike the Dutch EQ-5D valuation study, this study was performed in group sessions, which has previously been shown to produce comparable TTO results (120).

The TTO exercise was self-administered through a digital tool for TTO elicitation (computer assisted personal interviews) in groups with about 12 to 25 respondents per session. Each session was supervised by 3 to 4 researchers to offer assistance if needed. Prior to the task, respondents received 30 minutes of instructions by researchers MV or AL including examples of the TTO computer program projected on a large screen. The task was piloted by MV and AL in a sample of 18 respondents to ensure the introduction, the computer program and the organization of the task were feasible.

The three questionnaires were presented separately in the TTO exercise and in all possible orders (e.g. first HAQ, then MSIS then QLQ). Within the TTO exercises, health states were presented random to individuals.

Respondents

Respondents were selected by a marketing agency which requited a sample resembling the Dutch general population in age, gender and education. Respondents were approached by phone and asked if they were interested in contributing to a task to value descriptions of health states. Respondents received a financial reward of € 35, - upon completion of the three TTO exercises. Respondents were removed from the analyses when the results indicated they valued the majority of logically worse states higher than logically better states in a set (i.e. HAQ state 11112 is logically better than HAQ state 14444).

Modeling of the TTO values

Once the TTO study had been performed, the preference values observed for the selected health states were used to estimate values for all potential health states through statistical modeling. Because individuals value more than one health state there are multiple observations for each individual. Random effects models were estimated to assess how the predictors (the items and their levels) influence the dependent variable (the mean observed TTO value). In these random effects models, the item levels were treated as dummy variables with dummy coding. The constant term was treated as an additional decrement for having any item level other than the base case ('no problems'), which is similar to the EQ-5D model. The values predicted by this random effects model will be referred to as the preference -based measure results (e.g. HAQ-PBM). Models were required to have significant predictors and worse scores on the levels ought to be represented by larger utility decrements. Model performance was assessed by comparing the mean absolute error (MAE) of observed and predicted values. Models were estimated until meeting those criteria. Only the most parsimonious models are presented. To keep optimal comparability between the developed CS-PBMs, models were estimated from the items only, without interaction effects or a 'worst-value' dummy variable which is 1 for every item on the lowest level. Interaction effects were not estimated because the study design was a main effects design.

Hypotheses and analyses

To investigate the properties of preference-based measures developed from existing questionnaires several hypotheses were tested. First, it was tested whether the TTO values could be successfully modeled. For HAQ and MSIS the TTO random effects models were fitted on both the full dataset (including 'most observed' health states) and on the subset consisting of health states originating from the OMEP. This was done to test whether an OMEP alone is sufficient to estimate the utility values of the most occurring health states. Second, it was investigated whether CS-PBMs yielded lower mean utility values than a generic measure, which was hypothesized to reflect that a downward bias on utility values resulting from a focusing effect might outweigh the upward bias on utility values resulting from a narrower scope of the CS-PBM. Third it was tested

with Wilcoxon rank-sum tests, to account for the non-normal distribution of utility values, whether the developed CS-PBMs had a more narrow focus and were therefore less sensitive to comorbidities (in arthritis and multiple sclerosis datasets) or side-effects (Non-Hodgkin's lymphoma dataset) than EQ-5D. Side-effects had World Health Organization performance status 2 or higher, representing the inability to carry out work activities due to the condition. Fourth, we assessed discriminative properties of the new measures using clinical indicators. For arthritis the Disease Activity Score-28 was used, which is based on a count of tender joints and the erythrocyte sedimentation rate. It distinguished between high, moderate and low disease activity, and remission. For multiple scleroses the Expanded Disability Status Scale (EDSS) was used which, when rounded to integers, distinguishes 11 categories of increasing disability. For cancer we used the WHO performance status score (or ECOG) which distinguishes 6 categories, from 0 to 6 (death). Lastly, responsiveness was measured in the cancer population using effect-size (Cohen's *d*) and mean change in the cancer population, for which follow-up measurements were available in the data set.
All results were compared to utilities of the Dutch EQ-5D tariff (75).

Software
For Rasch analysis the RUMM2020 software (Rumm Laboratory Pty Ltd) was used. Psychometric analysis was performed in SPSS 17.0 (SPSS Inc.) and all hypothesis testing and modeling efforts in STATA 11.0 (StataCorp. 2009).

5.3 RESULTS

Item and level selection
The selected items per questionnaire are presented in Table 1, and all met the criteria of the Rasch analysis, psychometric analysis and expert opinion. The full results of the selection of items and levels are described in appendix A. A table of the results of the Rasch analysis is presented in online appendix B.

Table 5.1 Items selected for the TTO valuation exercise

HAQ-DI	MSIS-29	QLQ-C30
• HAQ1 Stand up from a straight chair • HAQ2 Walk outdoors on flat ground • HAQ3 Get on / off toilet • HAQ4 Reach and get down a 5-pound object (such as a bag of sugar) from just above your head • HAQ5 Open car doors	• MSIS1 Problems with your balance • MSIS2 Being clumsy • MSIS3 Limitations in your social and leisure activities at home • MSIS4 Difficulties using your hands in everyday tasks • MSIS5 Having to cut down the amount of time you spent on work or other daily activities • MSIS6 Feeling mentally fatigued • MSIS7 Feeling irritable, impatient or short tempered • MSIS8 Problems concentrating	• QLQ1 Trouble taking a long walk • QLQ2 Limited in doing either your work or other daily activities • QLQ3 Have you had pain • QLQ4 Have you felt nauseated • QLQ5 Were you tired • QLQ6 Difficulty in concentrating on things • QLQ7 Did you worry • QLQ8 Has your physical condition or medical treatment interfered with your social activities

Resulting study design

Given the many items and levels in the study we chose a fractional factorial blocked design. Health states were presented in blocks, so one individual values one block containing several health states. The design is summarized in Table 2.

Table 5.2 TTO study design following item selection

	HAQ	MSIS	QLQ
Number of items	5	8	8
Total amount of health states to be valued	56	100	105
States identified by OMEP (used in study after fold-over)	15 (30)	32 (64)	n/a
Number of most occuring states included	26	36	n/a
Number of states valued by one individual (total = 33)	8	10	15
Number of blocks[1]	7	10	7
Expected number of observations per health state (N=400 / number of blocks)	57	40	57

[1] One block consist of a number of states and all of the states in one block are valued by one individual

Data quality

Four hundred two respondents participated in the computer assisted TTO study and resembled the Dutch population (Table 3). Respondents were excluded from the analyses because they had valued the majority of logically better states lower than logically worse states in one block (8 exclusions for HAQ, 17 for MSIS and 7 for QLQ). Average time to value one health state in the TTO exercise was about 1 minute. Total time per block was highest for QLQ (15 health states, about 12 minutes), followed by MSIS (10 health states, 10 minutes) and HAQ (8 health states, 8 minutes). Although two separate researchers took turns in holding the introductory talks this

did not bias the TTO responses (Wilcoxon rank test p>0.05). On average, women had higher utility values (t-test, $p<0.00$) for all three questionnaires.

Modeling

TTO values were modeled for each of the three questionnaires with random effects mean prediction models. For the HAQ, only using the OMEP based health states had too much variation in TTO scores between respondents to estimate a model with significant predictors and logical negative signs for each of the dummy variables (increasing negative decrements per item level of severity). Estimating the model on all the available data (thus including the 'most observed' states) yielded a well-functioning mean prediction model. The pre-final MSIS-29 model had insignificant predictors for three variables MSIS3; the pre-final QLQ-C30 model had insignificant predictors for two variables. In all instances merging the levels with the adjacent categories resolved the problem. Model characteristics are summarized in Table 4 and full models are presented in Table 5.

Table 5.3 Respondent characteristics

	TTO study sample	Dutch population norms[1]
N	402	-
Gender M/F	46% / 54%	49.5% / 50.5%
Age		
mean (SD)	45 (15.5)	40.1
min-max	15-76	-
Agegroup		
<20	4.8	23.7
20-40	37.6	25.3
40-65	46.9	35.7
65-80	10.4	11.4
>80	0.3	3.9
Education		
High	34	27
Medium	35	31
Low	25	33
Missing / Else	6%	9%
Mean (SD) time to complete TTO		
HAQ 8 states	8 min (4.6min)	-
MSIS 10 states	10 min (5.8min)	-
QLQ 15 states	12.7 min (5.8min)	-

[1] Statistics Netherlands, 2009

Table 5.4 Final random effects model characteristics

	HAQ-PBM†	MSIS-PBM*	MSIS-PBM†	QLQ-PBM
Random effects mean models				
R†	0.94	0.68	0.78	0.88
MAE	0.028	0.034	0.04	0.033
MAE most observed states	0.022	0.057	0.043	-
Illogical sign or order of variables	0	0	0	0
Insignificant predictors	0	0	0	0
Possible range	0.32 - 1	0.40 - 1	0.42 - 1	0.34 - 1

* Model based on states from the orthogonal design † Model based on states from the orthogonal design and the most observed states

Table 5.5 Coefficients of random effects models with TTO value as dependent variable

HAQ-PBM			MSIS-PMB†			QLQ-PMB		
	Coefficient	Std. Err.		Coefficient	Std. Err.		Coefficient	Std. Err.
haq1_2	**-0.005**	0.001	ms1_2	**-0.016**	0.003	qlq1_2	**-0.027**	0.001
haq1_3	**-0.031**	0.002	ms1_3	**-0.043**	0.003	qlq2_2	**-0.020**	0.002
haq1_4	**-0.121**	0.002	ms1_4	**-0.089**	0.003	qlq2_3	**-0.047**	0.002
haq2_2	**-0.029**	0.001	ms2_2	**-0.018**	0.003	qlq2_4	**-0.068**	0.002
haq2_3	**-0.091**	0.002	ms2_3	**-0.047**	0.003	qlq3_3	**-0.079**	0.002
haq2_4	**-0.144**	0.002	ms2_4	**-0.047**	0.003	qlq3_4	**-0.213**	0.001
haq3_2	**-0.042**	0.001	ms3_3	**-0.055**	0.002	qlq4_2	**-0.018**	0.002
haq3_3	**-0.055**	0.002	ms3_4	**-0.071**	0.002	qlq4_3	**-0.055**	0.002
haq3_4	**-0.213**	0.002	ms4_2	**-0.061**	0.002	qlq4_4	**-0.089**	0.002
haq4_2	**-0.022**	0.001	ms4_3	**-0.101**	0.003	qlq5_2	**-0.021**	0.002
haq4_3	**-0.041**	0.002	ms4_4	**-0.108**	0.003	qlq5_3	**-0.031**	0.002
haq4_4	**-0.074**	0.002	ms5_2	**-0.032**	0.003	qlq5_4	**-0.037**	0.002
haq5_2	**-0.016**	0.001	ms5_3_4*	**-0.057**	0.002	qlq6_2	**-0.004**	0.002
haq5_3	**-0.038**	0.002	ms6_2	**-0.020**	0.003	qlq6_3	**-0.039**	0.002
haq5_4	**-0.044**	0.002	ms6_3	**-0.035**	0.003	qlq6_4	**-0.052**	0.002
Constant	**0.918**	0.002	ms6_4	**-0.059**	0.003	qlq7_3	**-0.009**	0.002
			ms7_3	**-0.024**	0.002	qlq7_4	**-0.047**	0.002
			ms7_4	**-0.038**	0.002	qlq8_2	**-0.008**	0.002
			ms8_2	**-0.037**	0.003	qlq8_3	**-0.041**	0.002
			ms8_3	**-0.059**	0.003	qlq8_4	**-0.060**	0.002
			ms8_4	**-0.073**	0.003	Constant	**0.944**	0.002
			Constant	**0.959**	0.005			

* Both ms5_3 and ms5_4 have the same decrement † Msis model with most observered health states included. Bold coefficients p<0.01

When the MSIS-29 prediction model was based on all the states (thus including the most observed states) the prediction error for the most observed states was reduced (MAE=0.043 compared to MAE=0.057). When the MSIS-29 TTO values were modeled without the 'most observed' states the utility values were generally higher which caused that the utility values of some of the 'most observed' states were overestimated (Fig. 5.1).

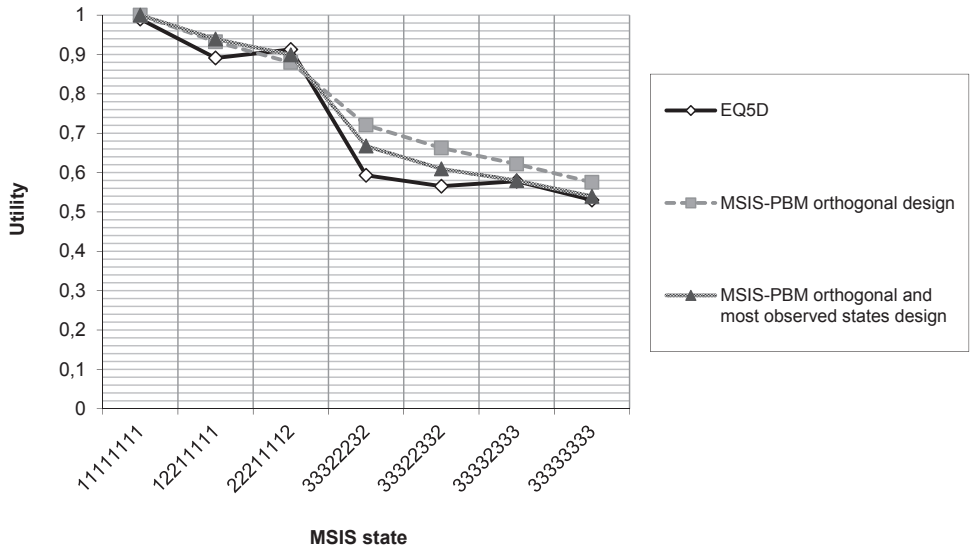

Figure 5.1 Including most observed states influences health state values

Comparability of mean utility values

In the four datasets, the developed CS-PBMs based on the models presented in table 5 produced higher mean utility score for patients than EQ-5D (Table 6). Especially the HAQ-PBM (mean = 0.91) had a much higher mean utility value than EQ-5D (mean = 0.68). Furthermore, the difference between the EQ-5D in arthritis and mean EQ-5D in MS was 0.06 while the differences between the HAQ-PBM and the MSIS-PBM are 0.24. The QLQ-C30-PBM based utility values had the highest correlation with EQ-5D utility values. Both the MSIS-PBM and the QLQ-PBM, however, have increased sensitivity compared to EQ-5D. Where EQ-5D scores full health (a utility value of 1) the MSIS-PBM and the QLQ-PBM report decrements in utility for respectively 99 and 185 patients (Table 7).

Table 5.6 Comparison of utility values derived from the new PBM measures with EQ-5D and SF-6D

	HAQ	MSIS	QLQ_MM‡	QLQ_NH‡
N	738	1295	716	789
Mean utility (SD) [range]				
EQ5D	0.68 (0.23) [-0.134 - 1]	0.62 (0.26) [-0.22 - 1]	0.74 (0.21) [-0.058 - 1]	0.73 (0.26) [-0.33 - 1]
SF6D	0.66 (0.10) [0.37 - 1]	-	-	-
PBM*	-	0.69 (0.13) [0.40 - 1]	0.84 (0.09) [0.44 - 1]	0.82 (0.11) [0.34 - 1]
PBM†	0.91 (0.09) [0.57 - 1]	0.67 (0.14) [0.42 - 1]	-	
Intraclass correlations				
EQ5D-PBM*†	ICC = 0.45	ICC = 0.62	ICC = 0.64	ICC = 0.67

* Model based on states from the orthogonal or balacanced design. † Model based on states from the orthogonal design and the most observed states. ‡ MM= Multiple Myeloma, NH = Non-hodgkin.

Table 5.7 MSIS-PBM and QLQ-PBM have increased sensitivity at the ceiling of EQ-5D

Total sample size		EQ5D =1	EQ5D <1	Worst state for which EQ-5D = 1
738	HAQ-PBM < 1	n = 7	-	21211
	HAQ-PBM = 1	-	n = 252	
1295	MSIS-PBM < 1	n = 99	-	33111222
	MSIS-PBM = 1	-	n = 2	
1505	QLQ-PBM < 1	n = 185	-	24334324
	QLQ-PBM = 1	-	n = 4	

Comorbidities and side-effects

The HAQ-PBM could not discriminate between patients with and without comorbidity (other vascular disorders and psychiatric disorders) when EQ-5D could (Table 8). For arthritis patients with diabetes, hypercholesterolemia or thyroid disease the HAQ-PBM showed higher utility values for individuals with the disorder while EQ-5D signaled the expected direction of differences. The MSIS-PBM also showed higher utilities for patients with asthma and high blood pressure (rather than without) but this was concordant with the differences indicated by EQ5D. Both the MSIS-PBM and the EQ5D picked up significant differences between multiple sclerosis patients with and without depression ($p<0.05$). In the Non-Hodgkin's lymphoma dataset, patients with side-effects and infections as result of treatment had lower ($p<0.05$) utility values in both EQ-5D and QLQ-C30 than patient without side-effects and infections, except for hair loss. All significant differences were at least half a standard deviation except for comorbidity 'depression' in the MS dataset and 'other side-effects' in the Non-Hodgkin's lymphoma dataset.

Discriminative ability and responsiveness

Utilities of all instruments decreased with an increase of severity as assessed by the clinical indicator (Table 9). The utilities of the HAQ-PBM, however, failed to distinguish between low and moderate disease activities. EQ-5D did so accurately. As has previously been shown, EQ-5D was unable to distinguish between categories 3, 4 and 5 on EDSS (114). This signifies the inability of EQ-5D to distinguish between fully ambulatory MS patients (EDSS 3) and patients whose disability is severe enough to impair full daily working activities (EDSS 5). The MSIS-PBM, of which the physical scale was known to be sensitive to changes between level 3 4 and 5, did pick up the deterioration in health. Neither the QLQ-PBM nor the EQ-5D adequately reflected the deterioration between level 0 and level 1 of the WHO performance status.

Table 5.8 Comorbidities and side-effects by PBM and EQ-5D

	HAQ-PBM				EQ5D			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Diabetes	0.92 (0.08)	0.9	0.89 (0.10)	0.89	0.66 (0.24)	0.79	0.71 (0.16)	0.78
Hypercholesterolemia	0.90 (0.09)	0.88	0.89 (0.10)	0.89	0.57 (0.27)	0.65	0.72 (0.15)	0.78
Thyroid disease	0.90 (0.10)	0.89	0.89 (0.10)	0.89	0.60 (0.29)	0.71	0.72 (0.16)	0.78
Other cardiac disease	0.86 (0.05)	0.88	0.89 (0.10)	0.89	0.60 (0.21)	0.68*	0.72 (0.15)	0.78*
Psychiatric disorder	0.84 (0.13)	0.89	0.89 (0.10)	0.89	0.54 (0.30)	0.67*	0.72 (0.16)	0.78*
	MSIS-PBM				EQ5D			
	Comorbidity		No comorbidity		Comorbidity		No comorbidity	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Depression	0.61 (0.12)	0.59*	0.68 (0.14)	0.68*	0.54 (0.26)	0.64*	0.63 (0.26)	0.67*
Asthma	0.67 (0.14)	0.71	0.67 (0.14)	0.68	0.63 (0.22)	0.68	0.62 (0.26)	0.67
HPB	0.68 (0.14)	0.73	0.65 (0.13)	0.68	0.64 (0.25)	0.64	0.60 (0.26)	0.67
	QLQ-PBM				EQ5D			
	Side-effects		No side-effects		Side-effects		No side-effects	
	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median	Mean (sd)	Median
Neurotoxicity†	0.76 (0.09)	0.76*	0.81 (0.09)	0.83*	0.56 (0.27)	0.65*	0.73 (0.25)	0.81*
Hair loss†	0.8 (0.08)	0.79	0.81 (0.1)	0.83	0.7 (0.23)	0.78	0.72 (0.26)	0.79
Nausea†	0.73 (0.13)	0.74*	0.81 (0.09)	0.83*	0.56 (0.31)	0.65*	0.72 (0.25)	0.81*
Other side-effects†	0.79 (0.09)	0.8*	0.81 (0.09)	0.83*	0.66 (0.23)	0.69*	0.72 (0.26)	0.81*
	Infection		No infection		Infection		No infection	
Ear / Nose / Throat[1]	0.72 (0.11)	0.72*	0.81 (0.09)	0.83*	0.42 (0.32)	0.31*	0.73 (0.25)	0.81*

* = significant difference between comorbidities / no comorbidities at ($p<.05$) Wilcoxon sum rank test. † WHO grade ≥ 2

Table 5.9 Discriminant validity

	HAQ-PBM		EQ-5D		N
DAS28	mean	sd	mean	sd	
Remission	0.98	0.04	0.76	0.20	11
Low DA	0.90	0.08	0.70	0.25	15
Moderate DA	0.90	0.09	0.67	0.22	70
High DA	0.83	0.07	0.51	0.29	27
	MSIS-PBM		EQ-5D		N
EDSS	mean	sd	mean	sd	
0	0.80	0.14	0.81	0.22	35
1	0.78	0.14	0.78	0.23	74
2	0.73	0.14	0.72	0.23	262
3	0.68	0.14	0.63	0.25	206
4	0.66	0.13	0.63	0.23	248
5	0.63	0.10	0.64	0.19	103
6	0.60	0.11	0.54	0.25	201
7	0.58	0.11	0.46	0.27	78
8	0.57	0.07	0.40	0.31	17
9	0.47	0.07	0.09	0.10	5
	QLQ-PBM		EQ-5D		N
WHO	mean	sd	mean	sd	
0	0.83	0.11	0.75	0.25	356
1	0.83	0.10	0.76	0.24	304
2	0.80	0.11	0.69	0.24	96
3	0.71	0.10	0.37	0.27	27

The QLQ-C30 was, in terms of effect-size measured with Cohen's *d*, at times more, and at times less sensitive to changes over time (table 10). However, the absolute differences indicated that even when the QLQ-PBM had a larger mean difference relative to the standard deviation, the EQ-5D still reported larger mean change scores.

Table 5.10 Responsiveness of utilities in non-hodgekin sample

	Cohen's d		Mean change	
Follow-up	QLQ-PBM	EQ-5D	QLQ-PBM	EQ-5D
2nd treatment cycle	0.13	0.17	0.02	0.05
4th treatment cycle	0.02	0.08	0.00	0.02
6th treatment cycle	-0.09	-0.06	-0.01	-0.01
3 months follow-up	0.33	0.22	0.03	0.06
6 months follow-up	0.25	0.10	0.02	0.02
10 months follow-up	-0.01	-0.09	0.00	-0.02
18 months follow-up	0.00	0.19	0.00	0.04

5.4 DISCUSSION

This study developed three CS-PBMs from existing questionnaires HAQ, MSIS-29 and QLQ-C30 to provide evidence concerning comparability of CS-PBM derived utility values with generic PBM derived utility values. CS-PBMs had different mean utility values within a disease and did not report equal differences in mean utility values between diseases. The CS-PBMs in this study did not seem to exaggerate health problems, but rather reported higher mean values. Capturing comorbidities and along that line: side-effects of interventions appeared problematic for the HAQ-PBM, but not for MSIS-PBM and QLQ-PBM. The MSIS-PBM and QLQ-PBM were more sensitive to very mild impairment than EQ-5D. The physical scale of the MSIS-29 questionnaire is known to be more sensitive in discriminating between clinical categories in multiple sclerosis than EQ-5D. The MSIS-PBM, derived from the MSIS-29, also has better discriminatory properties.

As the mean utility values of all three CS-PBMs were higher than those of generic instruments, it seems that a potential downward bias of a focusing effect may be smaller in size than the upward bias that results from a narrower scope of the condition specific measures. This is most clearly seen in the performance of the HAQ-PBM, which is developed from the HAQ-Disability Index (HAQ-DI) which measures functional ability (121). Consequently the HAQ-PBM indicates the utility decrements associated with these functional (dis)abilities. In the HAQ-PBM there is no dimension such as 'pain' or 'psychological state'. Since pain is a frequently occurring symptom in arthritis, it is not surprising that the mean utility value of the early arthritis cohort as measured by the HAQ-PBM is much higher than the mean utility value of the generic instruments; any additional utility decrement besides functional disabilities, such as pain, is not captured directly, if at all. In case of the HAQ this result could have been anticipated based on the fact that the HAQ-DI aims to offer a unidimensional assessment of functionalities and does not attempt to measure other dimensions of health since these are captured by other instruments that are

part of the minimum dataset internationally agreed on. The unidimensionality of the HAQ caused some problems in the valuation task. As all items aim to measure the same underlying latent variable (functional ability) they are highly related. OMEP generated states have favorable statistical properties, but do not consider the sensibility of the combination of item levels. Consequently, one health state in the valuation study consisted of the counter intuitive combination "able to get up from a chair" and "not able to get up from the toilet". This particular state caused confusion with some of the respondents.

The HAQ-DI does not intend to form a comprehensive assessment of relevant disease specific health outcomes in patients with rheumatoid arthritis, and therefore could be rejected as offering a suitable basis for development of CS-PBMs. The large deviations in mean utility values presented in this study between the HAQ-PBM and EQ-5D support this view. More generally, it can be concluded that instruments with a narrow scope, often identifiable through inspecting items or dimensions, are unsuitable as a base for CS-PBMs used for resource allocation.

The perceived insensitivity of existing generic instruments is an important motive for developing CS-PBMs. In this study sensitivity of the CS-PBM and EQ-5D was compared by investigating ceiling effects and discriminative ability of the instruments between patients with and without comorbidity or side-effects. A ceiling effect found in EQ-5D for mild impairments was not found in the MSIS-PBM and QLQ-PBM (table 7). One reason for this difference may be the descriptive system of the questionnaires: the three level system of EQ-5D might result in a lower likelihood of reporting problems than the 4 level systems of the CS-PBMs. Nevertheless, using CS-PBMs did not result in an exaggeration of health problems on average when compared to generic instruments in this study. Rather, the mean utility value of MSIS-PBM and QLQ-PBM was higher than EQ-5D. This may be a reflection of the smaller range in obtainable utility values, which skews the average upwards. Bad EQ-5D health states reflect very poor health, which is perhaps not captured in MSIS-PBM and QLQ-PBM. Indeed, the negative range of utility values as produced for EQ-5D has rarely been reproduced for other instruments. EQ-5D, MSIS-PBM and QLQ-PBM performed equally well in distinguishing patients with comorbidities / side-effects from patients without it. Only the HAQ-PBM performed poorly in this aspect. Interestingly, the MSIS-PBM and the QLQ-PBM displayed equal discriminative ability as EQ-5D despite having a much smaller total scale size due to a higher 'floor' (i.e. the lowest attainable value).

Superiority of CS-PBMs compared to EQ-5D in regard to their discriminative ability is not demonstrated for HAQ-PBM and equivalence has been shown for QLQ-PBM. MSIS-PBM showed better discriminative properties than EQ-5D in EDSS subcategories. With additional evidence on known-group differences this could prove the MSIS-PBM to be a contribution

to cost-utility analyses. The original preference-based questionnaire MSIS-29 was the only measure for which empirical evidence indicated better discriminative properties than EQ-5D in a multiple sclerosis data sets.

While a CS-PBM may have desirable statistical properties, such as expressed in effect-size or the ability to identify significant differences between groups with or without-side effects, partly due to a small standard deviation of mean values, these properties may not be reflected the absolute size of differences in utility values between groups. This has consequences for QALY computation. Imagine a new drug that reduces nausea from cancer treatments. Using the figures from table 8, the population not having nausea would have a higher utility with an effect-size (Cohen's *d* with pooled standard deviations) of 0.57 for EQ-5D but a larger 0.73 for QLQ-PBM. The absolute difference, however, would be 0.16 for EQ-5D and 0.08 for QLQ-PBM. An implication of these results is that if a CS-PBM is developed in order to increase sensitivity compared to EQ-5D, statistical sensitivity is not a sufficient criterion.

Rather than due to concerns about the sensitivity of an existing generic PBM, a CS-PBM may also be developed because a PBM was not administered in, for example, a clinical trial. In this case one could also choose to use the variation in responses on a condition specific measure to estimate what a generic utility instrument like EQ-5D would have been had it not been absent, a process called mapping (58). It is important to reflect on the question which strategy for deriving utilities from a disease specific instrument is most appropriate. The main difference between mapping and constructing a CS-PBM is that the development of a PBM assigns population weights (via TTO) to the item levels of a questionnaire, while a mapping function assigns weights to the items that are dependent on the generic measure it aims to estimate. As such, issues with insensitivity of the generic instrument are not resolved when mapping a condition specific measure onto a generic PBM. In our view, a well conducted and validated mapping function may be preferred to the development of a CS-PBM, because it yields utility values that compare better to the more frequently used generic instruments used in other economic evaluations, but only under the following circumstances: 1) there is no empirical evidence for insensitivity of the generic instrument, and 2) only use of mean utility values is intended rather than sub group analysis (122) and 3) the health status or disease subtype of the sample on which the function was estimated is comparable to the sample on which the function is applied (114).

Findings here underline that the TTO health state values as modeled from a fractional factorial design can differ from direct TTO valuations of those states. Often but not always an OMEP is applied to allow the estimation of TTO values for all theoretically possible health states from only a fraction of health states. This study adopted that technique but also valued directly a

selection of states that were observed frequently in patients. Using these states in the estimation of the preference algorithm resulted in lower scores for at least some of these states (Fig. 5.1). These results suggest that discrepancies exist between modeled TTO values and directly observed TTO values for the most occurring health states which may affect the validity of the measure. Little guidance is available for researchers who wish to design a valuation study for a CS-PBM using state of the art techniques, so it is not surprising that practices vary and this deserves more attention to ensure that high quality CS-PBMs are produced. Ideally the process of constructing the CS-PBM is supported by the original developers of the questionnaires. This is relevant for example to avoid wild grow of value sets (e.g. for the QLQ-C30 now multiple value sets exists derived via mappings (114, 123-125)), to further guarantee quality, and to offer support to users of the CS-PBM .

Constructing and using a CS-PBM for the purpose of resource allocation could be considered when the following conditions are met: empirical evidence disproves sensitivity of existing generic instruments, empirical evidence proves the superiority of the condition specific measure from which the new preference-based measure will be derived and the derived CS-PBM is shown to be superior to the existing CS-PBM, not just in terms of statistical sensitivity, but also in terms of absolute differences. The development of CS-PBMs is welcome from an academic point of view as it pushes methodological frontiers and introduces new data for comparing measures in a field where no gold standard PBM exists. Use in resource allocation of these instruments, however, is only warranted when the above mentioned conditions are met. The introduction of preference-based measures which are specific to a certain disease has the presupposed merit of sensitivity to disease specific effects of interventions, but this article shows that such an advantage is not necessarily achieved. Furthermore, the possible increase in sensitivity is traded off to the loss of comparability of absolute differences in utility values, which are most important for economic evaluations. It is argued here that without convincing empirical evidence on the insensitivity of a generic instrument, using a CS-PBM introduces confusion about the appropriate outcome measures in cost-utility analysis and health care decision making.

APPENDIX A: RESULTS FROM THE PROCESS OF ITEM SELECTION

Item selection: HAQ

The 20 items of the HAQ did not have disordered thresholds. The HAQ had some misfit to the Rasch model, caused by item 10 ('wash body') and 16 ('open jar') with significant fit residuals (>3). Removal of these items improved the fit of the scale to the Rasch model (Item fit residual = -.37, SD = 1.36, Item trait interaction χ^2=91, DF=90, p=0.44, Person separation index = 0.94). All remaining 18 items performed nicely in the Rasch model, and none of the remaining 18 items showed differential item functioning for gender or age group (consisting of two groups under and over the median age of 53).

Employing further psychometric criteria did not aid the selection of items as all HAQ items had >37% of the responses on the highest level. Linear correlation between individual HAQ items and EQ-5D index was <0.29, with an average of 0.21 (SD = 0.03) without any marked deviation.

On the basis of the analysis of three previous Rasch analyses (112, 126, 127)it was decided to go with the five items that are included in both the HAQ and its successor the HAQ-II. These items were 'Get on and off the toilet', 'Open car doors', 'stand up from a straight chair', 'walk outdoors on flat ground' and 'reach and get down a five pound object (such as a bag of sugar) from just above your head'. The last item (5 pound object), is different in the Dutch translation of the HAQ, where 5 pounds is changed in to 1 kg, meaning that it represents less disability than the original item (112), which may yield a smaller utility decrement in a TTO study for that item than when using the original English version.

Item selection: MSIS-29

The Rasch analysis was applied separately for the physical and psychological functioning scales of the MSIS-29.

Nearly all items of the physical scale showed difficulty for respondents to differentiate between the categories 'a little', 'moderate' & 'quite a bit' and some had reversed thresholds (no ordinal order of categories). All items were rescored to a 4-point scale merging 'moderate' with one of the adjacent categories. Rescoring improved the fit considerably but the scale continued to misfit the Rasch model. Several items were deleted for different reasons (table 2). Applying psychometric criteria suggested to the removal of item 17 'trouble using transport' for which 45% of respondents reported no problems. Item 15 was retained despite 29% of respondents reporting no problems on the item, as removal worsened scale fit. The resulting scale was found to be unidimensional, showed no DIF and fit the Rasch model. On the basis of the spread of

difficulty represented by the item, and advice from the clinical expert, 5 out of 8 items were selected for the vignette. The selected items are 4 ('Problems with your balance'), 6 ('Being clumsy'), 13 ('Limitations in your social and leisure activities at home'), 15 ('Difficulties using your hands in everyday tasks'), 16 ('Having to cut down the amount of time you spent on work or other daily activities').

The psychological scale of the MSIS-29 showed disordered thresholds for item 26. Items showed difficulty for respondents to distinguish between level 'moderate' and the two adjacent categories. All items were rescored to a 4-point scale merging either with the adjacent higher or lower level of level 'moderate'. Rescoring slightly improved model fit. Table 2 shows the process of deleting items and the effect on the total fit of the psychological scale. The resulting items fit the Rasch model individually and as a scale, showed no DIF and were unidimensional. On the basis of the psychometric criteria, item 22 ('problems sleeping') and 29 ('feeling depressed') were not considered to be a candidate for the selection on the vignette as 30% of respondents reported 'no problems'. Item 28 was not considered for the vignette on the basis of spread of difficulty of the item. The linear correlation (R^2) between the items (i) and sum of the other (not selected) items in the domain was used to inform a final decision. Linear correlation was highest (>.6) for item 25 ('feeling anxious or tense'), 26 ('feeling irritable, impatient or short tempered') and 27 ('problems concentrating'). Despite the high correlation for item 25, it was decided to go with items 26 and 27 as item 25 had a high fit residual (Fit. Res. -2.1, $\chi^2_{(df=5)}$= 9.7, p=.08). Upon consultation of the clinical MS expert it was decided to add item 23 'feeling mentally fatigued' to the final list, as this item was deemed a crucial element in MS.

Item selection: QLQ-C30

The QLQ-C30 questionnaire consists of five functional scales, nine symptom scales and one global health status. The aim of the item selection was to include only one item of each QLQ-C30 scale, and to have the health state represent all the dimensions identified by factor analysis. Rasch analysis was not performed on all of the scales, as some scales consist of only one item. Therefore, psychometric criteria combined with expert opinion were used as the main criteria for selection of the items.

A principal component analysis was performed on all items but global health status. The global health status scale is an assessment of the quality of life in general rather than a specific aspect of quality of life and was therefore not considered for the health state description. Five different factors were identified, which we summarize with the following factor identifiers: physical functioning, vitality, mental functioning, discomfort and pain (table 1).

Table 5.A1 Factor structure derived from PCA

Physical functioning		Vitality		Mental functioning		Discomfort		Pain	
Taking a long walk (PF)	0.72	Social activities (SF)	0.66	Depressed (EF)	0.75	Nauseated (NV)	0.8	Pain (PA)	0.84
Strenuous activities (PF)	0.68	Family life (SF)	0.66	Tense (EF)	0.74	Vomited (NV)	0.77	Pain interfere with daily activities (PA)	0.81
Tired (FA)	0.66	Help with eating, dressing washing or using the toilet (PF)	0.66	Irritable (EF)	0.74	Appetite loss (AP)	0.62		
Short of breath (DY)	0.6	Stay in bed or chair (PF)	0.64	Worry (EF)	0.73	Constipation (CO)	0.46		
Limited in work or other daily activities (RF)	0.6	Short walk (PF)	0.59	Concentrating on things (CF)	0.51	Diarrhea (DI)	0.36		
Need to rest (FA)	0.59	Limited in hobbies (RF)	0.57	Sleeping (SL)	0.49				
Felt weak (FA)	0.57	Financial difficulties (FI)	0.47	Remembering things (CF)	0.43				

QLQ-C30 scale abbreviations: PF=physical functioning; RF=role functioning; EF=emotional functioning; CF=cognitive functioning; SF=social functioning; FA=fatigue; NV=Nausea and vomiting; PA=pain; DY=dyspnoea; SL=insomnia; AP=appetite loss; CO=constipation; DI=diarrhea; FI=financial difficulties. Shaded items were selected for the QLQ-C30 health states.

The fourteen QLQ-C30 scales loaded on five factors. Within these factors, we aimed to select items that belonged to different QLQ-C30 scales to obtain a maximum representation of relevant items that impact on quality of life. When items loaded on the same factor but belonged to different QLQ-C30 scales, we accepted a maximum correlation of 0.6 between the two items.

The 14 QLQ-C30 scales were rank ordered based on their linear correlation with the global health scale. An arbitrary cut off point was that scales had to explain 15% of the variance in the global health scale to be selected for the health state. Following this strategy the following QLQ-C30 scales were not considered for the vignette 'Nausea and vomiting', 'Dyspnoea', 'Constipation', 'Diarrhea' and 'Financial difficulties' Some of these scales consist of multiple items so items were also rank ordered on the percentage of variance explained on the global health scale and on their distribution of responses on the item levels. We based our final selection of items within a QLQ-C30 scale on this rank order. In one instance we deviated from our strategy. We chose to include 'have you felt nauseated' rather than 'appetite loss'. Although the item 'have you felt nauseated' did not meet our prior requirements, the item 'appetite loss' was replaced with 'have you felt nauseated' on the advice of the cancer expert.

6

Mapping QLQ-C30, HAQ and MSIS-29 on EQ-5D

With Annemieke Leunis, Jolande Luime, Mike Boggild, Carin Uyl-de Groot and Elly Stolk.
Medical Decision Making, 2012

ABSTRACT

Responses on condition-specific instruments can be mapped on EQ-5D to estimate utility values for economic evaluation. Mapping functions differ in predictive quality and not all condition-specific measures are suitable for estimating EQ-5D utilities. We mapped QLQ-C30, HAQ and MSIS-29 on EQ-5D and compared the quality of the mapping functions with statistical and clinical indicators

We used four datasets that included both EQ-5D and a condition-specific measure to develop ordinary least squares regression equations. For the QLQ-C30, we used a multiple myeloma dataset and a non-Hodgkin's lymphoma one. An early arthritis cohort was used for the HAQ, and a cohort of patients with relapsing remitting or secondary progressive multiple sclerosis for the MSIS-29. We assessed the predictive quality of the mapping functions with the root mean square error (RMSE) and mean absolute error (MAE) and the ability to discriminate among relevant clinical subgroups. Pearson correlations between the condition-specific measures and items of the EQ-5D were used to determine if there is a relationship between the quality of the mapping functions and the amount of correlated content between the used measures.

QLQ-C30 had the highest correlation with EQ-5D items. Average %RMSE was best for QLQ-C30 with 10.9%, 12.2% for HAQ and 13.6% for MSIS-29. The mappings predicted mean EQ-5D utilities without significant differences with observed utilities and discriminated between relevant clinical groups, except for the HAQ model.

The preferred mapping functions in this study seem suitable for estimating EQ-5D utilities for economic evaluation. However, this research shows that lower correlations between instruments leads to less predictive quality. Using additional validation tests besides reporting statistical measures of error improves the assessment of predictive quality.

6.1 INTRODUCTION

Utility values (128), which are required to conduct cost-utility analyses, are usually measured by preference-based instruments like the EQ-5D, SF-6D or HUI III. Many clinical trials, however, use condition specific instruments which do not incorporate preferences in the scoring algorithms, rather than preference-based instruments (30). Utility values can be estimated from the answers on a condition specific instrument when preference-based instruments are absent (45). This technique is called 'mapping' and primarily serves the purpose of 'rescuing' trial data for economic evaluation when a preference based instrument is absent. This paper presents a study that uses data from three condition specific questionnaires (Cancer: EORTC QLQ-C30 version 2; Arthritis: HAQ and Multiple Sclerosis: MSIS-29) to predict outcomes on a preference based instrument (the EQ-5D). We compare the quality of the mapping functions with statistical and clinical indicators and explore the influence of overlap in dimensions. Condition specific measures do not necessarily measure the same dimensions of health as a preference-based measure. The amount of overlap in dimensions between instruments is considered to be of influence on the predictive ability of the mapping function (129).

'Mapping' comes down to giving weight to different independent variables (items of a scale, the 'starting' measure) to predict the dependent variable (the 'target measure', e.g. utility index) through regression techniques. The independent variables may be sum scores, item scores, demographic variables and/or other (clinical) predictors of health. The purpose of such a mapping effort is to enable the estimation of a utility index in other datasets that do not have the 'target measure' but do have the 'starting measure'. The utility index, derived from the mapping effort, can be used to calculate the quality adjustment necessary for the computation of a Quality Adjusted Life Year (QALY). Previous studies have found that mapping is a feasible approach, but mapping models differ in predictive ability (45). Disadvantages of mapping are that the estimated utility indices have far greater errors for severe health states, and EQ-5D models tend to overpredict low utilities and underpredict high utilities (45, 130). The performance of a mapping function may be tested by applying the algorithm to other (subset) data, which, like the sample on which the statistical association between dependent and independent variables was estimated, has data for both instruments.

Over the last few years several mapping algorithms have become available. In some circumstances mapping may be the only way to get utility data, but the current growth in use of this strategy requires a more critical attitude towards the problems and promises of its application as the quality of the mapping algorithms is highly variable (45). Assessing the quality of a mapping function, however, is not straightforward as different indicators may 'hide' certain flaws in a mapping function. First, an accurate prediction of mean utility values may cover that prediction

errors may be larger for particular subgroups of patients. Second, (statistical) error indicators may not be easily interpretable without a comparator. Third, a measure of error does not directly reflect the external validity of a mapping function.

Another issue that deserves attention is the amount of overlap between dimensions of health covered by the descriptive systems of the starting and target measure. If the instruments focus on different dimensions of health (e.g. pain and mobility) they have little overlap and hence a low correlation between the items of the two measures, which may negatively influence the quality of the derived mapping function. The EQ-5D index values are the result of an algorithm that transforms the answers on five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. To predict the index value, the variation of responses on those categories has to be predicted. A scale that only measures pain may have difficulty explaining variation in answers on a self-care domain. Consequently, the amount of overlap between instruments may influence the predictive ability of a mapping function.

This paper aims to present mapping functions for three condition-specific questionnaires onto EQ-5D suitable for use in economic evaluations, and to assess their quality through both statistical error measures and clinical indicators. We also test the hypothesis that the overlap of health domains is an important predictor for the predictive quality of a mapping function. If the amount of overlap between two instruments can be assessed at face value, it may inform a quick judgment about the expected quality of the derived mapping function.

6.2 METHODS

Instruments

Both the condition specific measures and the preference-based generic measure are patient reported outcome measures to assess health status. The measures have different properties as outlined below.

EQ-5D

The EQ-5D is a preference-based generic measure. It measures health related quality of life on five dimensions (mobility; self-care; usual activities; pain/discomfort and anxiety/depression) with 3 severity levels each. The measure was developed to provide a "simple abstracting device for use alongside other more detailed measures of health-related quality of life to serve as a basis for comparing health outcomes" (131). The main outcome of interest is the derived 'utility (or preference) index'; a single metric for quality of life derived by transforming the dimension scores with country-specific tariff. Utility values used in this study were computed using the Dutch EQ-5D tariff(75) and the UK EQ-5D tariff(26).

QLQ-C30

The EORTC-QLQ-C30 (version 2) is a cancer specific questionnaire and consists of 30 items, divided into three categories: functional scales (physical, role, emotional, cognitive and social functioning, total of 15 items), symptom scales (fatigue, nausea/vomiting, pain, dyspnoea, sleep, appetite, constipation, diarrhea and financial difficulties, total of 13 items) and a global health status scale (two items). Scale sum scores are transformed so that a high score on the functional scales represents a high level of functioning, a high score on the symptom scales represents a high level of symptomatology and a high score on the global health status represents a high quality of life(111). At face value it seems all health domains of EQ-5D are present in the QLQ-C30. The QLQ-C30 has been successfully mapped on the EQ-5D (UK tariff) before(124), but not for Dutch utilities nor for a lymphoma population.

HAQ

The Health Assessment Questionnaire (HAQ) is a widely used questionnaire in the field of rheumatology. The HAQ assesses the functional ability of patients using 20 items across eight domains (dressing, rising, eating, walking, hygiene, reach, grip and usual activities) (132). Items are scored on a four level disability scale from zero to three, where three represents the highest degree of disability. Scores are adjusted for the use of aids or devices and averaged into dimension sum scores and an overall disability index value, which represents the extent of functional ability of the patient. Values between one and two represent moderate to severe disability (121). A face value judgment of the items and sum scores of HAQ suggests the EQ-5D dimensions pain/discomfort and anxiety/depression are not represented in the HAQ. The HAQ has been mapped on the EQ-5D before and was able to predict mean utility values (133), but such a function has not been estimated for Dutch utilities.

MSIS-29

The MSIS-29 is a Multiple Sclerosis (MS) specific questionnaire with 29 items developed through reducing an item pool of 129 items concerning the health impact of multiple sclerosis(110). The MSIS-29 is a self-reported measure which measures the physical and psychological impact of MS on individuals. Items measure disease impact due to limitations in the past two weeks, scored on 5 levels from 'not at all' to 'extremely'. The first 20 items (physical scale) and the last 9 items (psychological scale) form two summary scores transformed to a 0-100 scale. The MSIS-29 has not been previously mapped to our knowledge. The EQ-5D dimensions mobility, self-care, and usual activities are not explicitly represented as dimensions of the scale, but the dimensions are tapped into by items like 'difficulty moving about indoors', 'having to depend on others to do things for you' and 'limitations in your social and leisure activities at home'. MSIS-29 items and sum scores suggests the EQ-5D dimension 'pain/discomfort' is not present in the MSIS-29.

Population

The EQ-5D and condition specific measure data were retrieved from three different datasets described below. An overview of characteristics of the populations is presented in table 1.

Table 6.1 Patient characteristics

Sample		Development set	Test set 1	Test set 2
QLQ-C30	N	723 (pooled)	789 (pooled)	
	Age (range)	54 (37 - 64)	72 (65-84)*	
EQ-5D†	Mobility %1/2/3	56.7 / 41.4 / 1.9	48 / 47.3 / 4.7	
	Self-care %1/2/3	85.8 / 12.8 / 1.4	81.4 / 13.9 / 4.7	
	Usual activities %1/2/3	30.1 / 51.1 / 18.8	38.1 / 43.3 / 18.6	
	Pain/Discomfort % 1/2/3	39.6 / 59 / 1.4	52.2 / 42.9 / 4.9	
	Depression/Anxiety % 1/2/3	69.4 / 29.6 / 1.0	70 / 29 / 1.0	
	EQ-5D-index	.742 (.21)	.735 (.26)	
	Cancer type	Multiple Myeloma	Non-Hodgkin lymphoma	
QLQ-C30 (pooled)	Physical functioning	64 (24.6)	57.3 (26.8)*	
	Role functioning	59.5 (28.9)	57.4 (31.5)	
	Emotional functioning	82.8 (18.9)	81.3 (20.6)	
	Cognitive functioning	82 (20.8)	81.9 (23.7)	
	Social functioning	76.2 (25.8)	75.7 (28.6)	
	Global health	68.7 (18.0)	62 (21.7)*	
	Fatigue	35.7 (25.0)	44.7 (29.4)*	
	Nausea / Vomiting	6.1 (14.3)	8 (16.9)*	
	Pain	25.2 (24.7)	18.7 (26.2)*	
	Dyspnoea	16.1 (24.9)	24.8 (28.9)*	
	Sleep	21.1 (27.3)	28.7 (31.8)*	
	Appetite	16 (27.2)	21.9 (32.6)*	
	Constipation	4 (15.4)	11.8 (22.8)*	
	Diarrhea	8.3 (18.7)	7 (18.5)*	
	Financial difficulties	12.5 (23.0)	6.3 (16.9)*	
HAQ	N	186	132	
	Age (range)	51 (16 - 82)	55 (25 – 78)	
EQ-5D†	Mobility %1/2/3	38.2 / 61.8 / 0	44.4 / 55.3 / 0	
	Self-care %1/2/3	68.5 / 30.4 / 1.1	58.0 / 41.2 / .8	
	Usual activities %1/2/3	33.9 / 60.1 / 6.0	27.5 / 67.9 / 4.6	
	Pain/Discomfort % 1/2/3	8.1 / 78.4 / 13.5	3.8 / 77.3/ 18.9	
	Depression/Anxiety % 1/2/3	72.4 / 25.4 / 2.2	80.3 / 17.4 / 2.3	
	EQ-5D-index	.67 (.24)	.64 (.26)	
	DAS 28	4.34 (1.30)	4.30 (1.27)	
	HAQ-DI	0.75 (.65)	0.81 (.65)	

HAQ	Dressing & Grooming	0.64 (.72)	0.76 (.75)	
	Arising	0.71 (.77)	0.72 (.75)	
	Eating	0.84 (.84)	0.95 (.83)	
	Walking	0.64 (.85)	0.58 (.81)	
	Hygiene	0.70 (.85)	0.79 (.86)	
	Reach	0.68 (.80)	0.74 (.81)	
	Grip	0.87 (.86)	0.93 (.86)	
	Activities	0.93 (.86)	1.04 (.88)	
MSIS-29	N	661	339	295
	Age (range)	40 (18-88)	40 (18 -87)	41 (18-88)
EQ-5D†	Mobility %1/2/3	21.2 / 76.6 / 2.2	26.7 / 70.9 / 2.4	25.0 / 74.3 / .7
	Self-care %1/2/3	62.4 / 35.8 / 1.8	68.2 / 29.4 / 2.4	63.0 / 35.6 / 1.4
	Usual activities %1/2/3	21.1 / 70.9 / 8.0	22.9 / 69.4 / 8.3	23.5 / 71.3 / 5.2
	Pain/Discomfort % 1/2/3	25.3 / 67 / 7.7	27.5 / 61.5 / 11	24.7 / 68.2 / 7.1
	Depression/Anxiety % 1/2/3	40.7 / 52.1 / 7.2	41.1 / 50.9 / 8.0	39.2 / 56.3 / 4.5
	EQ-5DUK/NL	.58 (.29)/ .63 (.62)	.57 (.31) / .61 (.20)	.60 (.25) / .65 (.23)
	Type of MS‡	RR = 81% SP = 19%	RR = 81% SP = 19%	RR = 82% SP = 18%
MSIS-29	Physical scale	47.1 (25.7)	45.4 (25.8)	44.9 (25.7)
	Psychological scale	45.4 (25.6)	44.7 (24.9)	44.4 (25.6)

* Significant difference (2-tailed t-test $p < .05$) with development set. † % at level 1 / 2 / 3 (not at all / some problems / extreme). ‡ RR= Relapsing Remitting, SP = Secondary Progressive

QLQ-C30: the HOVON study

Data for QLQ-C30 model were taken from two separate studies carried out by the Dutch association for hematology/oncology in adults (HOVON). The HOVON 24 (134) and HOVON 25(135) studies are randomized clinical trials that measure the effectiveness of different treatments in respectively Multiple Myeloma patients with previously untreated multiple myeloma (HOVON 24) and previously untreated Non-Hodgkin lymphoma patients (Ann Arbor stages II to IV, or intermediate or high-grade malignancy). The sample size of the clinical trial is larger than the sample size of patients that had concluded both an EORTC QLQ-C30 and an EQ5D instrument.

The mapping algorithm was developed on the multiple myeloma sample, and tested on the Non-Hodgkin lymphoma sample. The database for the multiple myeloma sample contained 137 patients at baseline with 6 follow-up measurements (three early follow-ups were missing), the latest two years after baseline. To increase the number of data-points per EQ-5D utility value – samples are often short on severe health states and thus lack predictive ability in the lower range – follow up measurements were pooled for the development of prediction models. The database for the Non-Hodgkin's sample contained 108 patients at baseline and had seven time

points at which the EORTC QLQ C-30 was administered. Three time-points were after the second, fourth and sixth treatment cycle, and four follow-up measurements were at baseline, three, six, ten and eighteen months after baseline. Predictive ability of the mapping is assessed per time point.

HAQ: the REACH study

The HAQ data were taken from the Rotterdam Early Arthritis CoHort (REACH) with patients recruited from the Erasmus Medical Centre in the Netherlands, with arthritis. As can be inferred from the name, one of the aims of the study is early detection of rheumatoid arthritis. Data are collected through three outpatient visits, during which respondents filled out a booklet of self-report questionnaires including HAQ, Hospital Anxiety and Depression Scale (HADS), Short Form-36 (SF-36) and EQ-5D. The mapping algorithm is developed on a randomly drawn sub-sample of the dataset (N=186) and tested on a remaining sample (N=132) for which most of the data of the questionnaires were available, both at baseline. A randomly drawn subsample is not expected to deviate much from the remaining sample.

MSIS-29: the MS risk sharing scheme monitoring study

The MSIS-29 data were taken from the Multiple Sclerosis Risk Sharing Scheme Monitoring Study in the UK (136), which aims to analyze the long term cost effectiveness of disease modifying treatments in patients with MS. Data for the study were collected from MS patients in 70 specialist MS centers in the UK and includes both relapsing remitting and secondary progressive MS patients. Cross-sectional cost and utility data for a sub-set of these MS patients was collected to enhance economic analysis. The MSIS-29 was administered once (N= 1295); hence there are no different time-points for this measure. The mapping algorithm was developed on a randomly drawn sub sample (N=661) and tested on two samples randomly drawn from the remaining respondents (N=339 and N=295). As with the HAQ, the randomly drawn samples are not expected to deviate much from the rest of the data.

Analysis

Pearson's r and Spearman's rho determined the amount of overlap between the instruments. Ordinary least squares regression models will be fitted to the data to generate mapping algorithms. All models were developed for the Dutch value set of the EQ-5D. As the MS dataset is based on the English versions of the MSIS-29 and EQ-5D, the algorithm was also estimated for the UK value set.

The models in this paper serve the purpose of calculating mean utility scores applicable in economic evaluation. The mapping models generate individual utility values, but these have uncertain estimates (137). Even if the estimates were less uncertain, the EQ-5D is a tool for

economic evaluation and priority setting in health care (29), not a key health indicator of individual health. Thus only aggregated utilities derived from mapping studies are to be used. Analyses were run in STATA 10.0 and SPSS 17.0.

Model development

This study follows the suggestions of a recent review of mapping studies which suggests estimating different types of models with increasing complexity and decreasing assumptions about the properties of the data (45). The strategy implies that a range of models is estimated, with increasing levels of complexity. All models that we developed aimed to compute EQ-5D utility scores from the condition specific measures with ordinary least squares regression. An alternative approach would be to predict the dimension scores of EQ-5D through multinomial regression. This implies that the models predict occurrence of level 1, 2 and 3 responses on the EQ-5D from which one indirectly computes the utility scores. We chose ordinary least squares regression to predict the EQ-5D utility index as predicting dimension levels has been shown to have equal or worse predictive performance (45) for both QLQ-C30(124) and HAQ (133).

All models were developed in similar order, relaxing the assumptions about the properties of the condition specific measures in each step, while eyeballing model performance (described below) as guidance for model selection. First the EQ-5D utility index was regressed on sum scores, then on item scores treated as continuous variables. Treating items as continuous variables assumes interval properties of the questionnaire, because only one coefficient is calculated for all changes in responses of one item (i.e. On the question 'have you been bothered by problems sleeping?' moving from answer category 'not at all' to 'a bit' receives the same decrement as moving from adjacent categories 'quite a bit' to 'extremely'). Lastly, by treating item scores as categorical (dummy) variables, this assumption was relaxed. Dummies are computed to let each subsequent level represent worse health compared to the reference category 'no problems'. Models were developed using backward selection procedures with probability of F to remove a variable at 0.10.

Models were required to be logically consistent meaning that a worse score on an item should lead to a larger utility decrement. For reasons of parsimony models were reduced to the smallest models that have similar predictive performance as larger models. In the final models achieving parsimony was attempted by merging item categories with the logically adjacent answer category when they a) did not meet probability <0.10 of F to remove or b) were logically inconsistent.

Clinical variables or additional questionnaires may correlate with a given dimension (muscle strength may capture mobility). Such variables may improve the mapping but might not be available as data in all clinical trials, potentially limiting the use of a function. As the HAQ is a

measure of disability and not a quality of life questionnaire we expected that HAQ based models might not yield favorable prediction results. We therefore selected 5 variables in our dataset that theoretically would correlate with EQ-5D dimensions. The variables are: sum scores of the Hospital Anxiety and Depression Scale, sum scores of the SF-36 and the Disease Activity Score (DAS28) and a count of tender and swollen joints. Analysis started with all predictor variables as regression coefficients and proceeded through backward step wise regression with probability of F to remove at 0.10. Missing values were not imputed in any of the datasets.

Model performance

There are no strict criteria for a utility function to be acceptable (45). QALYs are computed using mean scores, thus interesting indicators are differences between the predicted mean utility score and the observed mean utility score. However, correct prediction of a mean may still 'hide' differences between individual observed and predicted values across the entire scale. Model performance is therefore reported as root mean squared error (RMSE) and mean absolute error (MAE) of the predicted EQ-5D utility values (4, 45). RMSE is the root of the average of the squared differences between observed and predicted values, while MAE is the average of the roots of the squared difference between observed and predicted values. Lower values of RMSE and MAE indicate better model performance. As RMSE averages the squared differences it is sensitive to extreme deviations from the mean (e.g. outliers) and thus always equal to or larger than MAE. There is no definition of what level of RMSE or MAE is a threshold for model acceptance. Besides that, RMSE and MAE are not comparable for models with different preference-based instruments as dependent variables or models with a different range of observed values since larger scale size usually leads to a larger error figure. For instance: UK EQ-5D index values have a measurement range of 1.59 compared to 1.33 for the Dutch values. Consequently RMSE is also reported as a percentage of the scale size, the normalized RMSE (138). It is expected that a higher prediction error is positively associated with less overlap between EQ-5D and the disease specific instrument.

Discriminant validity

Statistical measures such as RMSE and MAE may be difficult to interpret and can be small due to method of testing, caused, for instance by using a randomly drawn subsample which does not deviate much from the development sample. Therefore, the validity of predicted index values is also inspected by checking the ability of the predicted values to discriminate between relevant clinical groups. Mean scores of observed and predicted EQ-5D index values are calculated per category of a relevant clinical indicator. For the cancer data this clinical indicator is the doctor reported World Health Organization performance status (or ECOG score) which distinguishes 6 categories from 0 (asymptomatic) to 6 (death). For the arthritis data the clinical indicator is the Disease Activity Score 28 (DAS28) and is based on a count of tender joints and the erythrocyte

sedimentation rate (ESR). It can be used to distinguish between high, moderate and low disease activity and remission. Lastly, for multiple sclerosis the data are compared to the categories of the Expended Disability Status Scale (EDSS) which range from 0 (no neurological deficit) and 10 (death). Pearson's r is used to measure the (linear) correlation between the clinical indicators and the observed and predicted scores.

6.3 RESULTS

Pearson correlations indicate that de condition specific measures differ in the amount of overlap with EQ-5D dimensions (Table 2). Table 2 suggests that Pearson correlations with EQ-5D dimensions were higher for QLQ-C30 and MSIS-29 than for HAQ and were nearly identical to Spearman's rho. For instance, none of the HAQ dimensions had a correlation coefficient >0.23 with the EQ-5D dimensions 'anxiety / depression', whilst this dimensions correlates with -0.70 and 0.68 for respectively QLQ-C30 and MSIS-29. QLQ-C30 has the highest correlations with the EQ-5D dimensions. Based on these results we would expect the mapping functions based on QLQ-C30 to outperform those based on the HAQ and MSIS-29.

Table 6.2 Pearson's correlation matrix between sum scores and EQ-5D dimensions

	EQ-5D domain				
QLQ-C30†	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Physical functioning	**-0.67****	-0.48**	**-0.64****	-0.45**	-0.27**
Role functioning	-0.54**	-0.38**	**-0.78****	-0.47**	-0.29**
Emotional functioning	-0.22**	-0.20*	-0.30**	-0.24**	**-0.70****
Cognitive functioning	-0.20*	-0.17*	-0.28**	-0.27**	-0.37**
Social functioning	-0.49**	-0.34**	**-0.55****	-0.30**	-0.40**
Global health status	-0.36**	-0.18*	-0.47**	-0.44**	-0.39**
Fatigue	0.35**	0.20*	0.51**	0.39**	0.34**
Nausea and vomiting	0.03	-0.08	0.11	0.21*	0.09
Pain	0.40**	0.17*	0.41**	**0.76****	0.19*
Dyspnoea	0.11	0.13	0.26**	0.08	-0.02
Sleep	0.08	0.10	0.19*	0.19*	0.30**
Appetite loss	0.25**	0.15	0.28**	0.25**	0.27**
Constipation	0.05	0.04	0.05	0.26**	0.04
Diarrhoea	0.21*	0.07	0.14	0.08	0.10
Financial difficulties	0.14	0.22**	0.22**	-0.01	0.12
HAQ	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Dressing	0.29**	**0.63****	0.42**	0.41**	0.17**
Rising	0.44**	0.51**	0.45**	0.37**	0.20**
Eating	0.27**	0.52**	0.44**	0.36**	0.11*
Walking	0.52**	0.48**	0.39**	0.32**	0.21**
Hygiene	0.35**	**0.62****	0.41**	0.39**	0.19**
Reach	0.35**	0.54**	0.39**	0.44**	0.23**
Grip	0.27**	0.45**	0.36**	0.40**	0.15**
Usual activities	0.42**	0.54**	0.51**	0.45**	0.18**
MSIS29	mobility	self care	usual activities	pain/ discomfort	anxiety/ depression
Physical scale	**0.67****	**0.59****	**0.67****	0.50**	0.40**
Psychological	0.43**	0.37**	0.46**	0.45**	**0.68****

** Correlation is significant at the 0.01 level (2-tailed).* Correlation is significant at the 0.05 level (2-tailed). † Time point = baseline. Bold > 0.55

Mappings

The best functioning models and their performance are summarized in table 3 and further discussed below. QLQ-C30 model 4, HAQ model 3 and MSIS-29 model 4 meet the requirements of logical consistency, significance of predictors, parsimony and were able to predict mean utility values in the test samples (table 4). Box plots of prediction errors per EQ-5D utility category are presented in figure 6.1.

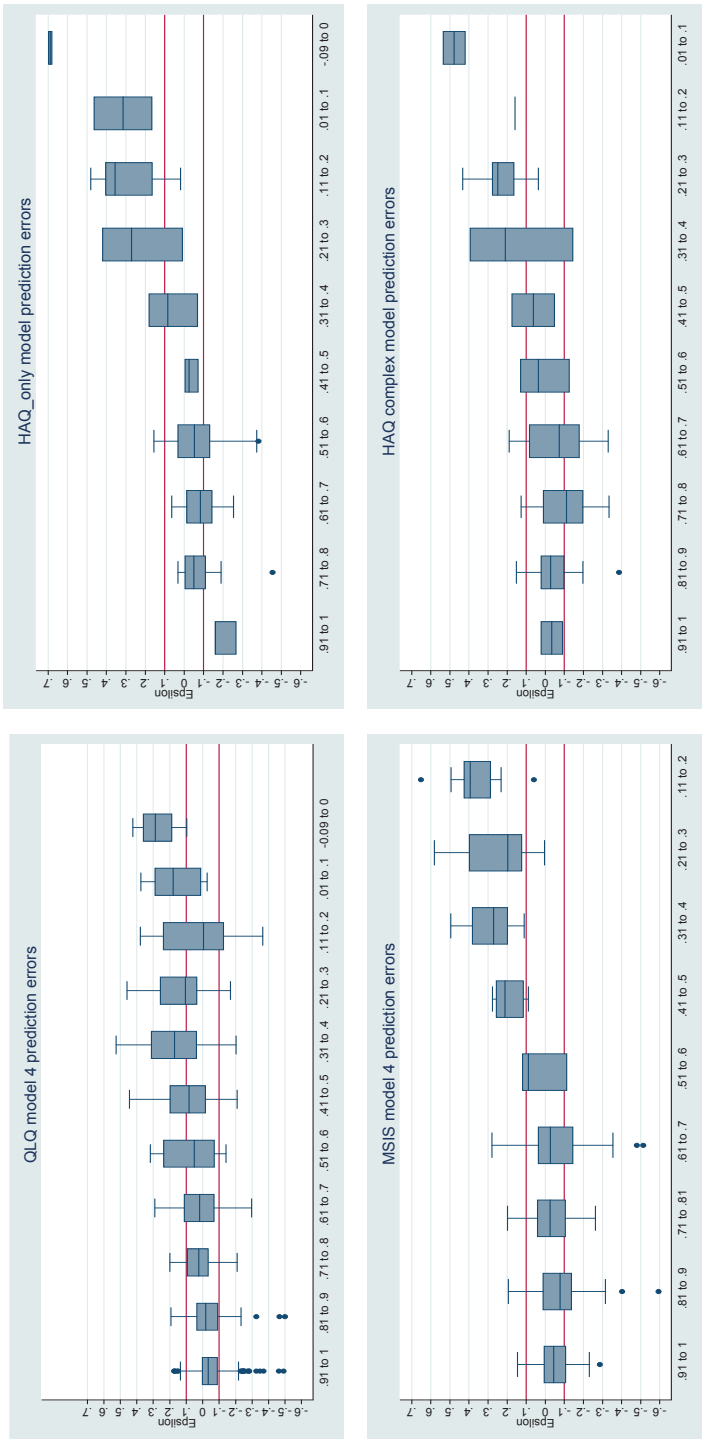

Figure 6.1 Box plots of Epsilon (Observed – Predicted) per EQ-5D utility category

QLQ-C30

Predictions were better when the assumptions concerning data were relaxed. As expected from the relatively low correlation with EQ-5D dimensions, the sum score of cognitive functioning scale was excluded after backward selection in model 1, as cognitive functioning is not represented in the EQ-5D. Using items as continuous predictor variables (model 2) reduced prediction errors, but performed worse than model 3 which used items as dummy variables. In model 3, after backward selection, it was decided to remove items from the model based on illogical signs (item 10, 22 and 25), at the cost of 2 percent explained variance. Dummy model 3 consists of items 1 to 5 (physical functioning); 6 and 7 (role functioning); 9 (pain); 16 (constipation); 23 & 24 (emotional functioning) and item 27 (social functioning). To achieve a more parsimonious model, non-significant, and remaining significant but illogically ordered (items 7 and 9) dummy categories were merged and item 16 (constipation) was dropped altogether, without effecting the predictive performance of the model. The model developed in the Multiple Myeloma patient sample was able to predict utilities in the Non-Hodgkin's sample. The largest RMSE value is at baseline and 10 months follow-up which is due to an outlier as can be seen from the relatively low MAE which is less sensitive to outliers. When the model could not predict a utility value because the NH sample had missing values in QLQ C-30 responses, or when the utility value from the EQ-5D was missing, cases were excluded leading to minimal differences in RMSE (smaller than 0.01) compared to including missing data-points as well.

HAQ

As the sample was focused on early arthritis the dummy items often lacked respondents which scored the lowest answer category (level 4). In a dummy model, the variables that represent the fourth category mostly score 0, yielding it impossible to estimate a decrement for a level 4 answer. Consequently a model with summed dimension scores had better RMSE scores in the test sample. Removing the insignificant sum variables in the prediction model did not improve predictions in the test sample. After backward stepwise regression, the model with the a priori selected predictor variables (extended model) performed better in terms of RMSE, MAE and R^2 and was capable of predicting a wider range of EQ-5D index values as can be seen in table 4. Stepwise selection of variables for the extended model resulted in the removal of all HAQ sum scores except 'usual activities'. Tender joint count and swollen joint count did not contribute to the model but are represented indirectly through the Disease Activity Score. The other variables are the transformed sum scores of the SF-36 ('physical functioning', 'role-physical', 'bodily pain', and 'role-emotional'), the depression sum score of the Hospital Anxiety & Depression Scale. Added variables had significant Pearson's correlations between 0.47 and 0.58 with at least one of EQ-5D dimensions. Adding a squared term for the SF36 bodily pain score did not reduce prediction errors. The extended model failed to predict 27 of the 132 EQ-5D index values due to missing data for one or more of the prediction variables.

MSIS-29

For the Dutch value set, MSIS-29 models 2 to 4 with items as predictors performed better than sum score model 1. Treating items as dummy variables did not reduce prediction error in terms of RMSE and MAE. However, the continuous model seemed illogical with an intercept of 1.17 while a value of 1 represents full health in the QALY-model. Model 3 contained many insignificant and illogical items, but removing them increased prediction errors. However, a more parsimonious model with significant predictors could be developed without losing predictive performance through merging categories. Only MSIS-29 item 10 and 28 could not be further reduced into smaller categories without losing predictive performance. Consequently 4 variables in the model are not logically ordered, however the largest difference is 0.002 between MSI10_4 and MSI10_5.

Results were similar for the UK value set. This model had, however, slightly larger prediction errors (0.002). The final algorithm contained 10 items, 7 items from the physical dimension and 3 from the psychological dimension.

Table 6.3 Preferred mapping functions

QLQ Model 4				HAQ Model 3				MSIS Model 4			
R^2 (A) =.74				R^2 (A) =.54				R^2 (A) =.58			
R^2 (B) = .73				R^2 (B) = .52				R^2 (B) = .49			
RMSE = .13				RMSE = .15				RMSE = .19			
Predictors	*B*	*SE*	*p*	Predictors	*B*	*SE*	*p*	Predictors	*B*	*SE*	*p*
(Constant)	.978	.008	.000	(Constant)	.527	.113	.000	(Constant)	.956	.021	.000
QLQ1	-.030	.010	.002	HAQ8	-.038	.024	.106	MSI3_5	-.075	.022	.001
QLQ2	-.025	.009	.007	HADS_D	-.017	.005	.001	MSI5_2	-.046	.021	.030
QLQ3	-.045	.010	.000	SF36_PF	.001	.001	.107	MSI5_3	-.055	.025	.029
QLQ4	-.069	.011	.000	SF36_RP	-.001	.000	.032	MSI5_4	-.056	.028	.046
QLQ5	-.159	.016	.000	SF36_BP	.005	.001	.000	MSI5_5	-.129	.035	.000
QLQ6_1	-.037	.010	.000	SF36_RE	.001	.000	.024	MSI6_2	-.062	.025	.013
QLQ6_2	-.077	.015	.000	DAS28	-.019	.012	.113	MSI6_3_4	-.071	.028	.013
QLQ6_3	-.187	.019	.000					MSI6_5	-.134	.036	.000
QLQ7_2_3	-.020	.011	.084					MSI10_2_3	-.049	.016	.003
QLQ9_1_2	-.076	.007	.000					MSI10_4	-.084	.023	.000
QLQ9_3	-.267	.019	.000					MSI10_5	-.082	.030	.006
QLQ23_1	-.020	.008	.015					MSI15_2	-.036	.019	.064
QLQ23_2	-.028	.016	.070					MSI15_3	-.068	.023	.004
QLQ23_3	-.267	.048	.000					MSI15_4	-.068	.025	.006
QLQ24_1	-.071	.009	.000					MSI15_5	-.105	.033	.001
QLQ24_2_3	-.144	.015	.000					MSI21_3	-.031	.018	.084
QLQ27_2	-.041	.010	.000					MSI21_4	-.044	.021	.037
QLQ27_3	-.063	.016	.000					MSI21_5	-.141	.029	.000
								MSI22_5	-.098	.024	.000
								MSI28_4	-.043	.020	.031
								MSI28_5	-.042	.026	.100
								MSI29_2	-.051	.018	.006
								MSI29_3_4	-.070	.020	.000
								MSI29_5	-.204	.029	.000

R^2 (A) = Adjusted R^2 in development sample. R^2 (B) = Adjusted R^2 in test sample. HADS_D= Depression sum score of HADS / SF36_PF=Physical Functioning / SF36_RP=Role-functioning / SF36BP=Bodily Pain/ SF36RE = Role-emotional / DAS28 = Disease Activity Score

Table 6.4 Summary of model performance in test samples

Model	Observed mean EQ-5D (SD)	Predicted mean EQ-5D (SD)	RMSE (normalized for range)	MAE	R^2	Min - Max observed	Min - Max predicted
QLQ-C30							
Baseline	0.66 (.30)	0.66 (.26)	0.16 (12.0%)	0.12	0.75	-0.06 -1	-0.30 - 0.98
2nd treatment cycle	0.70 (.26)	0.71 (.22)	0.13 (9.7%)	0.1	0.79	-0.13 - 1	0.10 - 0.98
4th treatment cycle	0.72 (.25)	0.72 (.21)	0.12 (9.0%)	0.08	0.79	-0.09 - 1	0.13 - 0.98
6th treatment cycle	0.70 (.26)	0.69 (.22)	0.15 (11.3%)	0.1	0.75	-0.13 -1	-0.06 - 0.98
3 months follow-up	0.77 (.26)	0.77 (.21)	0.10 (7.5%)	0.07	0.82	-0.13 - 1	-0.03 - 0.98
6 months follow-up	0.80 (.20)	0.79 (.18)	0.11 (8.3%)	0.07	0.74	0 - 1	-0.03 - 0.98
10 months follow-up	0.77 (.27)	0.80 (.19)	0.16 (12.0%)	0.09	0.68	-0.33 - 1	0.02 - 0.98
18 months follow-up	0.81 (.18)	0.82 (.18)	0.09 (6.7%)	0.06	0.8	0.22 - 1	0.01 - 0.98
HAQ							
HAQ-only model	0.64 (.26)	0.65 (.16)	0.17 (12.2%)	0.07	0.39	-0.13 - 1	0.22 - 0.84
Extended model	0.64 (.26)	0.65 (.20)	0.15 (10.8%)	0.04	0.54	-0.13 - 1	0.13 - 1.02
MSIS-29							
Test sample 1	0.62 (.28)	0.62 (.27)	0.2 (14.4%)	0.16	0.49	-0.13 - 1	-0.05 - 0.96
Test sample 2	0.65 (.23)	0.65 (0.19)	0.18 (12.9%)	0.13	0.49	0.01 - 1	0.06 - 0.96
MSIS-29 (UK)							
Test sample 1	0.57 (.31)	0.59 (.22)	0.22 (13.8%)	0.16	0.49	-0.32 - 1	-0.14 - 0.95
Test sample 2	0.60 (.26)	0.60 (.21)	0.18 (11.3%)	0.13	0.49	-0.17 - 1	0 - 0.95

Discriminant validity

The preferred models were tested for their ability to discriminate between relevant clinical subgroups (known-groups analysis). Results are satisfactory for all the preferred mapping models (QLQ-C30 model 4, HAQ model 3, MSIS-29 model 4 and the MSIS-29 UK model) as presented in table 5. For the QLQ-C30 the predicted values follow a similar pattern to the observed EQ-5D values. Both the predicted and the observed EQ-5D values can hardly distinguish between WHO categories 0 (fully active) and 1 (cannot do heavy physical work but can do everything else), but neither can the self-assessed global health sum score from the QLQ-C30.

Table 6.5 Comparison of predicted and observed EQ-5D index values by clinical indicators

QLQ-C30					
WHO	N (summed)		EQ-5D index	Mapped EQ-5D model 4	QLQ sum score global health
	0	356	0.75	0.75	63
	1	304	0.76	0.74	64
	2	96	0.69	0.72	59
	3	27	0.37	0.42	41
Pearson's r			-0.19**	-0.19**	-0.095**

MSIS-29					
EDSS	N		EQ-5D index	Mapped EQ-5D model 4	MSIS-PHY	MSIS-PSY
	0	9	0.81	0.81	34	21
	1	25	0.73	0.72	40	22
	2	71	0.74	0.70	44	22
	3	51	0.58	0.59	55	26
	4	56	0.63	0.58	57	26
	5	28	0.61	0.54	64	26
	6	53	0.55	0.49	69	28
	7	25	0.40	0.40	74	28
Pearson's r			-0.38**	-0.47**	0.57**	0.23**

HAQ					
DAS28	N	N Model 3	EQ-5D index	Mapped EQ-5D model 1	Mapped EQ-5D model 3
Remission	11	9	0.76	0.76	0.83
Low DA	15	12	0.70	0.68	0.73
Moderate DA	70	59	0.67	0.68	0.67
High DA	27	23	0.51	0.54	0.49
Pearson's r			-0.37**	-0.52**	-0.57**

** *p<.00*

The extended model (model 3) of the HAQ can adequately discriminate between the 4 categories of disease activity (DA). In contrast, the sum score model (model 1) does not discriminate between low and moderate DA. The extended model, which requires data from 3 questionnaires and as a result of missing values in one of those questionnaires, fails to predict 20 of the 123 EQ-5D index scores. The consequence is a large difference between observed (0.76) and predicted (0.83) scores for patients in the 'remission' category due to two missing cases with lower than average utility values (both 0.67).

The MSIS-29 is a slightly different story, as the predicted EQ-5D values have more discriminative ability than the observed EQ-5D between EDSS categories 3, 4 and 5, which was noticed in both test sets and for each of the two EQ-5D country tariffs. Like the predicted EQ-5D values, the sum score of the physical impact of MS also indicates decreasing health per EDSS category. In the distribution graph (fig. 6.2) it is seen that the predicted values do not follow the distribution of the observed values (a similar pattern was observed in the cancer and arthritis data). The EQ-5D seems to have a bimodal distribution with the observations between 0.4 and 0.6 either on the low end of 0.4 or the high end of 0.5. This is most likely the result of few respondents reporting exactly those health problems on the EQ-5D which are transformed into scores between about 0.45 and 0.55 on the EQ-5D tariff.

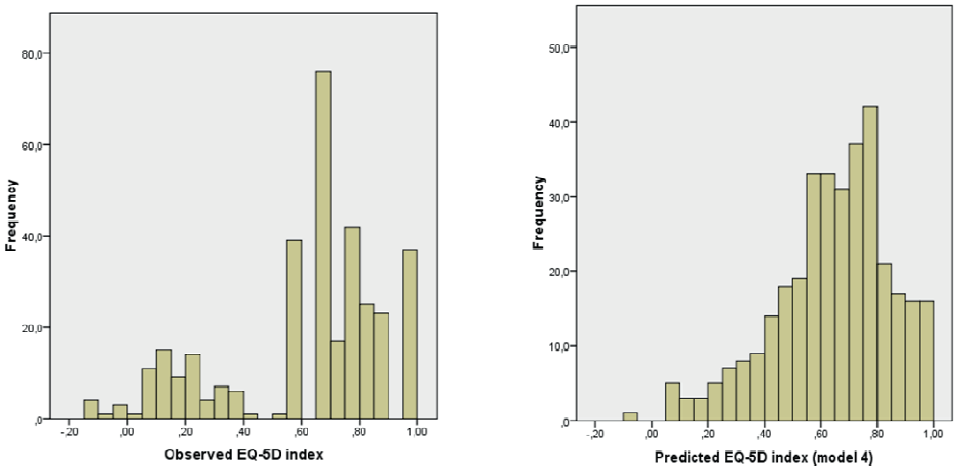

Figure 6.2 Distribution of observed and estimated utilities

6.4 DISCUSSION

This study aimed to develop mapping functions for HAQ, MSIS-29 and QLQ-C30. Quality of the functions was assessed with statistical indicators and performance in relevant clinical subgroups. It was also explored whether the amount of overlap between instruments could explain differences in predictive performance between mapping functions.

The best functioning mapping models are QLQ-C30 model 4, HAQ model 3 and MSIS-29 model 4. Based on the ability of the models to predict mean utility scores for the entire sample and for clinically relevant subgroups these mapping functions seem suitable for predicting utility values to rescue data for economic evaluation when a preference-based measure is absent. When correlations between the starting measure and the EQ-5D were relatively low, the mapping function performed worse. The QLQ-C30 had the highest correlation with EQ-5D dimensions and produced a function with the smallest prediction errors. The HAQ had relatively low correlations with the mobility dimension of the EQ-5D thus seemingly measuring other aspects of mobility. The content of the questionnaires reveal the differences between the measures. The HAQ sum score for mobility is made up by item 8 (walk outdoors on flat ground) and 9 (climb up 5 steps). These questions thus measure the ability to perform a specific mobility related task and differs from the interpretative EQ-5D level 2 'some problems walking about'. It may be due to this discrepancy that 114 respondents report 'no problems' on HAQ item 8, but score 'some problems walking about' on EQ-5D. The HAQ model required additional predictor variables from other questionnaires to successfully discriminate between relevant clinical subgroups.

The preferred mapping functions for the QLQ-C30, HAQ and MSIS-29 successfully predicted mean EQ-5D utility values of the test data-sets. In this study, the measures of error, RMSE, MAE (and epsilon in the box plot), represent the average differences between all the individual predicted and observed EQ-5D index scores. The interpretation of the figures in table 4 is that the mapping functions of MSIS-29 and HAQ-only models are less certain to perform well in samples that deviate 'too much' from the samples on which the models were generated or tested. All models have larger prediction errors for patients with low EQ-5D utility values, as is best represented in figure 6.1. The functions differed in quality, despite successful prediction of the mean and relatively small prediction errors. HAQ model 1 predicts the mean correctly and has a relatively suitable RMSE, but has a small range and cannot distinguish between relevant clinical categories. Drawing on the example of the HAQ mapping function in this study, it seems that only presenting statistical measures is not always sufficient to make an educated judgment about the quality of a mapping function. Additional (clinical) indicators are more easily interpretable and proved to be of added value for assessing the quality of the mapping functions in this study though known-groups analysis.

Improvement of the models was sought by using additional predictor variables, which were only available in the HAQ dataset. Improvement of the model was needed as HAQ model 3 (with the additional predictor variables from the SF-36, the Hospital Anxiety and Depression Scale and the Disease Activity Scale) outperformed the other HAQ models in terms of RMSE, range and ability to discriminate between relevant clinical subgroups. However, here a tradeoff is made between model performance and usability of the model, as it is not likely that many other trials included these additional predictor variables. Even in our own study the amount of missing predicted values by HAQ model 3 was much larger than for the other prediction models.

Several limitations of the study need to be discussed. First and most important, the performance of both the HAQ and MSIS-29 models is likely to be overestimated because test samples were very similar to the development samples as they where randomly drawn from the same original dataset, instead of originating from a different study as was the case with the QLQ-C30 test set. Second, the HAQ was administered in an early arthritis cohort with relatively few very ill patients. Consequently a dummy model could not be estimated. The presence of very ill patients with low utility values is the biggest contributor to prediction errors as these values are generally overestimated. Absence of these values in the HAQ dataset is thus likely to flatter the statistical measures of error. The extended model is recommended for use if predictor variables are available. Third, QLQ-C30 test sample was also a lymphoma type cancer. Performance in subjects with other cancer types is not tested, which consequently causes uncertainty about model performance in different kinds of cancer. Fourth, the quality of the starting measure, the condition specific questionnaires, will have effect on the quality of the mapping function. A notable issue with the MSIS-29 mapping model was that there where many illogically ordered variables. The coefficients of scoring category 2 'a little' were higher or equal to category 3 'moderate'. A recent Rasch analysis of the MSIS-29 suggests that indeed the MSIS-29 would be better of with less answer categories (the authors of the Rasch study suggest 3), as the current 5 could not adequately distinguish the levels 'a little', 'moderately' and 'quite a bit' (22). That issue is a likely cause of the illogically ordered variables in the regression analysis.

QLQ-C30 and HAQ have been previously mapped on $EQ\text{-}5D_{UK\ tariff}$ (124, 133). The QLQ-C30 mapping, developed on a sample (N=199) of patients with inoperable oesophageal cancer, did not report RMSE but reported the adjusted R^2. The adjusted R^2 was 0.61 which is somewhat lower than was found in this study, but the model could successfully predict mean scores. We applied their mapping model, which performed well in our sample (%RMSE range over time points = 11.3 – 7.5). This result may provide some first support for the generalizability of QLQ-C30 mapping models, as %RMSE was relatively favourable. However, a more complete analysis involving different mapping functions tested on different types of cancer is required to

investigate the important and yet unsettled issue of generalizability. The applied model, published by McKenzie and Van der Pol in 2009 (124), was based on sum scores of the QLQ-C30. It is possible that our model, based on items but not all items of all QLQ-C30 dimensions, may be less sensitive in other cancer types. The previous mapping of HAQ on EQ-5D (133) reported %RMSE of a model with dummy variables ranging from 11% to 15%, and a limited range of predicted values (0.2 -0.8). The values presented here do not deviate strongly, even though a dummy-model could not be estimated. In our study the limited range of predicted variables could be overcome by adding additional predictors that cover the other domains of general health, but this does create a problem with generalizability of the model to other datasets that do not hold all variables for instance due to the use of different instruments.

While testing the mapping functions for their discriminant validity it was noticed that MSIS-29 mapping model 4 was more sensitive than EQ-5D between categories 4 to 6 of the EDSS and had higher correlation with the EDSS. It could be argued that a change between levels 4 (able to walk 500 meters without aid) to 6 (assistance, like a cane, required to walk 100 meters without resting) on EDSS does not have impact on quality of life and is therefore not 'picked up' by EQ-5D. However, on these same levels there is a noticeable change on the physical sum score of MSIS-29. It was not anticipated that the predicted EQ-5D would be more sensitive to change in this area of EDSS. We hypothesize that the explanation for the difference is that the predicted and observed EQ-5D index scores have different distributions. The different distributions are most likely the result of only few respondents reporting health problems on the EQ-5D which are transformed by the country tariff into scores between about 0.4 and about 0.6 on the EQ-5D tariff (139). The predicted values of the mapping study have a different distribution than the observed values (figure 6.2) which can be interpreted as an unsuccessful reproduction of true EQ-5D values. However, in this case, it results in an anomalous finding where a mapped EQ-5D index is more sensitive to changes in clinical categories.

The EQ-5D utility index reflects quality of life as measured on five different dimensions of health. Condition specific measures can be used to estimate mean EQ-5D utility values, despite covering different dimensions of health. However, research in this paper suggests that lower degrees of overlap leads to poorer predictive quality. Face value assessment may not necessarily represent true overlap between the dimensions of the condition specific and preference-based instrument. Mapping functions derived from condition specific questionnaires with few dimensions may adequately predict a mean utility value, but should be used with caution in populations that deviate markedly from the population on which the function was estimated. As generalizability is a major issue for mapping functions, it ought to be tested how these models perform in different cancer, arthritis and multiple sclerosis populations. Because errors of the predicted values are larger for patients in poor health, these mapping functions may not perform

well in such populations. Results from this study suggest that the mapped EQ-5D index values of the preferred mapping models can discriminate between relevant clinical subgroups. An important next step is to investigate how using mapped EQ-5D values instead of observed EQ-5D values influences cost-utility analyses.

7

Mapping onto EQ-5D for patients in poor health

With Donna Rowen, John Brazier and Elly Stolk
Health and Quality of Life Outcomes, 2010

ABSTRACT

An increasing amount of studies report mapping algorithms which predict EQ-5D utility values using disease specific non-preference-based measures. Yet many mapping algorithms have been found to systematically overpredict EQ-5D utility values for patients in poor health. Currently there are no guidelines on how to deal with this problem. This paper is concerned with the question of why overestimation of EQ-5D utility values occurs for patients in poor health, and explores possible solutions.

Three existing datasets are used to estimate mapping algorithms and assess existing mapping algorithms from the literature mapping the cancer-specific EORTC-QLQ C-30 and the arthritis-specific Health Assessment Questionnaire (HAQ) onto the EQ-5D. Separate mapping algorithms are estimated for poor health states. Poor health states are defined using a cut-off point for QLQ-C30 and HAQ, which is determined using association with EQ-5D values.

All mapping algorithms suffer from overprediction of utility values for patients in poor health. The large decrement of reporting 'extreme problems' in the EQ-5D tariff, few observations with the most severe level in any EQ-5D dimension and many observations at the least severe level in any EQ-5D dimension led to a bimodal distribution of EQ-5D index values, which is related to the overprediction of utility values for patients in poor health. Separate algorithms are here proposed to predict utility values for patients in poor health, where these are selected using cut-off points for HAQ-DI (>2.0) and QLQ C-30 (<45 average of QLQ C-30 functioning scales). The QLQ-C30 separate algorithm performed better than existing mapping algorithms for predicting utility values for patients in poor health, but still did not accurately predict mean utility values. A HAQ separate algorithm could not be estimated due to data restrictions.

Mapping algorithms overpredict utility values for patients in poor health but are used in cost-effectiveness analyses nonetheless. Guidelines can be developed on when the use of a mapping algorithms is inappropriate, for instance through the identification of cut-off points. Cut-off points on a disease specific questionnaire can be identified through association with the causes of overprediction. The cut-off points found in this study represent severely impaired health. Specifying a separate mapping algorithm to predict utility values for individuals in poor health greatly reduces overprediction, but does not fully solve the problem.

7.1 INTRODUCTION

In recent years there has been an increasing amount of publications concerned with 'mapping' condition specific measures on EQ-5D to estimate EQ-5D utility values. Mapped EQ-5D utility values are accepted as evidence in cost-utility analyses by reimbursement agencies such as the National Institute of Health and Clinical Excellence (NICE) (46) (see § 5.4.6 of the guidance) but suffer from non-trivial problems like the overprediction of utility values for patients in poor health. A mapping algorithm can be estimated by regressing a non-preference-based measure onto a preference-based measure on a dataset external to your study dataset (58). The resulting mapping equation is used to estimate the utility values of the preferenced-based measure in the study dataset where such a measure is absent. Criteria for the quality of a mapping algorithm do not currently exist although it is well known that utilities estimated by mapping algorithms typically have larger errors for lower utility values (58) and mapped EQ-5D utilities show a systematic overprediction of utility values for patients in poor health (140). For instance, a study mapping SF-12 on EQ-5D report predicted values under 0.5 to be notably higher than observed values, for both 2nd and 4th order models (141). Another study, mapping the modified Rankin scale measurement, which assesses disability after stroke, on EQ-5D reports decreased accuracy for patients in poor health and significant overprediction of low values (142). While it is unlikely for such overprediction to be a problem in all samples, given that many studies have reasonably high mean EQ-5D values (143), it is likely to occur in patient (sub) samples containing a significant proportion of individuals in poor health. The current study explores whether the causes of overprediction of utility values for patients in poor health found in the literature can inform a method to minimize that overprediction. The proposed solution involves the use of a different algorithm for patients in poor health, where health status is determined using available information from a condition-specific non-preference-based measure.

There are several causes for the overprediction of low utility values. First, the non-preference based measure may have different severity content than the preference-based measure. For instance, the lowest possible range of scores on the Health Assessment Questionnaire Disability Index (HAQ-DI) is between 2.5 and 3.0 which is not necessarily associated with the lowest value of -.59 on the EQ-5D, but with a value near .1 (144), as the HAQ measures different dimensions of health (114). Adding additional covariates to the mapping functions, like clinical variables or dimension scores of other questionnaires may overcome this problem, but this limits the use of the function to datasets that hold all those variables.

Second, in many clinical studies, health states are not normally distributed: most patients typically experience mild to moderate health problems and few experience severe problems (107, 114). Progression from moderate to poor states, for instance moving from 'some

problems with washing or dressing myself' to 'unable to wash or dress myself', results in a steep drop in utilities. This 'drop' may not be adequately predicted in a linear model which is powered on the large group of patients which reports mild to moderate health problems. This has led to the suggestion that using Ordinary Least Squares regression on the entire sample, which is more accurate for mean values than for extremes, may contribute to the problem of overprediction (58). Specifying other models may lead to better predictions, but will rarely overcome overprediction.

Alternatively, one option is to specify a separate mapping function for patients in poor health whose utility values are overpredicted. Such an approach would require a method to identify the 'poor health' population. A study, mapping SF-36 onto EQ-5D, reported overprediction of utility values for poorer health states (EQ-5D index values < 0.5) for existing algorithms from the literature and algorithms estimated in the study (140). The study hypothesized that this may be observed because more severe health states (utility value <0.5) have at least one of five EQ-5D health dimensions at the most severe level causing the aforementioned steep decline in utility values. Further support for this hypothesis is that in many patient populations a 'bimodal distribution' of EQ-5D utility values is observed. Bimodal distribution refers to the observation of high (>0.5) mean utility values for EQ-5D states with no dimensions at the most severe level and low (<0.5) mean utility values for EQ-5D states with one or more dimensions at the most severe level. This bimodal distribution has a 'gap' in the distribution of EQ-5D utility values around the .5 value (107). This observation is limited to EQ-5D, as prediction errors are also increased for patients in poor health when mapping to SF-6D (145), but no systematic overprediction is present.

This suggests that the alternative mapping function ought to be estimated on the lower part of the bimodal distribution of EQ-5D values. However, as the EQ-5D is absent by definition if a mapping algorithm is applied, it is difficult to assess which predicted values are overpredicted. It is plausible that values can be identified on the condition-specific instrument that are associated with the lower part of the EQ-5D utility distribution, which represents 'poor health'. To this purpose mapping algorithms and datasets for three condition-specific measures, the arthritis Health Assessment Questionnaire (HAQ) and the cancer EORTC's Quality of Life Questionnaire C-30 (version 2) are investigated. When available mapping algorithms systematically overpredict utility values for patients in poor health, it is explored whether it is possible to identify the 'poor health' population by the health status reported on the condition specific measure. If so, we estimate a separate mapping algorithm for use in patients in poor health.

7.2 METHODS

Existing algorithms were applied to one sample of patients with arthritis (146) (arthritis sample) and two samples of patients with cancer: patients with Multiple Myeloma (MH sample) and patients with Non-Hodgkin's Lymphoma (NH sample) (134, 135). A short description of the population characteristics of the samples (pooled data for 8 follow-up time points of QLQ-C30, baseline of HAQ) on which the algorithms were run is presented in table 1. Thus all work presented in this paper was performed using these datasets, limiting generalizability to different types of cancer.

Instruments

The EuroQol EQ-5D is a generic preference-based measure of health related quality of life. It classifies health states on five dimensions (mobility; self-care; usual activities; pain/discomfort and anxiety/depression) with three severity levels each: level one represents no problems; level two represents some problems; and level three represents extreme problems. The classification system defines 243 unique health states which are given a utility score using an existing tariff. The EQ-5D tariff represents the preferences of the general public as elicited using time trade-off, and differs per country. Here the UK tariff (26) and Dutch tariff (75) are used.

The EORTC QLQ-C30 (version two) is a cancer specific questionnaire which consists of 30 items across 6 functioning scales (physical, role, cognitive, emotional, social, global quality of life) and 9 symptom scales (fatigue, nausea and vomiting, pain, dyspnoea, sleep disturbance, appetite loss, constipation, diarrhoea, financial impact). High scores on the functioning and global health status scales reflect good quality of life, while high scores on the symptom scales represent a high level of symptoms (111).

The Health Assessment Questionnaire (HAQ) was first developed for use in patients in rheumatology. The most widely used version of the HAQ assesses the functional ability of patients using 20 items across eight domains (dressing, arising, eating, walking, hygiene, reach, grip and usual activities) (132). Questions are scored on a four level disability scale from zero to three, where three represents the highest degree of disability. Scores for the eight domains are adjusted for the use of aids or devices and averaged into an overall disability index value, HAQ-DI (Health Assessment Questionnaire Disability Index), with a range from zero to three and adjacent steps of 0.125 (e.g. 0, 0.125, 0.250), which represents the extent of functional ability of the patient. A value between one and two represents moderate to severe disability (121).

Table 7.1 Patient characteristics

EQ-5D		N	Mean	% at level 1 / 2 / 3*
Multiple Myeloma population (pooled)				
Age (range)		652	54 (37 - 65)	
EQ-5D	Mobility			56.7 / 41.4 / 1.9
	Self-are			85.8 / 12.8 / 1.4
	Usual activities			30.1 / 51.1 / 18.8
	Pain/Discomfort			39.6 / 59 / 1.4
	Depression/Anxiety			69.4 / 29.6 / 1.0
	EQ-5D utility (UK tariff)		.69 (-.32 - 1)	
Male / Female		381 / 252		
Follow-up series		t=0, 1, 2, 3, 4, 5, 6, 7		
Non-Hodgkin population (pooled)				
Age (range)		789	72 (65 - 84)	
EQ-5D	Mobility			48 / 47.3 / 4.7
	Self-care			81.4 / 13.9 / 4.7
	Usual activities			38.1 / 43.3 / 18.6
	Pain/Discomfort			52.2 / 42.9 / 4.9
	Depression/Anxiety			70 / 29 / 1.0
	EQ-5D utility (UK tariff)		.68 (-.59 - 1)	
Male / Female		480 / 351		
Follow-up series		t=0, 1, 2, 3, 4, 5, 6, 7, 8		
Arthritis population				
Age (range)		457	50 (16 - 88)	
EQ-5D	Mobility			58.5 / 41.5 / 0
	Self-care			75.3 / 24.3 / .4
	Usual activities			37.1 / 58.2 / 4.7
	Pain/Discomfort			9 / 77.4 / 13.6
	Depression/Anxiety			70.7 / 27.1 / 2.2
	EQ5D utility (UK tariff)		.62 (-.24 - 1)	
Male / Female		133 / 333		
Follow-up series		t=0		

Condition specific instruments				
EORTC QLQ-C30 (Sum scores)			HAQ (Domain scores)	
	MM population mean (SD)	*NH Population mean (SD)*	*Arthritis population mean (SD)*	
Physical functioning	64 (24.6)	57.3 (26.8)	Dressing & Grooming	0.58 (.71)
Role functioning	59.5 (28.9)	57.4 (31.5)	Arising	0.65 (.73)
Emotional functioning	82.8 (18.9)	81.3 (20.6)	Eating	0.75 (.82)
Cognitive functioning	82 (20.8)	81.9 (23.7)	Walking	0.54 (.78)
Social functioning	76.2 (25.8)	75.7 (28.6)	Hygiene	0.64 (.81)
Global health	68.7 (18.0)	62 (21.7)	Reach	0.64 (.75)
Fatigue	35.7 (25.0)	44.7 (44.7)	Grip	0.78 (.85)
Nausea / Vomiting	6.1 (14.3)	8 (16.9)	Activities	0.94 (.88)
Pain	25.2 (24.7)	18.7 (26.2)		
Dyspnoea	16.1 (24.9)	24.8 (28.9)		
Sleep	21.1 (27.3)	28.7 (31.8)		
Appetite	16 (27.2)	21.9 (32.6)		
Constipation	4 (15.4)	11.8 (22.8)		
Diarrhea	8.3 (18.7)	7 (18.5)		
Financial difficulties	12.5 (23.0)	6.3 (16.9)		

* EQ-5D: 1 / 2 / 3= no problems / moderate problems / severe problems

Algorithms

Algorithms are taken from the literature and predict EQ-5D index values from either the QLQ-C30 (version 2) or the HAQ. All algorithms have been tested on another dataset with the exception of one HAQ model that was developed for this article, from now on referred to as a test model.

The original articles in which the algorithms were presented labelled them as suitable for estimating utility values (114, 124, 147). Details of the algorithms are presented in table 2. All models were developed using ordinary least squares regression. The HAQ algorithm developed and tested by Bansback et al. (147) was estimated on patient samples from Canada (N=319) and the United Kingdom (N=151) who were clinically diagnosed with rheumatoid arthritis (RA). The algorithm computes EQ-5D utility values based on the UK tariff. We estimated one additional HAQ algorithm, the test model, for this article based on a larger group of patients than was used for the published algorithm, as this sample holds more patients in severe conditions (114). The test model was developed using the Rotterdam Early Arthritis Cohort with 493 patients with and without clinically diagnosed RA recruited from the Erasmus Medical Centre in the Netherlands. It is not recommended for use as not all patients are clinically diagnosed with RA. A tested HAQ model that predicts Dutch utilities is presented elsewhere (114). The

QLQ-C30 algorithm by McKenzie & Van der Pol (124) was developed on a sample of 199 patients with inoperable esophageal cancer. The algorithm computes EQ-5D utility values based on the UK tariff. The QLQ-C30 algorithm by Versteegh et al. (114) was developed and tested on pooled data from two clinical trials for patients with multiple myeloma (pooled N=723) and patients with aggressive non-Hodgkin's lymphoma (pooled N=789). It computes EQ-5D utility values based on the Dutch tariff.

Table 7.2 Mapping algorithm specifications

Measure		Algorithms
HAQ	Bansback (2006)[1]	EQ-5D index (UK tariff)= .80 + (h1_2*-.15) + (h4_1*-.08) + (h4_2*-.12) + (h4_3* -.59) + (h6*-.15) + (h7_1*-.04) + (h7_2*-.08) + (h8*-.10) + (h9*.12) + (h13* -.14) + (h16*.07) + (h23*-.05) + (h24_1*-.05) + (h24_2*-.11) + (h26_2*-.14) + (h26_3*-.13) + (h27_2*-.08) + (h27_3*-.20) + (h30_1*-.05) + (h31_1*-.07) + (h31_2*-.08) + (h32*.09)
	Test model[2*]	EQ-5D index (Dutch tariff)= 0,858 + (haq1* -0,027) + (haq2*-0,035) + (haq3*-0,025) + (haq4*-0,033) + (haq5*-0,001) + (haq6*-0,035) + (haq7*-0,031) + (haq8*-0,057)
QLQ-C30	McKenzie (2009)[3]	EQ-5D index (UK tariff)= .2376 + (ql*.0016) + (pf*.0004) + (rf*.0022) + (ef*.0028) + (cf*.0009) + (sf*.0002) + (fa*-.0021) + (nv*.0005) + (pa*-.0024) + (dysp*.0004) + (sleep*.00004) + (eat*.0003) + (obsti*.0001) + (diarr*-.0003) + (finan*-.0006)
	Versteegh (2012)[4]	EQ-5D index (Dutch tariff)= 0.985 = (1*-.037) + (2*-.025) + (3*-.059) + (4*-.033) + (5*-.134) + (6_2*-.033) + (6_3*-.067) + (6_*-.180) + (7_2*-.013) + (7_3*-.037) + (7_*-.012) + (9_l2*-.065) + (9_3*-.053) + (9_*-.189) + (16_2*-.038) + (16_3*-.045) + (16_*-.126) + (23_2*-.028) + (23_3*-.049) + (23_*-.456) + (24_2*-.053) + (24_3*-.140) + (24_*-.232) + (27_2*-.027) + (27_3*-.091) + (27_*-.110)

[1] HAQ items as dummy variables: h1 = dressing & grooming; h4 = arising; h6-7 = eating; h8-9 = walking; h13-16 = aids or devices; h23-24 = hygiene; h26 = reach; h27-28 = grip; h30-32 = activities. (e.g. h1_2 = haq item one, answer level two).
[2] HAQ sum scores: haq1 = dressing & grooming; haq2= arising; haq3 = eating; haq4 = walking; haq5 = hygiene; haq6 = reach; haq 7 = grip; haq8 = activities.
[3] QLQ-C30 sum scores: ql = quality of life; pf = physical functioning; rf = role functioning; ef = emotional functioning; cf = cognitive functioning; sf = social functioning; fa = fatigue; nv = nausea & vomiting; pa = pain; dysp = dyspnea; sleep = sleeping; eat = eating; obst = obstipation; diarr = diarrhea; finan = financial difficulties.
[4] QLQ-C30 items as dummy variables: 1 to 5 = dichotomous items; 6 to 27 = four level items.
* Not tested on external data-set.

Thus, all models used in this study were taken from other studies. Despite their use to investigate our methodological point, generalizability of mapping functions between different types of cancer or arthritis is an empirical matter that still needs thorough investigation.

Analysis

First we determined if the mapping algorithms estimated on a relatively healthy patient sample overestimate utility values of patients in poor health. As the EQ-5D is absent by definition, we need to specify a threshold value on the condition specific measure for which we would expect a regular mapping algorithm to overpredict utility values to be able to anticipate whether a mapping algorithm is expected to be inaccurate in a certain population. Then we developed a mapping algorithm for that population. Five steps are described below, aimed at systematically exploring the topic.

Step one. Each published algorithm used here was found in its original article to be successful at predicting mean EQ-5D values. The same diagnostics have also been applied to the test model and indicate this model is successful at predicting mean EQ-5D values. However, a successful prediction of a mean EQ-5D utility value in a sample with a relatively high mean value does not guarantee a successful prediction in a sample with a much lower mean EQ-5D value. Therefore we compared the predicted values to the observed values over the range of observed EQ-5D values.

Step two. It has been suggested that reporting a level '3' answer on EQ-5D and the large utility decrement associated with it in the EQ-5D country tariff is a cause of overprediction (140). Using the UK tariff (26) an EQ-5D utility value of .52 is the lowest obtainable value without a level 3 answer (state 22222), and 0.56 is the highest obtainable value with a level 3 answer (state 11311), which is respectively 0.57 and 0.64 for the Dutch tariff. These values were used to interpret the distribution of utility values in the three samples.

Step three. The frequently observed bimodal distribution of utility values in patient samples has been associated with 'N3-term' (107) and the bimodal pattern has been presented by others as a specific feature of the EQ-5D (148). The N3 term is a model feature of the UK and Dutch EQ-5D country tariff and adds an extra utility decrement if any dimension on the EQ-5D scores a '3', representing extreme problems. However, it was hypothesized here that the 'N3' in itself does not cause a bimodal distribution. To test this, a random set of EQ-5D cases was generated (N=300) with an equal distribution of answer categories across the 5 domains.

Step four. Step one and two investigate whether the utility values of patients who report 'extreme problems' on at least one of the EQ-5D dimensions are overpredicted. The next step is to investigate which QLQ-C30 and HAQ value is associated with level '3' answers on the EQ-5D. The purpose of this exercise is to identify scores on the condition specific measure that are related to a possible cause of overprediction in mapped utility values: at those scores standard mapping algorithms might be inaccurate. As the QLQ-C30 provides no overall score, the functioning scale scores are used, since these have the highest correlation with EQ-5D scores (123). For the HAQ, the HAQ-DI value (which combines all items) was used.

Step five. The next step is to explore the performance of a separate algorithm for use on patients in poor health. An alternative algorithm was developed on a sample in poor health, in this case on a within sample selection of patients which are in poor health as determined by the cut-off point identified in step 4. The utility value of the EQ-5D, using the UK tariff, was regressed on the disease specific questionnaires. In the cancer population the algorithm was developed on the multiple myeloma sample and tested on the non-Hodgkin's sample. A variety of model specifications are estimated using OLS. All algorithms were applied at the individual level. Mean utility values were used to compare predicted and observed values.

7.3 RESULTS

All mapping algorithms applied here suffer from overprediction at the lower end of the scale, where predicted values are higher than observed values for observed EQ-5D utility values below ≈.5. Figure 7.1 compares predicted and observed EQ-5D utility values, and are representative for the other mapping algorithms investigated in this study.

Step one. Figure 7.1 indicates that overprediction begins to occur around EQ-5D utility value ≈.5. As is mentioned in the method section: the utility value of ≈.5 is related to the scoring 'extreme problems' on any EQ-5D dimension. Patients that have one or more dimensions at level 3 have a maximum observed $EQ\text{-}5D_{UK\ tariff}$ score of 0.56 in the MM and NH samples and of 0.43 in the Arthritis sample. Patients that have no dimensions at level '3' have a minimum observed $EQ\text{-}5D_{UK\ tariff}$ score of 0.52 in all samples (state 22222). A utility value of 0.52 and lower guarantees the presence of at least one level 3 answer in the UK tariff. Scores higher than 0.52 but below 0.57 do not guarantee the absence of at least one level 3 answer.

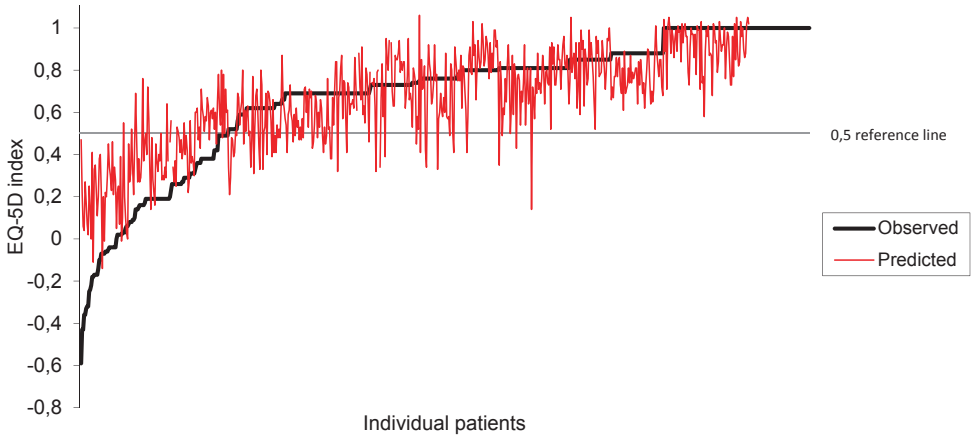

Figure 7.1 Overprediction of mapping algorithm

Step two. Minimum and maximum EQ-5D scores of patients with or without at least one dimension at level 3 on the EQ-5D inform our interpretation of figure 7.2, which indicates the bimodal distribution the arthritis sample, which is representative for the other samples. A patient with a 'level 3' answer on EQ-5D belongs to the left side 'poor health' distribution with a lower mean and less frequent observations than a patient without a 'level 3' answer. The area around a utility value of .5 can fall under either distribution, as indicated by the overlap in minimum and maximum values for cases with and without level 3 answers mentioned in step one.

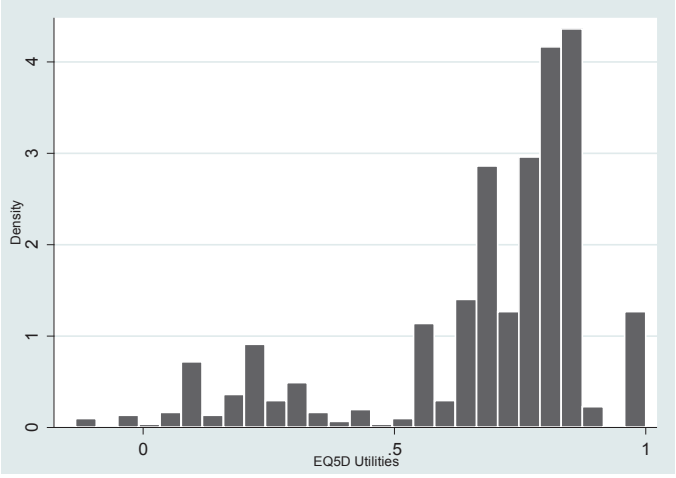

Figure 7.2 Bimodal distribution of utility values in arthritis sample

Step three. The randomly drawn distribution of EQ-5D answers have a normal distribution of utility values, suggesting that the bimodal distribution is not solely caused by the 'N3' term. The random sample (N=300) had 163 unique health states. The 34 most frequent health states account for 36% of the observations, which is in stark contrast to the other samples. The NH sample (pooled N=783) had 78 unique health states of which six states accounted for 53.5% of all observations. The MM sample (pooled N=716) had 59 unique states of which seven states accounted for 62.1% of observations. The Arthritis sample (N=488) had 49 unique states of which seven states accounted for 64% of the data. The combination of the EQ-5D country tariff and distribution of responses across severity levels seem to be the cause of the bimodal distribution of EQ-5D utility values. Few people have level '3' answers, many have level 1 or 2 answers and only a small amount of states cover most of the observations.

Step four. Mapping algorithms overpredict utility values under 0.5, which are for patients with 'extreme problems' on at least one of the five EQ-5D dimensions. This means that mapped utility

values are inaccurate for those patients with scores on the condition-specific measure that are associated with an EQ-5D utility value below 0.5. However, scores on the HAQ and QLQ-C30 do not provide a straightforward indication of the accuracy of the use of a mapping algorithm. For example, a patient average on the QLQ-C30 functioning scales of 70 could belong to an EQ-5D utility value between as low as .21 or as high as 1. At least half of the patients with an average value of the QLQ-C30 functioning scale lower than 55 have level 3 answers on the EQ-5D. Although it is a somewhat arbitrary cut-off point, an average of 45 on the functioning scales is a clear indication of the expected overprediction of a mapping algorithm, for at that value approximately 86% of patients in these samples have at least one level 3 response.

The HAQ-DI values faced similar problems: a HAQ-DI value of 1.5 (moderate to severe disability) can be associated with an EQ-5D utility value as low as .21 to .3 or as high as .71 to .8. At HAQ-DI values <1.6, over 50% of patients have at least one level 3 response on the EQ-5D. A HAQ-DI > 2.0 is a clear indication of the expected overprediction of a regular mapping algorithm, for at that value, approximately 72% of patients in this sample has at least one level 3 response.

Step five. The within sample population of cases in poor health (QLQ-C30 <45, HAQ-DI > 2.0) was relatively small (N=18 Arthritis sample, N=25 at t=0 NH-sample, N=40 at t=0 MM-sample). Within those subsamples, EQ-5D was regressed on QLQ-C30 and HAQ using a variety of regression model specifications. The mapping model was developed on the MM-sample, and tested on the NH-sample. The QLQ mapping algorithm contained 5 items after backwise selection, and included items as categorical variables. The mapping algorithm was applied on the NH sample for patients with QLQ average on the functioning scales < 45. In comparison to the standard mapping algorithms, the utility model for patients in poor health outperforms the model from the literature (table 3) for this selection of the sample and reduces root mean square error by .06 in the first 4 time points. As can be seen from the maximum score in table 3, 1 individual did not seem to have filled in the EQ-5D correctly and had a utility value of 1 (but a low score of 25 on the EQ-5D visual analogue scale). A similar pattern was observed for the last four time points, but not deemed trustworthy due to small sample size (N<8 for the last 4 time points of the QLQ-C30 follow up data). The predicted values showed less prediction error than the standard mapping algorithms, but still did not accurately predict mean utility samples in this selection of the sample with root mean squared error of 0.18.

For the REACH study, only a development dataset was available but for both cut-off points (HAQ-DI>1.6 and HAQ-DI >2.0) the regression model was underpowered with no significant predictor variables due to the small sample size and low correlations between HAQ sum scores and EQ-5D utilities.

Table 7.3 Predicted and observed values in N-H population with QLQ-C30 < 45

Time point		N	Minimum	Maximum	Mean	Std. Deviation
Baseline	Observed EQ-5D	25	-.36	1.00	**.18**	.39
	Predicted McKenzie & Van der Pol	24	-.14	.56	**.25**	.15
	Predicted 'low utility model'	25	-.34	.54	**.14**	.22
T=1	Observed EQ-5D	17	-.43	.64	**.16**	.26
	Predicted McKenzie & Van der Pol	17	-.01	.62	**.32**	.17
	Predicted 'low utility model'	17	-.07	.29	**.14**	.11
T=2	Observed EQ-5D	16	-.33	.38	**.10**	.18
	Predicted McKenzie & Van der Pol	16	.16	.56	**.32**	.12
	Predicted 'low utility model'	16	-.03	.41	**.16**	.11
T=3	Observed EQ-5D	13	-.24	.31	**.07**	.17
	Predicted McKenzie & Van der Pol	13	-.01	.55	**.31**	.14
	Predicted 'low utility model'	13	-.17	.42	**.18**	.17

7.4 DISCUSSION

This paper explored causes of EQ-5D utility values for patients in poor health when mapping from a non-preference-based measure, and investigated a possible solution to the problem. We examined the association between the cause of the overestimation and values on the condition specific questionnaire at which overprediction occurs. Our findings suggest that the main cause of overestimation is a combination of the large decrement in utility values in the UK and Dutch EQ-5D tariffs for having one or more dimensions at level '3', along with few observed responses at level '3'. We argue that this, alongside the large number of EQ-5D responses at the least severe level, leads to a bimodal distribution of the utility data. A result is that the most linear prediction models cannot adequately describe low utility values. We found that the values on the condition specific questionnaire can help inform decisions about the expected errors and hence accuracy of standard mapping algorithms, and that the use of a separate mapping algorithm specified for patients in poor health reduces the amount of overprediction for these patients.

Our findings, in accordance with the literature, suggest that the ≈.5 value of the $EQ\text{-}5D_{UK\ tariff}$ is the point at which mapping algorithms start to overpredict utility values. The reason it is the ≈.5 is due to the fact that values under ≈.5 belong to patients who have extreme problems on at least one dimension of EQ-5D. As the purpose of mapping algorithms is to predict EQ-5D values when EQ-5D was not included in the trial, such a value is not informative for the application of mapping algorithms. Here we explored the use of condition specific measures (that we are mapping from) to indicate the expected accuracy of a standard mapping algorithm. An alternative mapping algorithm can then be developed for use in patients in poor health. We found that the ≈.5 utility value itself is not a very useful measure of association with QLQ-C30 or HAQ-DI values, since there is not a one-to-one relationship between measures meaning that a large range of QLQ-C30 and HAQ scores are associated with the ≈.5 EQ-5D value. Since scoring a '3' on the descriptive system of EQ-5D is related to the problem of overprediction, we took an alternative approach using the scores on the condition-specific measure that correspond to having at least one level '3' response. Below a QLQ-C30 average of the functioning scale of 55, about half of the patients scores level 3 answers on the EQ-5D, as do patients with HAQ-DI > 1.6. At these scores, standard mapping algorithms are likely to overpredict utility values. More conservative and somewhat arbitrary cut-off values we determined are > 2.0 for HAQ-DI and < 45 for the average of the QLQ-C30 functional scales. These cut-off points represent very severe health problems: 45 for the QLQ-C30 is associated with severe cases like post-radiotherapy patients with metastatic and/or cardio-respiratory disease (149); a HAQ-DI value under 2.0 represents severe to very severe RA (121). At these more conservative values, a standard mapping algorithm is likely to be inaccurate.

A separate utility mapping algorithm estimated on a sample with poor health status is far better at predicting utility values for patients in poor health, when it is possible to estimate such a function (see table 3). However, using categorical variables introduced problems with perfect collinearity in the low utility model, and the HAQ sample did not allow the estimation of a low utility model due to poorer correlation with EQ-5D and smaller sample size than QLQ-C30. A model based on sum scores did not suffer from these restrictions but introduced larger prediction errors. The result is a model for low utilities that only uses 5 items of the QLQ-C30 as predictor variables. Item 3 (trouble taking a short walk), 4 (need to stay in bed or a chair), 5 (need help with eating, dressing, washing or using the toilet) 9 (pain) and 21 (feeling tense) together represent physical functioning, emotional functioning and pain. Consequently other quality of life drivers such as role functioning or fatigue are not represented which may lead to problems when applying the function in other cancer types. Furthermore, OLS models used in all mapping algorithms reported here are more precise around mean values than for extremes, which results also in underprediction for utility values near to 1, most notably when regressing EQ-5D on HAQ. Thus estimating and applying mapping algorithms on datasets with large

deviations in health status is likely to be problematic. The extent to which a deviation can be considered 'large' is difficult to assess, since it depends on how a change on the scale of the questionnaire relates to a change on the EQ-5D index values.

Cut-off points like the ones specified in this study can be used to inform whether a regular mapping algorithm from the literature would suffice or whether a 'low utility algorithm' is better at assessing the quality of life for those patients. Cut-off points can indicate whether there are patients in poor health and therefore whether predicted utility values are likely to suffer from overprediction if only a standard mapping algorithm has been used. Cut-off points can therefore inform users and policy makers whether mapped estimates should be treated with great caution. A weakness of the approach may be that there is no clear cut relation between the break point of utility values in the distribution and values on the condition specific measures. Besides, prediction errors might be reduced even more if there were several mapping functions for each 'severity group'. However, the relation between the condition specific measure and the preference-based measure may not be clear cut enough to identify more sub-groups.

Although overprediction proved to be less of a problem for patients in poor health with our combined prediction model, the largest part of the sample is not in very poor health. Nevertheless, predicted EQ-5D values do not capture the full range of observed EQ-5D values due to overprediction. As a consequence, they have 'tighter' confidence intervals around the QALY estimates.. In probabilistic sensitivity analysis this results in less uncertainty around the estimate of cost per QALY, but that is an incorrect representation of reality. In addition to the tighter confidence intervals, using mapped utility values may result in an underestimation of the utility-gain between time intervals. As the utility values of patients in poor health are systematically overpredicted, individuals who in reality would improve from poor health to better health (i.e. from a value <0.5 to a value >0.5) would have an underestimated utility gain when using mapped EQ-5D utilities.

A main point of concern in any effort to map onto a preference-based questionnaire is generalizability of the results. As mentioned earlier, it must be stressed that although the cut-off points presented here are empirically supported by our study, they cannot be considered transferable or generalizable to other types of cancer or arthritis samples prior to thorough empirical testing in different datasets.

The issue of generalizability also applies to the presented methodology. This study focused on mapping onto EQ-5D for patients in poor health. The methodology proposed here only applies to mapping onto EQ-5D using the UK or the Dutch country tariffs. We observed that individuals who report 'extreme problems' on one of the five EQ-5D dimensions receive

overestimated utility values from published mapping functions. Our suggestion is that this is caused by the large utility decrement applied to scoring 'extreme problems' in the UK and Dutch EQ-5D country tariff, combined with only a few observations of 'extreme problems'. However, other EQ-5D country tariffs may not have large utility decrements for all 'extreme problems' scores. For instance, the total decrement for scoring 13111 ('extreme problems' on the self-care dimension of EQ-5D) has a total utility decrement of 0.564 in the UK tariff and 0.254 in the Japanese tariff. These differences in preferences between populations may be of influence on the methodology used to identify the part of the population which is in poor health and where increased prediction errors are observed. However, if those patients can be identified, specifying a separate mapping function for that part of the populations is still a suggested option to reduce prediction error.

Further research is needed to determine if specifying more mapping functions for different severity classes is to be favored over other approaches. For instance, the problem mentioned above about the limited number of items available due to collinearity may be solved by using a larger dataset which provides more accurate predictions for summed scores. The approach could also be undertaken using regression techniques such as the probit model and a two-part model and this is an area for future research. An obvious attempt would be to raise variables to a power to allow non-linearity, but a recent study still reported overprediction under a utility value of around .6 for a model with significant second order predictors (125). Alternatively, stepped linear regression with a specified break-point may allow the utility function to 'curve' according to observed values, but specifying such a breakpoint is not clear cut as is shown in this study.

7.5 CONCLUSION

As the use of mapping in cost-effectiveness analyses of medical interventions is becoming more frequent, guidelines on the appropriateness of using mapping and specific mapping algorithms are needed. We investigated the often observed problem of overprediction in mapping and analyzed the use of cut-off scores for the condition specific measures QLQ-C30 and HAQ-DI to indicate when the use of a separate mapping algorithm for patients in poor health is the favored approach. Overprediction of utility values for patients in poor health can be greatly reduced by predicting the utility values of these patients using a separate mapping algorithm specified and estimated specifically for these patients, when deemed necessary.

8

Discussion

Health care expenditures are increasing in most countries and The Netherlands is not an exception. The availability and use of new medical technologies is considered to be an important driver of the increase in health expenditures (1). In this context, policy makers face the challenging task to allocate an optimal amount of resources to the health care sector and to ensure that these resources are used optimally within the health care sector. Both decisions receive considerable societal attention. Economic evaluations aim to inform policy makers regarding the efficiency of medical interventions, and can be used to assess whether the implementation of (new) health technologies and interventions is welfare improving.

In economic evaluations of medical interventions, a policy option is considered to be 'welfare improving' if the benefits associated with it outweigh the associated costs. In a cost-utility analysis (CUA), a popular form of economic evaluation of medical interventions, benefit is primarily captured using the Quality Adjusted Life Year (QALY) model. Hence, in CUA, a medical intervention is welfare improving if the cost per QALY is lower than what a society is willing to pay for that QALY.

The QALY is an outcome measure capturing length of life as well as quality of life. The past decades have shown that none of the available methods for measuring the quality adjustment part of the QALY is without problems (12). These problems relate to the source of preferences for health states (i.e. quality of life values), biases in valuation methods such as the TTO, descriptions of health states which may lead to insensitivity of instruments and the methods applied in estimating quality of life values when quality of life was not measured with preference-based instruments. Overcoming these problems is an important challenge and should ultimately improve QALY calculations and hence the validity of CUA as a source of information for policy makers. This thesis therefore addressed the overarching question: *How can the measurement and valuation of health for QALY computation in economic evaluations be improved?*

This thesis was structured around four research questions. Each of these questions focused on one of the key problems in the measurement and valuation of health. In this section, the main conclusions are presented and discussed, limitations are acknowledged, and research challenges for the future are highlighted.

8.1 RQ 1) Whose preference values should determine the value of health states?

In the QALY metric, life years are adjusted for the Health Related Quality of Life (HRQoL) experienced during those life years. Generally, HRQoL is equated with preferences for health states, i.e. the relative desirability of living in some state of health. This leaves open whose preferences should be used in deriving health state valuations. Often, the general public is asked

to express their preferences for health states and much used tariffs for generic instruments are based on such preferences. In other words, the general public determines the relative desirability of health states, and their corresponding value, anchored on a scale from 1 for perfect health to 0 for 'dead'. This is subsequently used to correct life years lived in different health states in QALY calculations.

Preferences for health states of patients generally differ from those observed in the general public. Over time, the feelings of patients about their state of health may change, but such information is not reflected in ex-ante health state valuations by the general public. Indeed, on average, patients seem to consider their own state of health to be more desirable (relative to perfect health) than how desirable that health state is in the view of the general public (49-52), although opposite observations also exist. As a result, health state values, and possibly the total benefit assigned to treatments in terms of QALYs, may differ depending on the source of preference values. The health economics literature acknowledges the discrepancy between preferences for health states of patients and the general public. Nevertheless, the standard practice in economic evaluations is to base the value of health on preferences obtained from the general public. For example, the most often used questionnaires to measure the benefit of treatments in economic evaluations (EQ-5D, SF-6D and HUI-III) transform measured health profiles into health state values using tariffs based on general public preferences.

Three key arguments for the use of general public preferences often put forward in the literature are 'adaptation' (48), the 'insurance principle' (84) and the 'societal perspective' (48). In chapter 2, it was contested that these arguments are sufficiently convincing, both combined and in themselves, to support the conclusion that general public preferences should solely determine the value of health states.

In contrast, it was concluded that completely disregarding either patient preferences or general public preferences cannot be based on the arguments put forward in the literature. More importantly, chapter 2 provided theoretical examples to show that using societal preferences may result in either higher or lower estimates of cost-effectiveness when compared to using values obtained in patients. If we assume that, due to adaptation, patients on average assign higher values to health states than the general public does, any additional life year lived receives 'more weight' with patient preferences than with general public preferences. As a consequence, life extending procedures, in conditions to which patients may adapt, such as paraplegia, result in a lower cost-effectiveness estimate when the quality adjustment of the QALY is based on patient preferences.

To date, the normative discussion concerning the appropriate source of preferences was focused on the discrepancy between patient preferences and general public preferences, framing the debate in a choice for either patients or the general public as a source of preferences. In The Netherlands and the United Kingdom, guidelines stipulate the use of general public preferences to determine the value of states of health, effectively excluding the use of patient preferences. However, it could be argued that any discrepancy between patient and general public preferences conveys that the preferences obtained in the two groups are potentially complementary sources of information on the desirability of states of health. Where general public preferences provide an ex-ante preference value for being in a state of health, patient preferences inform decision-makers about preferences of patients actually experiencing those states.

A limitation to the discussion of chapter 2 is that the literature was not gathered in a systematic fashion. It is therefore possible that some publications that contain relevant arguments for using general public preferences were missed. Nonetheless, the arguments that are most prevalent in the health economics community have been included in chapter 2.

Future avenues of research regarding RQ 1

How patient preferences for health states can be included in estimates of the benefit of medical interventions is an important avenue for future research. It is doubtful whether it is ethically acceptable to present structured choice tasks with questions about life and death to very ill patients (4). Nonetheless, the aim of chapter 2 was to reposition the experienced HRQoL by patients as a relevant source of information for societal decision making. The findings suggest that patient preferences *could* be used in assessing the benefit of health interventions, next to general public preferences. *How* that should be done was outside of the scope of chapter 2, but is an important area for future research. Several ways of including experienced HRQoL in health state valuation have already been explored. For example, respondents from the general public have been informed about the experiences of real patients prior to valuing health states (150). Also, it has been advocated that the sample from which respondents are selected should contain both patients and members of the general public (151). While these methods are to some extent appealing, they both try to incorporate experienced and expected HRQoL in a single index value. Following the argumentation in chapter 2, it may be more appropriate to treat patient and general public valuations as separate entities, rather than to try and incorporate both sources of value in one single index. After all, the differences between the two sources of valuation provide important information on expected and experienced utility, as well as the discrepancy between them.

A new research strategy to better understand the differences between patient and general public preferences is to estimate the impact that experienced health has on the valuation of hypothetical health. This may be achieved by asking respondents to indicate their own health state on a particular descriptive questionnaire, such as EQ-5D, and to consequently assign preferences to their own health state, for example using TTO. Then, respondents can assign preference values to hypothetical states of health which are described with the same questionnaire (i.e. EQ-5D).. The resulting dataset captures the potential adaptation, since a comparison can be made between health state values for own health, and health state values of others for that state of health. The dataset can also be used to model how owns own health state and the feelings about that health state affect ones feelings about the severity of other health states. It could, after all, very well be that preference values for health states are related to the 'distance' between ones current health and the health state under valuation. For example, perhaps health states are only considered very poor when they represent a large health loss compared to one's current health state. Such a study could help researchers understand to what extent a 'reference point' (i.e. one's own health) impacts the valuation of hypothetical health.

8.2 RQ 2) How can the TTO exercise be improved for the measurement of preferences for health states worse than dead?

Next to the question 'whose preference values should reflect the HRQoL of a health state' the question of 'how to measure preferences values for health states' is also important. One of the most used methods to elicit preferences for health states is the Time Trade-Off method (TTO). It has been used, for example, for the valuation of health states described with the widely used EQ-5D questionnaire. The TTO method involves asking respondents to trade-off living shorter in good health and living longer but in impaired health. The duration of the period in good health is varied in conventional TTO exercises, so that a point of indifference can be found. Essentially, if respondents prefer a relatively short period in good health to a long period in impaired health, it indicates that the impaired health state is considered to be very poor (i.e., of low value). With the TTO method, the value of an impaired health state can be expressed relative to the value of perfect health. Therefore, one year in impaired health may be equivalent to, for example, 0.8 years in good health. This 0.8, then, represents the quality adjustment part of the QALY for that particular health state.

Health states can be so poor that living in such a state of health for a period of, say, ten years, is considered to be worse than not living at all. These states of health are thus considered to be 'worse than dead' and are assigned a negative value (e.g. -0.2). The TTO has been criticized for using two separate valuation procedures; one for health states better than dead and one for health states worse than dead (27, 42, 43). Since this may lead to incomparability of valuations,

uniform procedures enabling valuations of both health states worse than dead and those better than dead have been developed. These procedures are labeled 'lead time TTO' and 'lag time TTO'. In the lead time TTO, the health state for which preference values are sought, is preceded by a period of good health, the so-called lead time. In the lag time TTO, this period of good health is placed after the period in the health state under valuation. As a consequence, one procedure and one utility equation can be used for calculating the health state preferences of both better than dead health states and worse than dead health states. In this thesis, the lead time TTO and lag time TTO were compared with the 'classic' TTO in an online experiment with 5208 respondents. Preference values elicited with lag time TTO were consistently lower than those elicited with lead time TTO. These novel methods may be more sensitive to behavioral characteristics such as time preferences than the classic TTO, as health states are explicitly placed in the future (lead time TTO) or in the present (lag time TTO).

The issue of the timing of health states is discussed in chapter 4, which describes a study based on the same data as collected in chapter 3. In chapter 4, the timing properties of lead time TTO and lag time TTO were used to quantify time preferences of respondents. Since lead time TTO and lag time TTO were identical in all aspects accept for the timing of the (again identical) health state, it could be argued that any difference in health state values is explained by the value respondents place on the timing of events. Normally, individuals in decisions place more weight on events occurring in the near future than on those occurring later. . A common way to capture time preferences is in a discount function, which indicates how individuals value timing. In chapter 4 a hyperbolic discount function and an exponential discount function were applied to see which function fitted the data best. Respondents indicated positive time preferences, indicating that they prefer the onset of disease to occur later rather than sooner, and both discount functions fitted the data equally well. The importance of chapter 4, hence, is that it is shown that lead time TTO and lag time TTO are influenced by preferences for timing of events, that these preferences can be measured and subsequently used to convert lead time TTO values into lag time TTO values and vice versa.

Lead or lag time TTO procedures may provide an improvement compared to the conventional two-step TTO procedure, due to the uniform method of measuring health state preferences better and worse than dead. They have been shown to be feasible in this thesis and elsewhere (e.g. (42)), but empirical tests are still scarce. For example, there is only one other study comparing lead time TTO to lag time TTO (91), and in that study, lead time TTO values were lower than lag time TTO values for very severe health states. This is opposite to what was found here and hence, results are mixed concerning the effect of applying either lead or lag time TTOs. In light of the contradictory available evidence and the results presented in this thesis, it seems preliminary to suggest that lead and lag time, although solving specific problems of

the conventional TTO, actually perform better than the conventional TTO and hence should replace it. Together, chapter 3 and 4 emphasize that even slight alterations to the TTO method can have large consequences for health state values and, by extension, for the benefit attributed to medical treatments in CUA.

There were some noteworthy limitations to design of the study presented in chapter 3, which also apply to chapter 4 as it was based on the same data. Although the software used in the TTO study was tested numerous times and included extensive instructions, the unsupervised nature of the online TTO experiment may have limited both the comprehension of the task and the level of engagement with which it was completed. Differences observed with other TTO studies (e.g. number of iterations used in the TTO procedure (96)) suggest that face-to-face TTO interviews provide superior data quality to unsupervised online TTO interviews.

Another limitation of the TTO study presented in chapter 3 was its 'between-subject design' design rather than a 'within-subject design'. This means that respondents filled out lead time TTO *or* lag time TTO, rather than both. The differences in health state values could thus be caused by differences between respondents, rather than by the specifications of the TTO. Since the sample size was rather large (approximately 1000 respondents for each valuation methodology), and since respondents were randomly assigned to a TTO specification, the chance that differences in health state values are caused mostly by respondent characteristics is small, but the possibility can only be excluded by more extensive multivariate tests where respondent characteristics are included as explanatory variables.

The between-subject design posed some specific challenges for the research in chapter 4. Due to this design, it was not possible to estimate individual discount parameters since not one individual had completed both the lead time TTO and the lag time TTO task. As a consequence, discount parameters were measured using the mean health state value of one group of respondents versus the mean health state value of the other group of respondents. Due to the nature of mean values, it is perfectly possible that a large set or respondents had negative or neutral time preferences, but that the mean discount value was positive as other respondents had larger positive discount values, thus distorting the mean towards a positive discount value.

Besides the limitations of the study design, there are also limitations inherent to the formulation of the research question. The research question focused on improvements *within* the TTO exercise, rather than *alternatives* to the TTO exercise, while in fact, there are other valuation techniques available such as the Standard Gamble, the Person Trade-off and the Visual Analogue Scale (4), as well as the newly emerging Discrete Choice experiments for the valuation of health states(152, 153). While an in depth discussion of the relative advantages and disadvantages of

these instruments is outside the scope of this thesis, it is important to acknowledge that this thesis only focused on the improvement of the TTO, as it is one of the most applied preference-elicitation methods, while other methods for the valuation of health are indeed available and command similar attention.

Future avenues of research regarding RQ 2

In the context of RQ 2 new TTO procedures were explored for the uniform measurement of health states better than dead and health states worse than dead. However, several issues remain unresolved and require further research.

First of all, new studies with lead time TTO and lag time TTO are encouraged, but these would preferably be conducted in a face-to-face setting (i.e. one interviewer and respondent), rather than in an online setting. Until more TTO studies with better quality have been conducted, it is not possible to come to a well-founded conclusion regarding the relative advantages of lead time TTO and lag time TTO over conventional TTO. Similarly, the study that formed the basis for chapters 3 and 4 used a between-subject design. A within-subject design also has disadvantages, such as a learning effect in respondents throughout the experiment, but this bias can (partly) be neutralized through randomization. Generally speaking, a within subject design would strengthen conclusions and may therefore be considered the preferred design for future TTO studies.

The thorny issue of 'worse than dead' health states is likely to be somewhat more complicated than can be solved by the provision of a uniform scale of measurement for health states worse and better than dead, or through a combination of the conventional TTO and new approaches[8]. Lead time TTO and lag time TTO still operate within the assumption of the traditional Quality Adjusted Life Year model, which assumes, amongst other things, constant proportional tradeoffs (CPTO). This means that "TTO scores should be the same no matter what time horizon is used in the elicitation process" (37). The CPTO assumption has often been rejected(37, 154), and indeed the assumption is unlikely to hold, especially for health states considered worse than dead. Many studies may have shown that individuals can prefer immediate death over having to live in a very poor health state for 10 years (and hence stating that the health state is equal to being dead or even worse than being dead) such as the EQ-5D valuation studies in the United Kingdom and in The Netherlands (25, 75, 86). But would these individuals also prefer

[8] The new EQ5D-5L, which will probably become the most often used preference-base measure for QALY computation, is currently being valued using a combination of 'common' TTO for health states better than dead and lead time TTO (10 years full health + 10 years disease time) for health states considered worse than dead, referred to as 'composite TTO'. Preliminary research results seem to indicate that the approach is feasible (162), but head to head comparisons of composite TTO with other specifications of TTO are currently not yet available.

to immediate death over living in that very poor health state for two days, which, for instance, provides the opportunity to say a last farewell to loved ones? The lead time TTO and the lag time TTO offer a good opportunity to research this issue of CPTO in relation to very poor health states further.

Concerning the impact of time preference on health state valuations, the application of differential discount functions for good health and impaired health is an important avenue of research. In chapter 4, it was assumed that individuals apply the same discount rate to years in good health as to years in poor health. However, it may equally well be that respondents perceive impaired health states differently from good health states, and hence assign a different weight to the timing of poor health relative to that of good health. The properties of lead time TTO and lag time TTO allow such further study of discounting, since both years in good health and in poor health are included in one hypothetical life. Further studies into time preference are important, in the context of economic evaluations, but also because it may, for instance, lead to a better understanding of healthy behavior. This would, however, require cooperation between researchers on time preferences and researchers from other fields, such as the area of health prevention. Often, healthy behavior has (uncertain) health benefits in the far future (it may, for example, reduce the chance of heart disease at older age), while it may have negative consequences in the present (such as eating less, or less tasty food). If respondents discount unhealthy life years to a larger extent than healthy life years, this could explain unhealthy behavior. Understanding healthy and unhealthy choices better, can for example help governments better design and target health prevention programs.

8.3 RQ 3) Are generic preference-based measures preferable to condition-specific preference-based measures?

Placing a value on a health state with a preference elicitation task such as TTO "requires some means of first describing [the health state]" (p. 55) (4). Most commonly, health states are described using a standardized questionnaire. On such a questionnaire, which is often self-administered, a patient answers questions (referred to as items) about several dimensions of health which, when combined, indicate the 'health state' of the patient. Typically, the content of this questionnaire reflects the dimensions of health or HRQoL.

The items of a questionnaire and the answer categories have to be sufficiently 'generic' to be applicable to patients with a very wide range of diseases. However, if the items are 'too generic', the instrument may lose sensitivity to disease specific dimensions of health related quality of life. Developing a questionnaire thus generally requires a difficult trade-off between disease specific sensitivity and a sufficient degree of generality of the questionnaire. Common practice is to use instruments which are 'generic' in nature.

The EQ-5D, one of the most popular preference-based questionnaires, measure 5 broad dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. A concern of patients, decision makers and doctors alike, is whether these 'generic' descriptions of health, based on the existing questionnaires such as the Health Utilities Index (155) and the EQ-5D (131), are sufficiently sensitive to the effects of specific medical interventions[9].

To answer research question 3, the study described in chapter 5 was conducted. For this study, three new condition-specific measures were developed and compared to EQ-5D performance in three patient populations: cancer, multiple sclerosis and arthritis. The results showed that two out of three instruments were better at identifying patients that had mild complaints than EQ-5D. Also, the instrument developed for multiple sclerosis showed better discriminative properties than EQ-5D. However, all instruments showed a 'floor-effect' when compared to EQ-5D, which means that the instruments were less capable in discriminating patients in poor health from those in very poor health. Mean values of the new instruments indicated a different mean quality adjustment than that observed using EQ-5D.

It seems that there are both relative advantages and disadvantages of using condition-specific preference-based instruments, compared to EQ-5D. Therefore, it is advised that condition-specific measures are only used for the quality adjustment of the QALY when there is evidence it has better measurement properties in a given patient population than the existing generic instruments.

There were several limitations to the approach used to answer RQ 3. First, it was decided to base the condition-specific instruments on existing questionnaires. This methodology has the merit of working with renowned and empirically supported questionnaires. On the other hand, a condition-specific measure could also have been developed *de novo*. Although this is much more work, as the new measure requires extensive empirical testing, it does provide more freedom in choosing relevant dimensions of health. Second, the selection of items from existing questionnaires was based on statistical properties and expert opinion of three researchers, each very experienced with studying the disease. In hindsight, it could have improved the validity of the set of selected items if a panel of patients (or the general public) had been involved to indicate which items they consider to be the most important. The third limitation of the study was that the TTO study presented in chapter 5 was performed in a supervised group setting using a highly standardized interview protocol on the computer. Although TTO in a group setting (i.e. more respondents at work with 2 supervisors present) has been shown to produce

[9] It should be noted that this concern may be sensible, but that any statement about the perceived (in) sensitivity of generic preference-based questionnaires requires empirical support.

good results, as was the case in a previous study (120), the face-to-face TTO interview remains preferable, since then it is easier to see if all respondents properly understand the TTO task.

Future avenues of research regarding RQ 3

There are different lines of future research possible in this context. Future research, could, for example, address the limitations of the study in chapter 5. First of all, it would be interesting to see whether patient and general public samples would select the same, or at least a similar, set of items for the condition specific instruments. Second, condition-specific preference-based instruments could be developed *de novo*, and it would be very interesting to see if the dimensions of health of such an instrument would differ much from existing measures (like the ones used in chapter 5). Third, the questionnaires developed and presented in chapter 5 could be assigned new preference values in a study using face-to-face TTOs rather than the group TTOs that were used here.

Another line of research would be to further analyze some of the interesting results found in chapter 5. There, it was shown that it matters which health states are included in a preference elicitation task. These health states are often selected from the total set of possible health states on statistical grounds, meaning that they all differ from each other in such a way that together they are a good representation of all possible health states. The preference values for the remaining health states are then estimated with statistical inference. However, the thus selected health states are not necessarily those that are most often experienced by patients. Chapter 5 showed that including the latter health states in the study design results in different values for these important health states. In chapter 7 it was shown that only 7 EQ-5D health states account for more than 60 percent of all observed health states. Since the EQ-5D is the most used generic preference-based measure, it is important to repeat the exercise of chapter 5 and include the most observed health states of EQ-5D in valuation studies.

Another future line of research could focus on the 'face validity' of generic preference-based measures. Generic measures are often *thought to be* insensitive to disease specific improvements although it is essentially an empirical matter whether this is, in fact, true. The perceived insensitivity may be one of the driving forces behind the proliferation of condition-specific preference-based instruments in the scientific literature. The concept of 'face validity' of an instrument suggests that it should not only *be* valid (i.e. that it measures what it is supposed to measure), "it should also *appear* valid" (italics in original) (156). Clearly, many elements concerning the validity of questionnaires are empirical matters, in the sense that it has to be tested if an instrument measures what it is supposed to measure. Nevertheless, *appearing valid* serves an important purpose. Although face validity may not reflect content validity of an instrument, it is a very important aspect of questionnaires since it may be "more likely that

policy makers and others accept the results" (Nevo, quoted in Norman and Streiner p. 83) (157). These 'others' may be relevant interest groups such as patients and doctors. These are key stakeholders who also may need to implement policy decisions based on CUA, such as prescribing only those treatments with favorable CUA results. The face validity of instruments may be more important than is generally assumed as it may ultimately impact on the support CUA results have in the medical community. Therefore, improving the face validity of existing generic measures may be an important challenge for future research.

8.4 RQ 4) Is it possible to estimate health state values using ordinary least squares regression?

Despite the considerable effort to develop high quality preference-based instruments and the increasing influence of CUA in policy decisions, preference-based instruments are not always used in clinical research. Obviously, without information on the benefit of treatments in terms of QALYs, health economists cannot perform CUAs to inform decision makers on the relative 'value for money' of interventions. A method that aims to repair this issue ex post is 'mapping' which uses regression techniques to *estimate* QALY gains based on other known characteristics of patients. Mapping therefore does not improve the measurement and valuation of HRQoL for QALY computation in any direct sense. Rather, it aims to overcome the practical problem of unavailability of health related quality of life data, often faced by health economists.

Mapping is feasible because health status information commonly measured in clinical trials often correlates with health state preference values. In other words, when a patient is in poor health, this is reflected in both clinical indicators and in quality of life. In chapter 6, three data sets in which both EQ-5D data and other information on patient health was present were used to establish the statistical relation between EQ-5D health state values and other information on patient health. This statistical relation (the mapping function) was used to estimate the mean EQ-5D based preference values in cancer, multiple sclerosis and arthritis samples. The mean estimated value did not differ from the observed value. However, in all instances, the quality of life of patients in poor health was overestimated with the mapping function when compared to observed EQ-5D values. The overprediction at the lower end of the distribution is offset by underpredictions at the higher end of the distribution and as a result, the mean predicted value is correct. The differences between observed values and predicted values at the higher end of the distribution are smaller in size, but greater in number.

This issue of 'overprediction' of was further explored in chapter 7. Several causes for overprediction were discussed there. First, the EQ-5D preference-based values for patients who report very poor health are very low, also relative to all other health state values. Second,

there are very few patients who report extreme health problems. As a consequence, patient QoL data measured with EQ-5D often has a bi-modal distribution, with the majority of the data representing patients in relatively good health, and this distribution is difficult to estimate with ordinary least squares regression. In chapter 7 it was shown that a stepped approach, with separate mapping functions for patients who are in poor health can partly overcome the issue of overprediction.

Estimating EQ-5D values is only a second best alternative to collecting EQ-5D data directly, and is only valid when EQ-5D itself is valid to measure HRQoL for the specific condition in which it is being applied. When EQ-5D is not valid, condition-specific measures such as those described in chapter 5, are to be preferred over both EQ-5D or estimated EQ-5D values.

The studies in chapter 6 and 7 used the same datasets and therefore share limitations that are related to the data. First of all, the data sets used for the development of the mapping functions contained particular patients. For cancer, these were multiple myeloma patients, for multiple sclerosis, these were patients that had either the relapsing remitting or the secondary progressive form and for arthritis, patients with early arthritis were included in the database. Since patients with other subtypes of these diseases may have other symptoms, it is unlikely that the developed mapping functions can also be applied in patient samples with other subtypes of the same disease. The mapping function developed in a multiple myeloma sample was also able to estimate mean EQ-5D values in a non-Hodgkin sample. It was concluded that the mapping function may be used in both these types of cancer. However, generalizability to other types of cancer has not been tested and should therefore only be attempted in an experimental setting, prior to use in QALY computation for CUA.

A second limitation is that only ordinary least square regression was applied, while several other regression functions could have been estimated. For example, ordered logistic regression could have been used to estimate answer categories on EQ-5D. Beta-regression or other specifications could have been applied to estimate EQ-5D health state values.

A third limitation is that all the datasets were relatively small in size. On larger datasets, it is more likely that statistically significant relations between EQ-5D and other sources of health information are found, and this can improve the predictive ability of mapping functions.

A fourth limitation is that the EQ-5D health state values were based on the Dutch tariff, and as such, the mapping functions developed here are only applicable to Dutch patients.

Future avenues of research regarding RQ 4

There are several opportunities for further research into the estimation of health state values. First, the external validity of mapping functions command attention. This means that mapping functions published in the literature have to be tested on other data sets. Unfortunately, most researchers present their own new mapping functions, rather than testing existing ones.

Second, the use of more advanced modeling techniques such as beta-regression may overcome some of the issues of mapping observed when using ordinary least squares regression, such as the overprediction of low utility values. These are, however, not necessarily an improvement in terms of predictive ability of mean preference values (158). Nevertheless, new regression techniques are the most viable route to improving estimated preference values.

Third, large data sets that contain many patients with different subtypes of a disease could be combined. Such a combined dataset could be used to estimate mapping functions with a dummy variable that specifies the subtype of illness from which a patient suffers. If the dataset is sufficiently large, it may be possible to use ordered logistic regression to estimate the answers of patients on the EQ-5D questionnaire, rather than estimate health state values with ordinary least squares regression. The advantage of such a strategy is that different 'tariffs' of different countries can then be attached to the estimated EQ-5D answers.

8.5 GENERAL CONCLUSION

Economic evaluations of medical interventions aim to aid welfare improving decisions, but to achieve this aim, economic evaluations require a valid operationalization of its constituting elements. As elaborated on in the discussion, CUA is a specific type of economic evaluation which is applied to health care interventions. In CUA, decisions are generally deemed welfare improving if the costs per QALY are lower than what a society is willing to pay for a (particular) QALY. The ratio of cost per QALY, naturally, depends on how much QALYs are gained, and this, in its turn, is influenced by a variety of theoretical and methodological choices made in the QALY computations. If these influences lead to biased or incorrect CUA results, then CUA is not a valid aid to support welfare improving decisions.

In this thesis it was investigated whether the measurement and valuation of health for QALY computation in economic evaluations can be improved. It seems that there indeed is room for improvement in several areas. First, despite a strong practical consensus to do otherwise, including patient preferences for health states in QALY computation can improve the measurement of benefit, if only by recognizing the alternative information patient preferences represent, such as the quality that life can have despite impaired health. Second, regarding preference elicitation

methods, new TTO methods have been developed but evidence is too scarce and inconclusive to advocate adopting lag time TTO or lead time TTO for all valuation studies. Third, regarding health state descriptive systems, condition-specific preference-based measures in certain instances can be an improvement to existing generic measures. Finally, if neither generic nor condition specific preference-based instruments were included in a study, mapping techniques can be used to estimate preference-based values to allow QALY computation.

That the measurement and valuation of health can be improved indicates that new scientific insights, that were not available in the past, need to find their way to CUA methodology. If researchers wish decision makers to use CUA results, it is of utmost importance that CUAs are methodologically valid and (as much as possible) unbiased, since decisions based on the results of CUAs may affect the welfare of many individuals. It is exactly for this reason that the continuous reflection on and improvement of the methodology of the constituent elements of CUA methodologies commands attention, now and in the future.

8.6 FINAL REMARKS

Measuring quality of life for economic evaluations of health care interventions is a challenging exercise. Research in this field integrates some of the most fascinating academic disciplines including economics, psychology and philosophy. Although the field is developing into a discipline in its own right, it has only existed for about 40 years (in its application to economic evaluations) and is therefore swiftly changing with many avenues of research that remain relatively unexplored. Sometimes, however, it seems that there is more attention for the latest methodological challenge, than for the philosophical foundations of the field, and this thesis is no exception.

In this thesis, following the tradition in health economics, the value of a health state was assumed to be equated to the preferences for – or desirability of – that health state. All research questions operated within this assumption. RQ 1 asked *who's preferences* should determine the value of health states, and hence it is not discussed whether 'preferences' for health states are indeed the best proxy for the quality of life in health states. Similarly, RQ 2, asked for an improvement in preference-elicitation techniques, rather than alternatives to preferences. For RQ 3 new preference-based measures were developed, and RQ 4 asked how to best estimate preference-based health state values when they have not been measured. While results in chapter 5 showed that preference-based measures are indeed capable of measuring improvements in quality of life, it is important to acknowledge that alternatives exist.

Indeed, there has been a lively debate on the limit to the use of preferences as an indication of quality of life. In the capability approach, as advocated by Amartya Sen, it is argued that there is an important distinction to be made between the *capabilities* a certain individual has, and the *feelings or preferences* an individual may have about his or her capabilities (159, 160). The potential of the 'capability approach' for health economics was endorsed by Coast and colleagues in 2008 (161). The field of health economics could be greatly enriched with future interdisciplinary research that does not restrict itself to the preference-based framework or to the utilitarian framework of maximization.

All of the work performed here owes gratitude to the existing body of knowledge in the health economics discipline. In all modesty, this thesis hopes to form a stepping stone towards a better understanding of the measurement of quality of life for economic evaluations of health.

9

Samenvatting

De stijging van de zorguitgaven in Nederland wordt voor een deel veroorzaakt door nieuwe medisch- technologische interventies. Deze interventies kunnen een belangrijke bijdrage leveren aan de gezondheid en het welzijn van de Nederlandse bevolking. Met economische evaluaties wordt onderzocht hoe de kosten van een interventie zich verhouden tot de gezondheidswinst. Die informatie wordt gebruikt om de Nederlandse overheid te adviseren over de samenstelling van het basispakket van de zorgverzekering.

Een belangrijke vorm van economische evaluaties van medische interventies is de kostenutiliteitsanalyse (KUA). In KUAs wordt de gezondheidswinst van interventies uitgedrukt in het aantal gewonnen Quality Adjusted Life Years (QALYs). De 'quality adjustment' van de QALY is een getal dat de kwaliteit van leven uitdrukt van de gezondheidstoestand van patiënten. In dit proefschrift werd onderzocht hoe kwaliteit van leven het best gekwantificeerd kan worden voor gebruik in KUA.

Het meten van een veelzijdig concept als kwaliteit van leven is op zichzelf al vrij complex. Kwaliteit van leven kwantificeren voor gebruik in KUAs van medische interventies is nog net iets complexer door de gevolgen die de meting kan hebben voor vergoedingsbesluiten. Een voorbeeld: een patiënt met ernstige beperkingen die, dankzij een moeizaam proces van adaptatie, revalidatie en een bewonderingswaardige menselijke veerkracht, een zinvol en bevredigend leven weet te leiden, geeft zichzelf een hoog cijfer voor kwaliteit van leven. Een behandeling die deze patiënt volledig geneest kan dat cijfer niet heel veel hoger maken, omdat het al zo hoog was. De patiënt heeft in die situatie wel baat bij de aangeboden behandeling, maar die baten vallen in de meting laag uit, omdat de patiënt zichzelf ook *voor* de behandeling al een hoog kwaliteit van leven cijfer had gegeven. De ratio tussen kosten en gewonnen QALYs van een behandeling zou dan onnodig hoog uitvallen met als gevolg een mogelijk negatief advies voor opname in het basispakket van de zorgverzekering. In KUAs wordt getracht het bovenstaande probleem van 'adaptatie aan slechte gezondheid' op te lossen door de kwaliteit van leven die hoort bij bepaalde gezondheidstoestanden niet vast te laten stellen door patiënten zelf, maar door een representatieve steekproef van het Nederlands algemeen publiek. Zij hebben, gemiddeld genomen, immers niet het proces van adaptatie ervaren en zullen daarom niet zo snel hoge waarden toekennen aan gezondheidstoestanden die slechter zijn dan die van henzelf.

In **hoofdstuk 2** van dit proefschrift is betoogd dat de kwaliteit van leven die patiënten toekennen aan gezondheidstoestanden *ook* van belang zijn voor economische evaluaties, ondanks, of misschien wel dankzij, adaptatie. De kostenutiliteitsratio van een levensverlengende behandeling kan namelijk ook onnodig hoog uitvallen, doordat het algemeen publiek niet voldoende ervaring met ziekte heeft om in te zien dat het leven voor de patiënt zelf, ondanks bepaalde beperkingen, nog uiterst waardevol is. Het algemeen publiek zou de kwaliteit van leven van de patiënt in door

behandeling gewonnen levensjaren te laag waarderen, omdat er geen rekening wordt gehouden met de nieuwe levensdoelen en zingeving die een patiënt ervaart ondanks ziekte en/of handicap. In hoofdstuk 2 wordt dan ook betoogd om gezondheidstoestanden te waarderen vanuit zowel het perspectief van de patiënt als het perspectief van het algemeen publiek. Op die wijze zou de veelvormigheid en de veranderlijkheid van de betekenis van gezondheid en kwaliteit van leven voor mensen, in voor en tegenspoed, het meest volledig in kaart worden gebracht, zonder dat patiënten in economische evaluaties worden benadeeld voor een bewonderingswaardige aanpassing aan hun omstandigheden.

De kwaliteit van leven in een bepaalde gezondheidstoestand wordt vastgesteld door een groep individuen, bijvoorbeeld patiënten of Nederlands algemeen publiek, gezondheidstoestanden te laten rangschikken op 'wenselijkheid'. In **hoofdstuk 3** van dit proefschrift zijn twee onderzoeken beschreven naar nieuwe onderzoeksmethoden om een 'kwaliteit van leven getal', ook wel een 'utiliteit' genoemd, aan een gezondheidstoestand toe te kennen. Deze methoden bouwen voort op een bestaande keuzetaak, waarin respondenten wordt gevraagd te kiezen tussen twee hypothetische levens: een korter gezond leven en een langer leven in een slechtere gezondheidstoestand. Als de respondenten kiezen voor een (veel) korter leven geeft dit aan dat ze het alternatief, de slechtere gezondheidstoestand, zo onaantrekkelijk vinden dat ze liever korter in goede gezondheid leven. Deze methode, die de Time Trade off (TTO) heet, wordt al sinds de jaren '80 gebruikt om een kwaliteit van leven getal toe te kennen aan een gezondheidstoestand. Recentelijk is er echter commentaar op een deel van die methode gekomen. Voornamelijk op de manier waarop een kwaliteit van leven getal werd toegekend aan gezondheidstoestanden die zo slecht bevonden worden dat respondenten aangaven liever helemaal niet te leven, dan een langere periode in de zeer slechte gezondheidstoestand te moeten leven. **In hoofdstuk 3** is aangetoond dat twee nieuwe varianten van de TTO methode, de lead time TTO en de lag time TTO, in staat zijn om zowel goede als hele slechte gezondheidstoestanden op een uniforme wijze in kaart te brengen. Echter, deze methoden zijn nog onvoldoende vaak getoetst om te kunnen concluderen dat ze de standaard moeten vormen. Een specifiek probleem dat werd besproken in **hoofdstuk 4** is dat respondenten een (hypothetisch) gezondheidsprobleem minder erg vinden als het pas over een aantal jaren optreedt. In de lead time TTO wordt een gezondheidsprobleem omschreven alsof het pas over een aantal jaar optreedt en als gevolg daarvan worden er hogere kwaliteit van leven getallen gevonden in lead time TTO dan in andere TTO varianten voor identieke gezondheidstoestanden.

De omschrijving van gezondheidstoestanden in een TTO taak is veelal gebaseerd op vragenlijsten over gezondheid. De meest gangbare omschrijving is gebaseerd op vragenlijsten die vrij algemene (generieke) omschrijvingen van dimensies van gezondheid hanteren zoals mobiliteit, vermoeidheid en pijn. Het voordeel van dat soort vragenlijsten is dat ze toepasbaar zijn bij vrijwel

alle aandoeningen en zodoende voor alle aandoeningen vergelijkbare gezondheidstoestanden omschrijven. Een nadeel dat sinds kort veel aandacht in de internationale literatuur krijgt, is dat deze vragenlijsten mogelijk ongevoelig zijn voor gezondheidswinsten die specifiek zijn voor bepaalde aandoeningen, zoals misselijkheid bij bepaalde vormen van chemotherapie. In dit proefschrift zijn drie ziekte-specifieke instrumenten ontwikkeld die zijn beschreven in **hoofdstuk 5**. Aan de gezondheidstoestanden die gebaseerd zijn op deze ziekte-specifieke instrumenten zijn kwaliteit van leven getallen toegekend door 402 respondenten uit het Nederlands algemeen publiek op basis van de TTO methode. Voor twee van de drie nieuwe instrumenten geldt dat ze gevoeliger zijn voor milde gezondheidsachteruitgang dan het veelgebruikte generieke instrument 'EQ-5D'. Ze zijn zodoende een verbetering voor het meten van de 'quality adjustment' van de QALY. Het blijkt echter wel dat deze nieuwe instrumenten gemiddeld hogere getallen toekennen aan de kwaliteit van leven van de gezondheidstoestand van patiënten dan de bestaande generieke instrumenten. Daarom is het onverstandig nieuwe ziekte-specifieke instrumenten te gebruiken, tenzij er empirisch gefundeerde aanwijzingen zijn dat bestaande generieke instrumenten niet volstaan.

De methodologische vernieuwing uit de hoofstukken 3 tot en met 5 vormen een belangrijke verbetering in het vaststellen van de gezondheidswinst van medische interventies. Maar aangezien kosten-utiliteits analyses van medische interventies nog een relatief jonge wetenschap is, is de meting van gezondheidswinst in termen van QALYs nog niet zo wijdverspreid. In de praktijk hebben gezondheidseconomen dus vaak niet de juiste gegevens om QALYs uit te rekenen. In **hoofdstuk 6 en 7** is getracht om via statistische methoden de 'quality adjustement' van de QALY te schatten in onderzoeken waar die niet was gemeten. Het blijkt goed mogelijk om op basis van andere gegevens over de gezondheid de gemiddelde 'quality adjustment' voor de kwaliteit van leven van een patiëntenpopulatie te schatten. Echter, de methoden die zijn gehanteerd in dit proefschrift (ordinary least squares regression) zijn alleen geschikt voor het goed schatten van de gemiddelde kwaliteit van leven in een patiëntengroep en niet voor het schatten van de individuele kwaliteit van leven.

In dit proefschrift is getracht de methodologie en theorie van het meten van kwaliteit van leven voor economische evaluaties van medische interventies te verbeteren. De resulterende methoden dragen bij aan een beter zicht op de gezondheidswinst van medische interventies.

10

Summary

The increase in health expenditures is partly caused by new medical technological interventions. These interventions typically contribute to population health and well-being in The Netherlands. Economic evaluations are performed to identify the ratio between the cost of an intervention and the health benefit it yields. Information derived from these economic evaluations is used to advise the Dutch government on the content of the basic benefit package of the Health Care Insurance act.

One important form of economic evaluation is the cost-utility analysis (CUA). In CUA, health benefit is expressed in terms of Quality Adjusted Life Years (QALYs). The 'quality adjustment' part of the QALY is a number, which reflects the quality of life related to a certain health state. Measuring the 'quality adjustment' of the QALY is not without problems. In this thesis, several improvements to measuring the quality adjustment were sought. These improvements apply to four key issues: 1) who should value health; 2) how should health be valued; 3) which aspects of health should be valued and 4) what can be done when no quality adjustment has been measured.

Measuring a multidimensional concept such as 'quality of life' is sufficiently challenging in itself. Measuring quality of life in a way that is applicable in CUA is even more challenging, due to the potential consequences for reimbursement decisions: if a CUA suggests that the costs of a medical treatment do not outweigh the health benefits it yields, it may not be collectively reimbursed.

An example of the complexity of measuring and valuing quality of life for economic evaluations is the following. Patients who, despite impaired health, enjoy a good quality of life, due to admirable human resilience in times of adverse events, pose a problem for economic evaluations. A patient who, due to a difficult and time-consuming process of adaptation and rehabilitation, enjoys a meaningful and satisfying life, may score unexpectedly high when quality of life is quantified. Any treatment, even if it cures the patient completely, cannot increase that quality of life score by much, since it was very high to start with. The ratio of cost to benefits of such a treatment would, thus, be very high, but 'erroneously' so. In CUA this problem is tackled by having a representative sample of the general population determine the quality of life experienced in certain states of health. They have, at least on average, not experienced adaptation to the disease at hand and do not assign 'unlikely high' quality of life values to health states that are worse than their own. This strategy, however, also has its pitfalls. The cost-utility ratio can equally well be erroneously high for life extending treatments, because the general public does not have sufficient experience with disease to see that a patient feels that life is very valuable, despite certain impairments. The general public would then underestimate the benefit in the life years gained due to treatment, because they underestimate the quality of life experienced by

a patient despite illness or disability. In **Chapter 2** of this thesis it is argued that the quality of life patients attach to their own state of health is also of importance for economic evaluations, despite, or even due to the issue of adaptation. Therefore, it is argued in chapter 2 that health state values used in cost-utility analysis should be based on valuations by patients and by the general public. That way, the multiplicity and versatility of the meaning of 'health' and 'quality of life' for individuals, in good times and in bad, is most complete, and it prevents that patients who have gone through an admirable process of adaptation to impaired conditions, would be disadvantaged in economic evaluations.

The quality of life in certain states of health is determined with preference elicitation studies. These studies measure the desirability of certain states of health according to groups of individuals, such as patients or the Dutch general public. **Chapter 3** of this thesis describes two research projects focused on new methodologies for obtaining a 'quality of life value' for health states, also referred to as a 'preference value'. These new methodologies are an extension of existing choice-tasks in which respondents are asked to choose between two alternative and hypothetical 'lives': a shorter healthy life or a longer life in poor health. Respondents who opt for a (much) shorter life in good health indicate that they consider the alternative, the poor health state, very undesirable. This method, called the Time tradeoff (TTO), has been used since the '80 to obtain 'quality of life' values for health states. Recently the TTO method has been criticized, mostly on the methodology for valuing health states so poor they are considered 'worse than death'. In **chapter 3** it is shown that alternative specification of the TTO, the lead time TTO and the lag time TTO, are capable measuring the desirability of health state values, whether very poor or very good in a uniform fashion. However, these methods have not been tested extensively enough be adopted as a complete replacement of old TTO. One specific issue, discussed in **chapter 4**, is that respondents consider (hypothetical) health states less problematic when they occur in the future. In lead time TTO, health impairments start in the future, after a period of good health. As a consequence, identical health states receive higher values in lead time TTO studies than in other specifications of TTO.

Descriptions of a health state used in preference elicitation tasks such as TTO are generally based on questionnaires with generic dimensions of health, such as mobility, fatigue and pain. The main advantage of those questionnaires is that they are applicable to several conditions and, thus, are able to describe health in a uniform fashion for different health conditions. A disadvantage, which has recently gained more attention in the literature, is that these questionnaires may be insensitive to health effects specific to certain conditions, such as nausea during chemotherapy. In this thesis, three new condition-specific measures have been developed, which are described in **chapter 5**. The health states derived from these questionnaires have received 'quality of life values' based on a preference elicitation study in the Dutch general public using TTO. Two out

of three of the questionnaires were more sensitive to mild health impairments than existing generic instruments. These instruments are, thus, an improvement for measuring the quality adjustment part of the QALY. However, these new instruments indicate rather different health state values than existing generic instruments. Therefore, it would be wise to only use new condition-specific measures for the computation of QALYs when there is empirical evidence that existing generic measures are not valid.

The methodological improvements from chapters 2 through 4 are an important contribution to the measurement of benefit in medical interventions. However, since the cost-utility analysis of medical interventions is a relatively young practice, the measurement of benefit in terms of QALYs is not common practice. In practice, health economists do not have the required data to compute QALYs. In **Chapter 6 and 7**, statistical methods are applied to estimate the quality adjustment part of the QALY in data sets where this information was absent. It seems possible to estimate the average 'quality adjustment' of a patient population using information on other aspects of their health. However, the methodological approach that was applied in this thesis (ordinary least squares regression) is only suitable for estimation of the mean quality of life in a patient sample, and not for the estimation of individual values.

In this thesis it was attempted to improve the existing methodology for measuring quality of life for use in economic evaluations of medical interventions. The methods in this thesis contribute to a better measurement of the benefit of medical interventions.

11

Acknowledgements

Het vak gezondheidseconomie is mij bijgebracht door dr. Elly Stolk en prof. dr. Werner Brouwer. Ik ben jullie beide zeer dankbaar voor de inhoudelijke begeleiding, de persoonlijke aandacht, de kansen en de vrijheid die jullie mij geboden hebben. Ik heb me gezien gevoeld.

Door samenwerkingsverbanden en discussies heb ik veel geleerd van mijn collega's. Bijzondere dank gaat uit naar dr. Marc Koopmanschap, dr. Arthur Attema, dr. Pieter van Baal en Prof. Carin Uyl-de Groot. Prof. Uyl de Groot stelde haar kennis en onderzoeksnetwerk belangeloos ter beschikking en daar heb ik veel profijt van gehad in de voortgang van dit onderzoek. Annemieke Leunis en Jolanda Luime waren belangrijke mede-onderzoekers.

Met een aantal collega's heb ik de afgelopen vier jaar een vriendschap opgebouwd die het werk overstijgt. Samen hebben we onze tijd op het iBMG -en ver buiten de landsgrenzen- tot een feestje gemaakt en tot tranen toe gelachen. Ik hoop nog lang van ons 'collectief' te genieten.

At Sheffield University, Donna Rowen, John Brazier and Aki Tsuchiya introduced me to their specific areas of research, which turned out to be a kick-start to this thesis. Thank you for your supervision.

I also owe gratitude to the critical and skilled members of the EuroQoL group. I enjoyed the wit and wisdom of Paul Kind, Nancy Devlin, Jan van Busschbach, Paul Krabbe, Mark Oppe, Frank de Charro, Peep Stalmeijer, Ben van Hout and Kim Rand-Hendriksen at conferences all over the world.

Dank ook aan de medewerkers van het College Voor Zorgverzekeringen die mij veel geleerd hebben over het beheren van het verzekerde pakket in de praktijk.

Ook mijn nieuwe collega's bij Ecorys krijgen mijn dank: Wija, Kim en Marijke, ik ben blij dat ik met jullie in een ijzersterk team terecht ben gekomen na het verlaten van de universiteit.

—

In mijn privéleven gaat mijn dank en liefde uit naar mijn moeder Maria en mijn broer Daniël. Natuurlijk ook naar mijn nichtje Petra, mijn geweldige opa Lau, mijn neven Nicolaï, Marius en Gabriël, mijn oom en tante Helene en Bart, oom Laurens en dierbare familievriend Wienke. Jullie vormen de rijke voedingsbodem van mijn leven, waarin ik wortel en waardoor ik groei.

Mijn vrienden uit Utrecht, alfabetisch Bob, Dirk, Evert, Jasper, Jorik en Teun hebben nagenoeg niets bijgedragen aan de totstandkoming van dit proefschrift. Jullie liggen echter aan de bron van vrijwel alles buiten mijn werk, waarvoor ik dankbaar ben. Het leven met jullie is warm en mateloos grappig. Ik mis jullie als jullie op vakantie zijn.

Hiske en Cathelijn, mijn dierbare vrienden voor het leven; jullie zijn beiden een bron van inspiratie en van plezier. Er zijn weinig mensen die ik zo weinig zie en die tegelijkertijd zo belangrijk voor me zijn.

Bij Freek en Truus vind ik de gouden combinatie van liefde voor autotechniek, zingeving en onwaarschijnlijk lekker avondeten. Maastricht heeft echt bijna alles.

Als echte stadsjongen beweeg ik me in een groep mensen die vaak van samenstelling verandert, maar een vaste kern heeft. Hier wil ik in ieder geval graag bedanken voor de tijd, liefde en aandacht die ze me schenken: Ike, Maartje, Fleur, de vriendinnen van de Boys (Janna, Louise, Laura, Anneke, Nathalie en Paulien), mijn Amsterdamse vrienden Bram, Marc (en Arnout en Michiel, die nog moeten verhuizen), Lysanne, Bennie en mijn buren Ray en Mariëlle.

Een warme groet natuurlijk ook aan mijn oud collega's van het Instituut voor Interdisciplinaire Studies van de UvA, waar de academische vlam dagelijks van zuurstof werd voorzien en het plan om te promoveren gesmeed werd.

Mijn dank gaat ook uit naar mijn oude middelbare school, de Werkplaats Kindergemeenschap in Bilthoven, waar mijn studerende leven op een hele fijne manier begon.

Mijn lieve vriendin Andrea. Samen maken we ons huis tot een thuis en de wereld tot ons speelveld. We dansen en we zoeken en kijken samen onze ogen uit. Ik ben je daar iedere dag dankbaar voor en ik geniet, altijd weer, van de vonkjes van plezier die uit je ogen stralen.

Tot slot wil ik graag mijn vader bedanken voor alles wat hij mij heeft geleerd in de tijd die hem gegeven was.

12

List of scientific publications

Versteegh, M.M. & Brouwer, W.B.F. The roal road or the middle way: patient and general public preferences for health states. *Submitted*.

Janssen, M.F.B., Oppe, M., **Versteegh, M.M.** & Stolk, E.A. Introducing the composite time trade-off: A test of feasibility and face validity.
European Journal of Health Economics, 2013

Attema, A., Ydelaar-Peeters, Y., **Versteegh, M.M.** & Stolk, E.A. Elements of TTO that affect health state values: a checklist.
European Journal of Health Economics, 2013

Versteegh, M.M., Attema, A.E., Oppe, M., Devlin, N. & Stolk, E.A.
Time to tweak the TTO: Results from a comparison of alternative specifications of the TTO.
European Journal of Health Economics, 2013

Attema, A. & **Versteegh, M.M.**
Would you rather be ill now, or later?
Health Economics, 2012

Versteegh, M.M., Leunis, A., Luime, J.J., Boggild, M., Uyl-de Groot, C.A. & Stolk, E.A. Mapping QLQ-C30, HAQ and MSIS-29 on EQ-5D.
Medical Decision Making, 2012

Crott, R., **Versteegh, M.M.** & Uyl-de Groot, C.A.
An Assessment of the External Validity of Mapping QLQ-C30 to EQ-5D Preferences.
Quality of Life Research, June 2012

Versteegh, M.M., Leunis, A., Uyl-de Groot, C.A. & Stolk, E.A.
Condition specific preference-based measures: benefit or burden?
Value in Health, 2012

Attema, A., **Versteegh, M.M.**, Oppe, M., Brouwer, W.B.F. & Stolk, E.A.
Lead-time TTO: leading to better health state valuations?
Health Economics, 2012.

Koedoot, C., **Versteegh, M.M.** & Yaruss, J.S.
Validity of Dutch OASES-A.
Journal of Fluency Disorders, 2011

Versteegh, M.M., Rowen, D., Brazier, J.E. & Stolk, E.A.
Mapping onto EQ-5D for patients in poor health.
Health and Quality of Life Outcomes, 2010

12.1 List of policy related publications for Matthijs Michaël Versteegh

Versteegh, M.M., Weistra, K., de Groot, S., Redekop, K.,, Davis, P., Rutten, F. & Oortwijn, W. Highly specialised and cost-intensive medical equipment.
Advies aan Europese Commissie, 2014

Versteegh, M.M. & Oortwijn, W. Electronic cigarettes.
Workshop voor Europees Parlement, 2013

Versteegh, M.M., Weistra, K. & Oortwijn, W. Evaluatie transkatheter aortaklepvervanging.
Advies aan College voor Zorgverzekeringen, 2013

De Voldere, I., Durinck, E., Mertens, K., Cardon, C., Maenhout, T., Warmerdam, S., **Versteegh, M.M.** & Canton, E. Survey on access to finance for cultural and creative sectors.
Advies aan Europese Commissie, 2013

Versteegh, M.M., Koopmanschap, M., Mastenbroek, C., Latta, J., de Wit, J., Gaasbeek Janzen, M., Witteveen, S. & Kuipers, M.
Adviezen laten werken.
Advies aan College Voor Zorgverzekeringen, 2012

de Groot, S., Rijnsburger, R., Redekop, K., **Versteegh, M.M.**, Heymans, J., Link, A., Verstijnen, I., Boksteijn, B., Kleijnen, S. & Staal, P.
Effectiviteit en Evidence Based Medicine.
Advies aan College Voor Zorgverzekeringen, 2012

Ligtenberg, G., Zwaap, J., Enzing, J., Saase van, L., Stolk, E.A. & **Versteegh, M.M.**
Haalbaarheidstoets uitsluiten DBC's voor aandoeningen met een lage ziektelast.
Advies aan de werkgroep 'Heroverweging Curatieve Zorg', 2010

Commissie Veerman
Bouwen aan verbinding.
Advies over Noord-Zuidlijn aan de Gemeente Amsterdam, 2009

13

Curriculum Vitae

Matthijs Michaël Versteegh (1984, Utrecht, The Netherlands) holds a B.Sc. in Health Sciences (2002-2005) and an M.A. in Arts & Culture (cum laude, 2005-2006), both from Maastricht University. Upon graduation Matthijs worked as a junior teacher at the Institute for Interdisciplinary Studies of the University of Amsterdam (2007-2009). During this period, he interned with the Commissie Veerman, which advised the city of Amsterdam on the future progress of the Noord/Zuidlijn, a subway under the city of Amsterdam.

From 2009 to 2012 Matthijs worked on his PhD on quality of life in economic evaluations of health at the Erasmus University of Rotterdam. During this period he was a visiting PhD-student at Sheffield University in the UK and worked on two advisory projects for the Dutch Board of Health Insurance (CVZ), which focused on feasibility and evidence based medicine in reimbursement decisions. Matthijs is a member of the EuroQoL group, an international research group for the measurement and valuation of quality of life. He is part of the core research group for the valuation of the new Dutch EQ5D-5L, the most important quality of life instrument used in economic evaluations of health.

On October 1st, 2012, Matthijs started working for the health unit of Ecorys research and consultancy. In this position he provides health policy advice to the European Commission, the World Health Organization and the Dutch government.

Matthijs lives in Amsterdam and continues to work on scientific publications with former colleagues in his spare time.

14

PhD Portfolio

Name PhD candidate:	Matthijs Versteegh
Eramus University of Rotterdam department:	iBMG
PhD period:	2009-2012
Promotor:	Prof.dr. Werner Brouwer
Supervisor:	Dr. Elly Stolk

1. PhD training

2009	Health Technology Assessment, Radboud University
2009	Regression analyses, NIHES summer school
2009	Rasch analysis, Leeds University
2011	Discrete choice analysis, Erasmus University of Rotterdam

2. Seminars and workshops

2010	iMTA lunch lecture
2012	iMTA lunch lecture
2012	iBMG lunch lecture

3. Non-scientific presentations

2010	ZonMW: preliminary project results
2010-2011	CVZ: case studies on feasibility issues in reimbursement
2011	CVZ: Severity of illness
2011	Erasmuc MC: mapping EQ-5D in arthritis patients
2011	CG-raad & NPCF: government cuts & severity of illness

4. International conferences

2010	ECHE conference, Helsinki. Oral presentation
2010	EuroQoL plenary, Athens. Poster presentation
2011	iHEA conference, Toronto. Oral presentation
2011	EuroQoL plenary, Oxford. Poster presentation
2011	ISPOR conference, Madrid. Poster presentation
2013	EuroQol plenary, Montreal. Oral presentation

5. National conferences

2010	LoLa HESG conference, Egmond aan zee. Paper presentation
2011	LoLa HESG conference, Soest. Paper presentation

6. Teaching

2009-2011	Sociology of medical science (tutor)
2010-2011	HEPL master thesis contact
2009-2010	Supervisor and co-evaluator for bachelor and master theses
2011	Startweek (tutor)

15

References

1. Poos MJJC, Smit JM, Groen J, Kommer GJ, Slobbe LCJ. Kosten van ziekten in nederland 2005. Bilthoven: RIVM; 2008.
2. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford: Oxford University Press; 2005.
3. Gyrd-Hansen D. Willingness to pay for a QALY. Health Econ. 2003 Dec;12(12):1049-60.
4. Brazier J, Ratcliffe J, Salomon JA, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. New York: Oxford University press; 2007.
5. Bobinac A, van Exel NJ, Rutten FF, Brouwer WB. Valuing qaly gains by applying a societal perspective. Health Econ. 2012 Oct 19.
6. Torrance GW. Utility approach to measuring health-related quality of life. J Chronic Dis. 1987;40(6):593-603.
7. Brouwer WBF, Koopmanschap MA. On the economic foundations of CEA. ladies and gentlemen, take your positions! J Health Econ. 2000;19:439-59.
8. Meltzer D, Johannesson M. Inconsistencies in the "societal perspective" on costs of the panel on cost-effectiveness in health and medicine. Med Decis Making. 1999;19:371-77.
9. Koopmanschap MA, Rutten FFH, van Ineveld BM, van Roijen L. The friction cost method for measuring indirect costs of disease. J Health Econ. 1995 6;14(2):171-89.
10. Briggs A. Handling uncertainty in economic evaluation. Br Med J. 1999;319:120.
11. Claxton K, Paulden M, Gravelle H, Brouwer W, Culyer AJ. Discounting and decision making in the economic evaluation of health-care technologies. Health Econ. 2011 Jan;20(1):2-15.
12. Weinstein MMC, Torrance G, McGuire A. QALYs: The basics. Value in health. 2009;12(s1 moving the qaly forward: building a pragmatic road):S5-9.
13. Boadway RF, Niel B. Welfare economics. Wiley; 1991.
14. Dolan P. The measurement of health-related quality of life. In: Culyer AJ, Newhouse JP, editors. Handbook of Health Economics. The Netherlands: North-Holland; 2000. p. 1723.
15. Ruger JP. Health and social justice. OUP Oxford; 2009.
16. Nord E. The trade-off between severity of illness and treatment effect in cost-value analysis of health care. Health Policy. 1993;24:227-38.
17. Dolan P. Utilitarianism and the measurement and aggregation of quality – adjusted life years. Health Care Analysis. 2001;9(1):65-76.
18. Dolan P, Robinson A. The measurement of preferences over the distribution of benefits: The importance of the reference points. Euorpean Economic Review. 2001;45:1697-709.
19. Dolan P, Olsen JA. Equity in health: The importance of different health streams. J Health Econ. 2001;20:823-34.
20. Wagstaff A. QALYs and the equity-efficiency trade-off. J Health Econ. 1991;10:21-41.
21. Stolk EA, Van Donselaar G, Brouwer WBF, Busschbach JJV. Reconciliation of economic concerns and health policy: Illustration of an equity adjustment procedure using proportional shortfall. Pharmacoeconomics. 2004;22(17):1097-107.
22. van de Wetering EJ, Stolk EA, van Exel NJ, Brouwer WB. Balancing equity and efficiency in the dutch basic benefits package using the principle of proportional shortfall. Eur J Health Econ. 2013 Feb;14(1):107-15.
23. CVZ. Pakketbeheer in de praktijk 2. 2009.
24. National Institute for Health and Clinical Excellence. Appraising life-extending, end of life treatments. 2009.
25. Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: Results from a general population study. Health Econ. 1996;5:141-54.
26. Dolan P. Modeling valuations for the EuroQol health states. Med Care. 1997;35:1095-108.
27. Tilling C, Devlin N, Tsuchiya A, Buckingham K. Protocols for time tradeoff valuations of health states worse than dead: A literature review. Med Decis Making. 2010 Sep-Oct;30(5):610-9.
28. Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: A methodological review of the literature. Health Econ. 2005 Feb;14(2):197-208.

29. Szende A, Oppe M, Devlin N, editors. EQ-5D value sets. inventory, comparative review and user guide. ; 2007.
30. Brazier J, Roberts J, Deverill M. The estimation of a preference based measure of health from the SF-36. J Health Econ. 2002;21:271-92.
31. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002 Feb;40(2):113-28.
32. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: What happens to cross programme comparability? Health Econ. 2010 -02;19(2):125-9.
33. Fryback DG, Lawrence WF. Dollars may not buy as many QALYs as we think. A problem with defining quality of lfie adjustments. Medical Decision Making. 1997;17:276-84.
34. Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: Adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. Med Decis Making. 1994 Jan-Mar;14(1):82-90.
35. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs. J Health Econ. 1997a;16:155-75.
36. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972;7(2):118-33.
37. Attema AE, Brouwer WB. Constantly proving the opposite? A test of CPTO using a broad time horizon and correcting for discounting. Qual Life Res. 2012 Feb;21(1):25-34.
38. van Nooten FE, Koolman X, Brouwer WB. The influence of subjective life expectancy on health state valuations using a 10 year TTO. Health Econ. 2009 May;18(5):549-58.
39. Robinson A, Spencer A. Exploring challenges to TTO utilities: Valuing states worse than dead. Health Econ. 2006 Apr;15(4):393-402.
40. Arnesen T, Trommald M. Are QALYs based on time trade-off comparable?--A systematic review of TTO methodologies. Health Econ. 2005 Jan;14(1):39-53.
41. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002 Jul;11(5):447-56.
42. Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: Feasibility study of the 'lead time' approach. Health Econ. 2011 Mar;20(3):348-61.
43. Attema AE, Versteegh MM, Oppe M, Brouwer WBF, Stolk EA. Lead time TTO, leading to better health state valuations? Health Econ. 2012:n/a,n/a.
44. Dolan P, Lee H, King D, Metcalfe R. Valuing health directly. BMJ. 2009(339).
45. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2010 Apr;11(2):215-25.
46. NICE. Guide to the methods of technology appraisal. London: NHS. National Institute for Health and Clinical Excellence; 2008.
47. College voor zorgverzekeringen. Richtlijnen voor farmaco-economisch onderzoek, geactualiseerde versie. 2006.
48. Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.
49. Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: A meta-analytical comparison of patient and population health state utilities. Value in Health. 2010;13(2):306-9.
50. Arnold D, Girling A, Stevens A, Lilford R. Comparison of direct and indirect methods of estimating health state utilities for resource allocation: Review and empirical analysis. BMJ. 2009 January 01;339.
51. Zethraeus N, Johannesson M. A comparison of patient and social tariff values derived from the time trade-off method. Health Econ. 1999;8:541-5.
52. Tengs TO, Wallace A. One thousand health-related quality-of-life estimates. Med Care. 2000;38(6):583-637.
53. Devlin NJ, Buckingham K, Shah K, Tsuchiya A, Tilling C, Wilkinson G, et al. A comparison of alternative variants of the lead and lag time TTO. OHE Research Paper 10/02 London: Office of Health Economics. In press 2010.

54. Yang Y, Brazier JE, Tsuchiya A, Young TA. Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. Medical Decision Making. March/April 2011 March/April 2011;31(2):281-91.
55. Brazier J, Czoski-Murray C, Roberts J, Brown M, Symonds T, Kelleher C. Estimation of a preference-based index from a condition-specific measure: The king's health questionnaire. Med Decis Making. 2008 Jan-Feb;28(1):113-26.
56. Stolk EA, Busschbach JJV. Converting clinical outcomes into utilities: The valuation of the international index of erectile function (IIEF). In: Stolk E, Busschbach J, editors. The cost-utility of Viagra in The Netherlands (appendix A). Rotterdam: institute for Medical Technology Assessment, Erasmus University Rotterdam; 1999. p. 45-63.
57. Mulhern B, Rowen D, Jacoby A, Marson T, Snape D, Hughes D, et al. The development of a QALY measure for epilepsy: NEWQOL-6D. Epilepsy Behav. 2012 Apr 12.
58. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ. 2009 Jul 8.
59. Nord E, Badia X, Rue M, Sintonen H. Hypothetical valuations of health states versus patients' self-ratings. In: Kind P, Brooks R, Rabin R, editors. EQ-5D concepts and methods: a developmental history. The Netherlands: Springer; 2005. p. 125-138.
60. Stiggelbout AM, De Vogel-Voogt E. Health state utilities: A framework for studying the gap between the imagined and the real. Value in Health. 2008;11(1):76-87.
61. Dolan P, Kahneman D. Interpretations of utility and their implications for the valuation of health. The Economic Journal. 2008;118:215-34.
62. Nord E, Daniels N, Kamlet M. QALYs: Some challenges. Value in Health. 2009;12:S10-5.
63. Dolan P. Thinking about it: Thoughts about health and valuing QALYs. Health Econ. 2010 Oct 22.
64. McTaggart-Cowan H, Tsuchiya A, O'Cathain A, Brazier J. Understanding the effect of disease adaptation information on general population values for hypothetical health states. Soc Sci Med. 2011 Jun;72(11):1904-12.
65. Dolan P, Metcalfe R. Valuing health: A brief report on subjective well-being versus preferences. Med Decis Making. 2012 Feb 2.
66. Sharma R, Stano M, Haas M. Adjusting to changes in health: Implications for cost-effectiveness analysis. J Health Econ. 2004 3;23(2):335-51.
67. Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality of life research: A theoretical model. Soc Sci Med. 1999 6;48(11):1507-15.
68. Krahn M, P PR, Irvine J, Tomlinson G, Bezjak A, Trachtenberg J, et al. Patient and community preferences for outcomes in prostate cancer: Implications for clinical policy. Med Care. 2003;41(1):153-64.
69. Lloyd A, van Hanswijck de Jonge P, Doyle S, Cornes P. Health state utility scores for cancer-related anemia through societal and patient valuations. Value Health. 2008 Dec;11(7):1178-85.
70. Stolk EA, Busschbach JJV. Are patients and the general public like-minded about the effect of erectile dysfunction on quality of life? Urology. 2003;61(4):810-5.
71. Pyne JM, Fortney JC, Tripathi S, Feeny D, Ubel P, Brazier J. How bad is depression? preference score estimates from depressed patients and the general population. Health Serv Res. 2009 Aug;44(4):1406-23.
72. Krabbe PF, Tromp N, Ruers TJ, van Riel PL. Are patients' judgments of health status really different from the general population? Health and quality of life outcomes. 2011;9(1):31.
73. Oldridge N, Furlong W, Perkins A, Feeny D, Torrance GW. Community or patient preferences for cost-effectiveness of cardiac rehabilitation: Does it matter? European Journal of Cardiovascular Prevention & Rehabilitation. 2008 October 01;15(5):608-15.
74. Schackman BR, Goldie SJ, Freedberg KA, Losina E, Brazier J, Weinstein MC. Comparison of health state utilities using community and patient preference weights derived from a survey of patients with HIV/AIDS. Medical Decision Making. 2002 February 01;22(1):27-38.

75. 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):1121-32.
76. Harris R, Olewiler N. The welfare economics of ex post optimality. Economica. 1979 May;46(182):pp. 137-147.
77. Robinson J. Economic philosophy. AldineTransaction; 2006.
78. Kahneman D, Sugden R. Experienced utility as a standard of policy evaluation. Environmental and Resource Economics. 2005;32(1):161-81.
79. Rawls J. A theory of justice. Belknap Press of Harvard University Press; 1999.
80. Brouwer WBF, Van Exel NJA, Koopmanschap MA, Rutten FFH. The valuation of informal care in economic appraisal. A consideration of individual choice and societal costs of time. International Journal of Technology Asessment in Health Care. 1999;15(1):147-60.
81. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Quality of Life Research. 2003;12:599-607.
82. Cohen GA. Equality of what? on welfare, goods and capabilities. In: Nussbaum MC, Sen AK, editors. The Quality of Life. Oxford: Clarendon press; 1993. p. 9.
83. Verkerk M, Busschbach JJV, Karssing ED. Health-related quality of life research and the capability approach of amartya sen. Quality of Life Research. 2001;10(1):49-55.
84. Hadorn DC. The role of public values in setting health care priorities. 781. 1991;32(7):773.
85. Sen A. Development as freedom. Oxford University Press; 1999.
86. Lamers LM. The transformation of utilities for health states worse than death: Consequences for the estimation of EQ-5D value sets. Med Care. 2007 Mar;45(3):238-44.
87. Verschuuren M. Quality adjusted life years and time trade off exercises: Exploring methodology and validity.[dissertation]. Utrecht, The Netherlands: PhD Thesis, University of Utrecht; 2006.
88. Olsen JA. Persons vs years: Two ways of eliciting implicit weights. Health Econ. 1994;3(1):39-46.
89. Gyrd-Hansen D. Comparing the results of applying different methods of eliciting time preferences for health. Eur J Health Econ. 2002;3(1):10-6.
90. Dolan P, Gudex C. Time preference, duration and health state valuations. Health Econ. 1995 Jul-Aug;4(4):289-99.
91. Devlin N, Buckingham K, Shah K, Tsuchiya A, Tilling C, Wilkinson G, et al. A comparison of alternative variants of the lead and lag time tto. Health Econ. 2012 Jun 19.
92. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011 Dec;20(10):1727-36.
93. Oppe M, Van Hout BA. The optimal hybrid: Experimental design and modeling of a combination of TTO and DCE. 27th scientific plenary meeting of the EuroQol group - proceedings; ; 2009.
94. Attema AE, Edelaar-Peeters Y, Versteegh MM, Stolk EA. What's affecting the TTO? . Forthcoming.
95. van Hout B, Janssen MF, Feng Y, Kohlman T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in health. In press.
96. Augestad LA, Rand-Hendriksen K, Kristiansen IS, Stavem K. Learning effects in time trade-off based valuation of EQ-5D health states. Value in Health(0).
97. Norman R, King MT, Clarke D, Viney R, Cronin P, Street D. Does mode of administration matter? comparison of online and face-to-face administration of a time trade-off task. Qual Life Res. 2010 May;19(4):499-508.
98. Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian valuation of EQ-5D health states: Preliminary value set and considerations for future valuation studies. PLoS One. 2012;7(2):e31115.
99. MacKeigan LD, Gafni A, O'Brien BJ. Double discounting of QALYs. Health Econ. 2003;12(2):165-9.
100. Dolan P, Gudex C. Time preference, duration and health state valuations. Health Econ. 1995;4(4):289-99.

101. Cairns J, Pol Mvd. Do people value their own future health differently from other's future health? Med Decis Making. 1999;19(4):466-72.
102. Attema AE, Bleichrodt H, Wakker PP. A direct method for measuring discounting and QALYs more easily and reliably. Medical decision making : an international journal of the Society for Medical Decision Making. 2012;32(4):583-93.
103. Versteegh MM, Attema AE, Oppe M, Devlin NJ, Stolk EA. Time to tweak the TTO: But how? . Forthcoming.
104. Loewenstein G, Prelec D. Negative time preference. Am Econ Rev. 1991;81 2:347-52.
105. Chapman GB. Temporal discounting and utility for health and money. J Exp Psychol Learn Mem Cogn. 1996 May;22(3):771-91.
106. Ganiats TG, Carson RT, Hamm RM, Cantor SB, Sumner W, Spann SJ, et al. Population-based time preferences for future health outcomes. Med Decis Making. 2000 Jul-Sep;20(3):263-70.
107. Brazier JE, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2004;13(9):873-84.
108. Kind P, Brooks R, Rabin R, editors. EQ-5D concepts and methods: A developmental history. The Netherlands: Springer; 2005.
109. Bruce B, Fries JF. The stanford health assessment questionnaire: A review of its history, issues, progress, and documentation. J Rheumatol. 2003 Jan;30(1):167-78.
110. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The multiple sclerosis impact scale (MSIS-29): A new patient-based outcome measure. Brain. 2001;124(5):962-73.
111. Aaronson NNK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The european organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst. 1993;85(5):365-76.
112. ten Klooster PM, TAAL E, van de Laar MAFJ. Rasch analysis of the dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. Arthritis Care & Research. 2008;59(12):1721-8.
113. Ramp M, Khan F, Misajon RA, Pallant JF. Rasch analysis of the multiple sclerosis impact scale MSIS-29. Health Qual Life Outcomes. 2009 Jun 22;7:58.
114. Versteegh M, Leunis A, Luime J, Boggild M, Uyl-De Groot CA, Stolk E. Mapping QLQ-C30, HAQ and MSIS-29 on EQ-5D. In press .
115. Young TA, Yang Y, Brazier JE, Tsuchiya A. The use of rasch analysis in reducing a large condition-specific instrument for preference valuation. Medical Decision Making. January/February 2011 January/February 2011;31(1):195-210.
116. Mavranezouli I, Brazier JE, Young TA, Barkham M. Using rasch analysis to form plausible health states amenable to valuation: The development of CORE-6D from a measure of common mental health problems (CORE-OM). Qual Life Res. 2010 Oct 23.
117. Pallant JF, Tennant A. An introduction to the rasch measurement model: An example using the hospital anxiety and depression scale (HADS). Br J Clin Psychol. 2007 Mar;46(Pt 1):1-18.
118. Tennant A, McKenna SP, Hagell P. Application of rasch analysis in the development and application of quality of life instruments. Value in health. 2004;7:s22-26.
119. Brazier JE, Roberts J, Platts M, Zoellner YF. Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health Qual Life Outcomes. 2005 Mar 15;3:13.
120. Stolk EA, Busschbach JJ. Validity and feasibility of the use of condition-specific outcome measures in economic evaluation. Qual Life Res. 2003 Jun;12(4):363-71.
121. Bruce B, Fries JF. The stanford health assessment questionnaire: Dimensions and practical applications. Health Qual Life Outcomes. 2003 Jun 9;1:20.
122. Versteegh M, Rowen D, Brazier J, Stolk E. Mapping onto EQ-5D for patients in poor health. In press 2010.
123. Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value Health. 2009 Jun 25.
124. McKenzie L, Van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D

instrument: The potential to estimate QALYs without generic preference data. Value in healthlue. 2009;12:167-71.
125. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. The European journal of health economics. 2010;11(4):427-34.
126. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Rheumatology. 1996 June 01;35(6):574-8.
127. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: A revised version of the health assessment questionnaire. Arthritis & Rheumatism. 2004;50(10):3296-305.
128. Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ. 1986;5:1-30.
129. Petrillo J, Cairns J. Converting condition-specific measures into preference-based outcomes for use in economic evaluation. Expert Rev Pharmacoeconomics Outcomes Res. 2008;8(3):453.
130. Versteegh MM, Rowen D, Brazier JE, Stolk EA. Mapping onto eq-5 D for patients in poor health. Health Qual Life Outcomes. 2010 Nov 26;8:141.
131. Williams A. The EuroQol instrument. In: Kind P, Brooks R, Rabin R, editors. EQ-5D concepts and methods: a developmental history. Dordrecht: Springer; 2005.
132. Bruce B, Fries J. The stanford health assessment questionnaire (HAQ): A review of its history, issues, progress, and documentation. J Rheumatol. 2003;30(4):167-78.
133. Bansback N, Marra C, Tsuchiya A, Anis A, Guh D, Hammond T, et al. Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis Rheum. 2007 Aug 15;57(6):963-71.
134. Segeren CM, Sonneveld P, van der Holt B, Vellenga E, Croockewit AJ, Verhoef GE, et al. Overall and event-free survival are not improved by the use of myeloablative therapy following intensified chemotherapy in previously untreated patients with multiple myeloma: A prospective randomized phase 3 study. Blood. 2003 Mar 15;101(6):2144-51.
135. Doorduijn JK, van der Holt B, van Imhoff GW, van der Hem KG, Kramer MH, van Oers MH, et al. CHOP compared with CHOP plus granulocyte colony-stimulating factor in elderly patients with aggressive non-hodgkin's lymphoma. J Clin Oncol. 2003 Aug 15;21(16):3041-50.
136. Boggild M, Palace J, Barton P, Ben-Shlomo Y, Bregenzer T, Dobson C, et al. Multiple sclerosis risk sharing scheme: Two year results of clinical cohort study with historical comparator. BMJ.British medical journal (Clinical research ed.). 2009;339(dec02 1):b4677-.
137. Tsuchiya A, Brazier J, McColl E, Parkin D. Deriving preference-based single indices from non-preference based condition-specific instruments: Converting AQLQ into EQ5D indices. Sheffield: ScHARR, Sheffield Health Economics Group, University of Sheffield, UK; 2002. Report No.: Discussion Paper Series 02/1.
138. Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value Health. 2009;**:**,**.
139. Versteegh MM, Leunis A, Birnie E, Oppe M, Stolk EA. Generalizability and development of dutch utilities for the EORTC QLQ-C30. In press 2010.
140. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: How reliable is the relationship? Health Qual Life Outcomes. 2009 Mar 31;7:27.
141. Franks PP, Lubetkin E, Gold M, Tancredi D, Jia H. Mapping the SF-12 to the EuroQol EQ-5D index in a national US sample. Medical decision making. 2004;24(3):247-54.
142. Rivero-Arias O, Ouellet M, Gray A, Wolstenholme J, Rothwell P, Luengo-Fernandez R. Mapping the modified rankin scale (mRS) measurement into the generic EuroQol (EQ-5D) health outcome. Medical decision making. 2010;30(3):341-54.
143. Ara R, Brazier J. Deriving an algorithm to convert the eight mean SF-36 dimension scores into a mean EQ-5D preference-based score from published studies (where patient level data are not available). Value Health. 2008 May 16.
144. Benito-Garcia E, Pedro S, Vasconcelos J, Marques RA, Rodrigues A, Chaves

I, et al. Different utility measurement instruments may not disciriminate across disease severity: Results from a cohort of rheumatoid arthritis patients in portugal. ISPOR 12th annual european congress; 24-27 October, 2009; Paris, France. ; 2009.
145. Yang M, Dubois D, Kosinski M, Sun X, Gajria K. Mapping MOS sleep scale scores to SF-6D utility index. Curr Med Res Opin. 2007 Sep;23(9):2269-82.
146. Geuskens GA, Hazes JMW, Barendregt PJ, Burdorf A. Work and sick leave among patients with early inflammatory joint conditions. Arthritis Rheum. 2008;59(10):1458.
147. Bansback N, Marra C, Tsuchiya A, Anis A, Guy D, Hammond T, et al. Using the health assessment questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis. Arthritis & Rheumatism. 2007;57(6):963-71.
148. Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, et al. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. Med Care. 2004;42(11):1125.
149. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Quality of life research. 1996;5(6):555-67.
150. Brazier J, Akehurst R, Brennan A, Dolan P, Claxton K, McCabe C, et al. Should patients have a greater role in valuing health states? Appl Health Econ Health Policy. 2005;4(4):201-8.
151. Krabbe PFM. A generalized measurement model to quantify health: The multi-attribute preference response model. Discussion paper for the EuroQol 29th plenary meeting. 2012.
152. Flynn TN, Louviere JJ, Marley AA, Coast J, Peters TJ. Rescaling quality of life values from discrete choice experiments for use as QALYs: A cautionary tale. Popul Health Metr. 2008 Oct 22;6:6.
153. Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate health state utility values. J Health Econ. 2012 Jan;31(1):306-18.
154. Attema AE, Brouwer WB. On the (not so) constant proportional trade-off in TTO. Qual Life Res. 2010 May;19(4):489-97.
155. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care. 2002 Feb;40(2):113-28.
156. Mosier CI. A critical examination of the concepts of face validity. Educational and Psychological Measurement. 1947;7:191-206.
157. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995.
158. Hunger M, Baumert J, Holle R. Analysis of SF-6D index data: Is beta regression appropriate? Value Health. 2011 Jul-Aug;14(5):759-67.
159. Nussbaum MC, Sen AK. The quality of life. Oxford: Clarendon Press; 1993.
160. Sen AK. Commodities and capabilities. Oxford: Oxford University Press; 1985.
161. Coast J, Smith R, Lorgelly P. Should the capability approach be applied in health economics? Health Econ. 2008 Jun;17(6):667-70.
162. Janssen MF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: A test of feasibility and face validity. Forthcoming in European Journal of Health Economics. 2013.

Health care expenditures have increased rapidly over the last decades in the Netherlands, in absolute terms and as percentage of gross domestic product. Curbing the rising health care costs has proven to be a very sensitive and complex societal issue. An important driver of rising costs is the availability and use of new and expensive medical technologies, causing a greater number of patients to be treated with more expensive interventions. Typically, these interventions do not only increase expenditures but also contribute to societal health and well-being.

Since health care expenditures are high on the political agenda, policy makers are interested in the relative effectiveness and efficiency of new medical interventions: do they achieve larger health effects than other treatments, and if so, at what additional costs? Economic evaluations address this question and the outcome of the evaluations is used to advise the Dutch government on the content of the basic benefit package of the Health Care Insurance act.

In this thesis it was attempted to improve the existing methodology for measuring quality of life for use in economic evaluations of medical interventions. The methods in this thesis contribute to a better measurement of the benefit of medical interventions in terms of quality of life.