UiT
THE ARCTIC UNIVERSITY OF NORWAY

Department of Community Medicine

Three essays on measuring health-related quality of life

External and internal relationships of the EQ-5D-5L

—

Thor Gamst-Klaussen

A dissertation for the degree of Philosophiae Doctor – August 2018

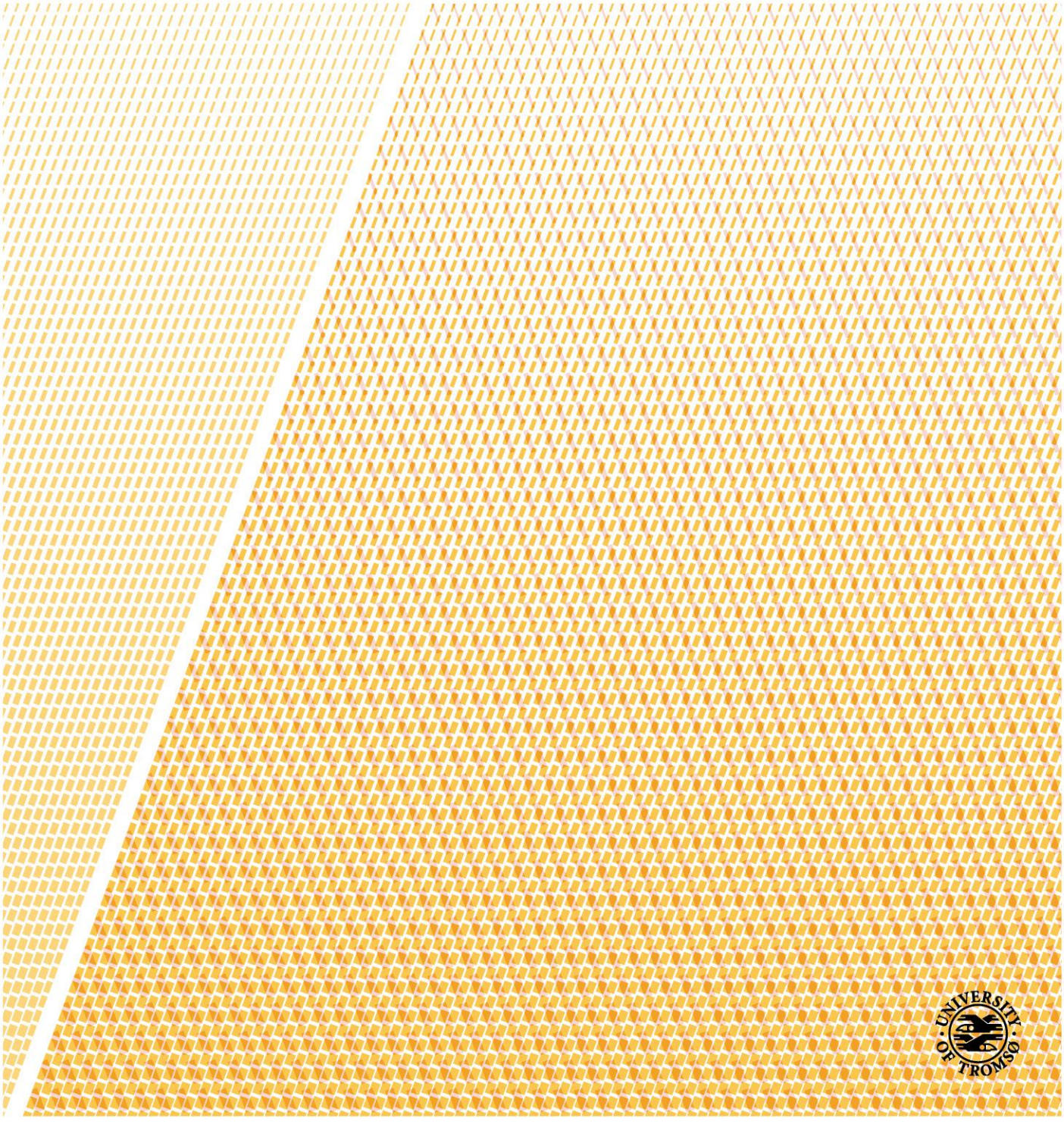

Table of Contents

Acknowledgments

I would like to thank the Research Council of Norway and the University of Tromsø (UiT) for providing financial support for the work described in this thesis. My thanks also goes to the Australian National Health and Medical Research Council and UiT for funding the Multi Instrument Comparison (MIC) project. I would especially like to thank Professor Jeff Richardson, at Monash University, Australia, who was the Principal Investigator of the very ambitious MIC project.

I would like to thank my PhD supervisor, Jan Abel Olsen. He always encouraged my work, and generously shared his ideas, knowledge and network with me. I am also grateful for the constructive comments and helpful insights from my coauthors Gang Chen and Claire Gudex.

I would also like to thank my coauthor, Admassu Nadew Lamu. We worked on the same project and shared office during our PhD studies. I am very thankful for his many valuable contributions and support during these years.

My deepest gratitude to my wife Sunniva, my son Arthur and my daughter Martha. Thank you for your love, encouragement and support.

List of abbreviations

15D	15 dimensional questionnaire
AQoL	Assessment of Quality of Life
BB	beta binomial
CFA	confirmatory factor analysis
CTA	confirmatory tetrad analysis
DASS-21	Depression Anxiety and Stress Scale 21-items
DCE	discrete choice experiment
df	degree of freedom
EQ-5D	EuroQoL five-dimension questionnaire (3L= three level; 5L= five level)
ER	exchange rate
FRM	fractional regression model
GPBM	generic preference-based measures
HRQoL	health-related quality of life
HUI2/3	Health Utility Index Mark 2 or Mark 3
K10	Kessler Psychological Distress Scale
MAE	mean absolute error
MDDC	maximum degree of differences in coefficients
MIC	Multi Instrument Comparison
PROM	patient-reported outcome measure
QALY	quality-adjusted life year
QRM	quantile regression model
QoL	quality of Life
QWB-SA	Self-Assessed Quality of Well-Being Scale
RMSE	root mean square error
SEM	structural equation model
SF-36	Medical Outcomes Study 36-item Short Form questionnaire
SF-6D	Short Form 6 dimensional questionnaire
SG	standard gamble
TTO	time trade-off
VAS	visual analogue scale
WHO	World Health Organization

List of publications

1. Gamst-Klaussen, T., Chen, G., Lamu, A. N., & Olsen, J. A. (2016). Health state utility instruments compared: inquiring into nonlinearity across EQ-5D-5L, SF-6D, HUI-3 and 15D. *Quality of Life Research, 25*(7), 1667-1678. https://doi.org/10.1007/s11136-015-1212-3
2. Gamst-Klaussen, T., Lamu, A. N., Chen, G., & Olsen, J. A. (2018). Assessment of outcome measures for cost-utility analysis in depression: mapping depression scales onto the EQ-5D-5L. *BJPsych Open, 4*(4), 160-166. http://doi.org/10.1192/bjo.2018.21 doi:10.1192/bjo.2018.21.
3. Gamst-Klaussen, T., Gudex, C., & Olsen, J. A. (2018). Exploring the causal and effect nature of EQ-5D dimensions: an application of confirmatory tetrad analysis and confirmatory factor analysis. *Health and Quality of Life Outcomes, 16*(1), 153. https://doi.org/10.1186/s12955-018-0975-y

Abstract

The use of quality-adjusted life years (QALYs) as a commensurable health outcome measure has been encouraged by health authorities in many countries in order to aid decisions on healthcare priorities. A key methodological challenge is to estimate the weights used for valuing health-related quality of life, i.e. the "Q" in QALY, based on people's preferences. Such generic preference-based measures (GPBMs) comprise a descriptive system and a value set that assign a value to each health state description on a 0 to 1 scale.

The objective of this thesis was to provide improved knowledge of the usefulness of GPBMs, with an emphasis on the most widely applied instrument, the EQ-5D. More specifically, the thesis aims to i) investigate into the degree of non-linear relationships across GPBMs and provide exchange rates that differ depending on disease severity (Paper 1); ii) develop mapping algorithms from depression scales (DASS-21 and K10) onto the EQ-5D (Paper 2) and iii) explore the causal and effect nature of EQ-5D dimensions (Paper 3). The analysis are based on an international sample from the Multi Instrument Comparison (MIC) project. A total of 7933 participants aged 18 years and above were included and separated into a non-diagnosed healthy group (n=1760) and seven disease groups (n=6173).

In Paper 1, quantile regression was used to investigate the degree of non-linear relationships between GPBMs (EQ-5D, SF-6D, HUI, and 15D) at nine different quantiles. Furthermore, the health state utility scale was split into intervals with 0.2 successive utility decrements to compare the GPBMs across different disease severities. The ER was calculated as the mean utility difference between two utility intervals on one GPBM divided by the difference in mean utility on another GPBM. The result revealed significant non-linear relationships across all four GPBMs. The degrees of non-linearity differed, with a maximum degree of difference in the coefficients (measured by the ratio of the largest to the smallest coefficient). ERs also

differed by disease severity: at the lower end of the health state utility scale, the ER from SF-6D to EQ-5D was 2.19, while at the upper end it was 0.35. These results illustrate the inaccuracy of using linear functions as cross-walks between GPBMs and suggest that level-specific exchange rates should be used when converting a change in utility on one GPBM onto a corresponding utility change on another GPBM.

Paper 2 aimed to develop mapping algorithms from two widely used depression scales: the Depression Anxiety Stress Scales (DASS-21) and the Kessler Psychological Distress Scale (K10) onto the EQ-5D-5L. Eight country-specific value sets (England, the Netherlands, Spain, Canada, China, Japan, Korea, and Uruguay) were applied. Data was based on the depression subgroup (n=917) of the MIC study. Six regression models were employed, including ordinary least squares regression, generalized linear models, beta binomial (BB) regression, fractional regression model, the MM-estimator, and censored least absolute deviation. Three model performance criteria were calculated to select the optimal mapping function for each country-specific value set: root mean square error, mean absolute error, and adjusted-r^2. Generally, the results revealed that the fractional regression model was preferred in predicting EQ-5D-5L utility values from both the DASS-21 and K10. The only exception was the Japanese value set, for which BB regression model performed best. The mapping algorithms can adequately predict EQ-5D-5L utility values from scores on the DASS-21 and K10. This enables disease-specific data from clinical studies to be applied to estimate outcomes in terms of QALYs for use in economic evaluations.

Paper 3 aimed to develop a conceptual framework for causal and effect relationships among the five dimensions of the EQ-5D (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) based on theoretical models of HRQoL, and test this framework using empirical data. The conceptual framework depicted the dimensions pain/discomfort and

anxiety/depression as causal indicators that drive a change in the effect indicators of activity/participation, mobility, self-care, and usual activities. Note that mobility has also an intermediate position between pain/discomfort and the other two effect dimensions (self-care and usual activities). Confirmatory tetrad analysis (CTA) and confirmatory factor analysis (CFA) were used to test this framework using the full sample from the MIC project (N=7933). CTA produced the best fit for a model specifying self-care and usual activities as effect indicators and pain/discomfort, anxiety/depression, and mobility as causal indicators. This was supported by CFA, which revealed a satisfactory fit to the data based on the comparative fit index=0.992, the Tucker-Lewis index =0.972, the root-mean square error of approximation =0.075 (90% CI 0.062-0.088), and the standardized root-mean square residual =0.012. The EQ-5D-5L appears to include both causal indicators (pain/discomfort and anxiety/depression) and effect indicators (self-care and usual activities). Although mobility played an intermediate role in our conceptual framework, the analysis suggested that it is mostly a causal indicator.

1 Introduction

The demand for healthcare service is growing continuously, and the healthcare sector has insufficient resources to meet these demands. Resources like staff, facilities, and equipment are limited, and decisions about which treatments to provide, for whom, where, and when are based on these resources [1]. Any one course of action will lead to fewer available resources for pursuing alternative services, so limited resources should be allocated in the best possible manner to produce the best health outcome. Therefore, evaluating the impact on both costs and health outcomes is a necessary part of choosing between competing services and interventions, or prioritizing different patients (i.e. rationing) [2].

In order to aid decision-makers in making efficient (i.e. value for money) and fair resource allocations, economic evaluations are required by government agencies such as the National Institute for Health and Care Excellence in the United Kingdom [3], the Norwegian Medicines Agency in Norway [4], and other similar agencies around the world [5,6]. While the overall purpose of economic evaluations is the comparative assessment of the costs and benefits of alternative healthcare interventions [7], the health consequences of an intervention are often less evident than the resource consequences. Since a health effect may be multi-dimensional, and in such a case is uncertain and may change, the measurement of the benefits of healthcare interventions is a critical part of an economic evaluation [1].

Different techniques of economic evaluation use different units to measure health benefits [7]. Indeed, in order to compare health interventions, the outcome must be measured on a common metric to identify the one that is the least costly per unit of outcome [1]. In a cost-utility analysis, this common metric is the quality-adjusted life year (QALY), which is a generic measure of health gain that combines the effects of an intervention on quality of life (QoL) and quantity of life. This is achieved by multiplying quality, i.e. the desirability of a

health state in terms of health-related QoL (HRQoL) by the duration of that health state (e.g. in years) [8]. The difficult task is measuring the quality weight (or the Q) in QALY.

There are a vast number of measures that have been developed to capture treatment effects as expressed by patients' own experiences, often referred to as patient-reported outcome measures (PROMs) [9]. PROMs may comprise one or multiple dimensions of health, assessing symptom(s), functional and health status, HRQoL, or QoL [10]. PROMs allow individuals to report their own experience on various health dimensions using a descriptive system; they provide a numeric value of health, which can be used to assess the efficacy and efficiency of interventions from a patient perspective [11-13]. However, most PROMs are disease-specific, making them less relevant for comparison across patient groups with different diseases. For this purpose, a generic measure is required. Furthermore, to be commensurable, trade-offs between health dimensions must be made to indicate the relative importance that people place on these dimensions. These measures are referred to as generic preference-based measures (GPBMs).

GPBMs have been developed to obtain the quality weights (also referred to as utility values) needed to calculate the QALY [14]. Utility values are derived from two components: a generic descriptive system that allows patients to report their health state and a pre-determined value set that provides values for each health state produced by the descriptive system. The values reflect an average of individuals' preferences for the health states, which are elicited using health state valuation techniques like standard gamble (SG), time trade-off (TTO), visual analogue scale (VAS), or discrete choice experiment (DCE) [15]. There are six primary GPBMs in use, including the EuroQoL 5 dimensional questionnaire (EQ-5D), the Short Form 6 Dimensional Questionnaire (SF-6D), the Health Utility Index Mark 2 or Mark 3 (HUI2/3), the 15 Dimensional Questionnaire (15D), the Assessment of Quality of Life (AQoL), and the Self-Assessed Quality of Well-Being Scale (QWB-SA) [14]. These

instruments can be applied across a range of patient groups and health conditions. All GPBMs purport to measure the same construct, which is utility. Here, utility is understood as a preference-based health state value that is anchored at 1 (full health) and 0 (being dead). However, studies indicate major discrepancies in the health state values produced by the different GPBMs for the same respondents. This is because GPBMs differ considerably in terms of the content and size of the descriptive system they use, as well as in the methodologies used for eliciting preference weights [16,17]. Thus, the intended comparability of studies is problematic when different GPBMs have been applied to measure the Q in the QALY.

The problem of incommensurability of studies using different GPBMs has led some reimbursement agencies to choose a single GPBM for consistency in utility values. For instance, the EQ-5D is preferred by reimbursement agencies in the United Kingdom and Norway [3,4] and is the most widely used GPBM. A review by Richardson et al. [18] found that the EQ-5D was applied in 63% of studies that applied a GPBM during the period 2005 to 2010, followed by the HUI-3 (9.8%), SF-6D (8.8%), and 15D (6.9%). Furthermore, the EQ-5D has dominated in most countries, except for the HUI in Canada and 15D in Finland. Another review by Wisloff and colleagues confirmed the dominant position of EQ-5D by revealing its application in 77% of cost-utility analysis published in 2010 [19].

Another problem is that clinical trials more often include a disease-specific measure (DSM) than a GPBM [20]. Since a DSM is incommensurable, one solution is to develop transformations (or exchange rates) that enable the estimation of utility data based on responses given on a DSM. However, even with available utility data, transformations are necessary to either estimate health state utility values for the GPBM preferred by a health authority, or to enable comparisons of health effects [21]. This procedure is commonly referred to as mapping or cross-walking [20,22], which is the main focus of this thesis.

Moreover, due to the central role of the EQ-5D as the preferred GPBM among health authorities and its widespread use in applied studies, this thesis will concentrate particularly on this GPBM and will focus mainly on the new 5-level version (EQ-5D-5L), which includes the application of recently developed country-specific value sets, making this thesis timely and highly relevant.

Mapping helps to reconcile the differences in health effects measured by different GPBMs. However, for mapping to be valid, there are some caveats that need particular attention. Studies have indicated non-linear associations between different GPBMs, and between GPBMs and DSMs [20,22-24]. However, previous studies on mapping have mostly applied linear transformations [20,22]. This implies that linear transformations would produce biased estimates at some part of the scale, usually at the top and/or bottom end. Hence, if mapping is to improve the comparison of health effects produced by different GPBMs, the critical fact that the strength of the association across GPBMs has been shown to vary at different disease severity levels should not be ignored. More knowledge about the presence of non-linear relationships across GPBMs is important, since it would advocate the use of non-linear transformations that could better harmonize the magnitude of units across GPBMs at different severity levels. Thus, Paper 1 of this thesis is the first study to specifically investigate non-linearity across GPBMs (EQ-5D-5L, SF-6D, HUI-3 and 15D) using a novel approach, quantile regression models (QRMs). QRMs allow researchers to investigate the effect of one measure across the whole distribution of another measure. Furthermore, Paper 1 explored exchange rates between GPBMs at different severity levels, which has not been previously done. This has important policy implications, particularly when decision-makers are comparing alternative programs whose QALY calculations are based on different GPBMs.

Paper 2 focused on developing mapping algorithms from two DSMs, the Depression Anxiety and Stress Scales 21-items (DASS-21) and Kessler Psychological Distress Scale

(K10), which are widely used measures of depression. This disease group was selected for several reasons. First, depression is a prevalent condition across all age groups, peaking in older adulthood. Globally, depressive disorders have been increasing in the last decade [25], and they are the single largest contributor to non-fatal health loss. The condition is different in the sense that it might last for longer periods, or may reoccur, significantly impairing an individual's ability to function at work or school or to cope with daily life [26]. Depression can range from mild to severe; at its most severe, it can lead to suicide. Secondly, mental health is receiving increasing health policy attention, which will raise the demand for comparative assessments of healthcare interventions that target this patient group. Lastly, as a psychologist, this disease group has been of prime personal interest, as was the goal of contributing knowledge about DSMs applied in mental health, and my interest in investigating mapping from mental health measures onto GPBMs. Furthermore, based on the knowledge from Paper 1, in addition to other commonly applied models in mapping studies, two novel regression models were applied to seek out optimal transformations: a fractional regression model (FRM) and a beta binomial (BB) regression model that both account for the non-linearity in the data.

While Papers 1 and 2 in this thesis concentrate on mapping, Paper 3 is more conceptual and reflects on how different dimensions of health are interconnected. Based on recommended models for conceptualizing the relationships between dimensions of HRQoL, Paper 3 is the first study to develop and empirically test a conceptual framework for causal and effect indicators among the five dimensions of the EQ-5D-5L. More knowledge on the causal pattern provides a better conceptualization of the underlying structure of the EQ-5D-5L, and might provide a better understanding of the relative importance of the five health dimensions as reflected in the preference-based value sets, as well as give insights into how to extend the descriptive system. A relatively new approach, referred to as confirmatory tetrad analysis,

was applied to determine whether EQ-5D-5L dimensions should be treated as causal or effect indicators.

2 Background

2.1 Concepts and definition of terms

While health and QoL are everyday concepts used by laypersons, HRQoL is a concept used more among researchers. Although these terms are conceptually different, they are often used interchangeably, which can create confusion about their meaning. There is no single definition for either of these terms and still a debate about how to define them [27,28]. To aid in the understanding of this thesis, a brief definition of each term is given below.

2.1.1 Health

The World Health Organization (WHO) broadly defined the term *health* in 1948 as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" [29]. This has remained a highly influential definition. Yet, others have left out the mention of social well-being, defining health as "an individual's level of function", where "optimum function" is judged in comparison to "society's standard of physical and mental well-being" [30]. Other more recent definitions have emphasized social and personal resources, as well as physical capacity [31], putting more emphasis on the capacity to cope autonomously with life's ever-changing physical, emotional, and social challenges [32].

2.1.2 Quality of life

QoL is a broad-ranging concept that covers all aspects of people's lives [33]. Although there are several definitions [28], the WHO defined QoL as "individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by the persons' physical health, psychological state, level of

independence, social relationships, personal beliefs and their relationship to salient futures of their environment" [34].

2.1.3 Health-related quality of life

The term HRQoL first appeared in the literature on health status measures in which HRQoL was used in the discussion of QALY as a measure of the value of 1 year of full health [28]. Thereafter, the use of HRQoL spread and refers to QoL when considered in the context of health and disease [35]. Thus, the term distinguishes the effects of illness and treatment from aspects of life that are beyond health care, e.g. political, societal, or cultural circumstances [36]. HRQoL is a dynamic multi-dimensional concept [37] and can be defined as "how well a person functions in her life and her perceived well-being in physical, mental, and social domains of health" [38]. Functioning refers to observable behaviors such as an individual's ability to perform pre-defined activities, including the ability to interact with family and friends, and to participate in one's work or studies [38,39], while well-being refers to the individual's internal subjective perceptions and feelings [38]. Accordingly, HRQoL goes beyond direct measures of health and focuses on the consequences of health status on QoL [35,36]. Another definition focuses on the value of health. Here HRQOL can refer to "the value assigned to different health states", and these values, or utilities, are on a 0-1 dead-healthy scale needed for QALY estimation, where values less than 1 indicate the loss of QoL or living in ill health [28]. Thus, as stated by Karimi and Brazier [28], if respondents' preferences for health states reflect the impact of health on QoL, and they are able to estimate that impact correctly, then the utility of health states could be referred to as HRQoL. Although defining HRQoL has proven challenging, and several definitions have been proposed in the literature [40-43], generally there is a consensus that HRQoL is a multi-dimensional concept that at least includes physical, mental, and social dimensions [36].

2.2 Health-related quality of life measures

2.2.1 Types of health-related quality of life measures

The field of HRQOL assessment has become more sophisticated and methodologically rigorous [44], and there is a wide range of measures available [45,46]. These measures can be broadly divided into disease-specific versus generic. The vast majority are disease-specific, which measure how patients perceive the impact of a certain disease or health condition. Thus, the content of a DSM should be relevant for patients suffering from that health problem. Since all or most of the content comprising such a measure is relevant for the patients under study, the measure is generally thought to have a greater degree of precision to detect differences in severity and important changes over time [47]. The disadvantage of DSMs is the fundamental problem of comparability of outcomes of different treatments across patients groups with different health problems and diagnoses. Furthermore, DSMs may miss the impact of unanticipated problems related to the disease or side effects of treatments, as well as the impact of possible comorbidities [47].

Generic measures assess a broad range of different health aspects across all types of morbidity and are often applicable to the general population [46]. These measures allow for comparison of scores across patients with various diseases or against the general population. Generic HRQoL measures can further be divided into non-preference-based and preference-based measures [14], also referred to as psychometric profiles or utility measures [48]. The most widely used non-preference-based measure in clinical trials is the Medical Outcomes Study 36-item Short Form (SF-36) [49,50]. SF-36 provides a profile or description for assessing a patient's health across eight different dimensions, i.e. physical functioning, social functioning, role limitations-physical, role limitations-emotional, bodily pain, vitality, mental health, and general health. The profile scores for each dimension indicate performance relative to both the

maximal and minimal level and, if calibrated to a population standard, the degree of health impairment in comparison to a population of interest [48]. However, although it is not uncommon in the literature [51], combining the dimensions of the SF-36 into an overall score, or total score, to measure health changes is not advisable, as it could lead to misinterpretation of any change [51]. Additionally, non-preference-based measures are not commensurable with lifetime gains. As a result, for health economic evaluations, i.e. cost-utility analyses, a preference-based HRQoL measure is essential to produce a cardinal index of health on a 0-1 dead-healthy scale, where changes on this quality scale are commensurable with changes on the quantity of life scale.

2.3 Generic preference-based measures

In health economic evaluations, it is essential to make healthcare programs comparable in terms of their cost-effectiveness. Effectiveness in producing health outcomes is measured by the QALY, and quality adjustment in the QALY needs to be measured in a way that systematically indicates the significance of various health effects in terms of HRQoL [8]. In this context, HRQoL is measured using a GPBM, also referred to as a multi-attribute utility instrument [15] or health state utility instrument [52]. Hereafter the term GPBM is used when referring to generic preference-based HRQoL measures in this thesis [1,7]. A GPBM consists of a descriptive system and a value set that assigns preference weights, or utility values, to each health state produced by the descriptive system.

2.3.1 Descriptive systems

There are six GPBMs described in the literature, and most provide more than one version [14]. These include the EQ-5D (EQ-5D-3L and EQ-5D-5L) [53,54], the SF-6D, derived from either the SF-12 or SF-36 [55,56], the HUI-2/HUI-3 [57,58], the 15D [59], the QWB-SA [60] and AQoL (AQoL-4D/AQoL-6D/AQOL-7D/AQoL-8D) [61-64]. They differ in terms of

descriptive systems, with a differing number of items/dimensions. Some dimensions are unique to one measure, while similar dimensions include different items, and there may be a different number of severity levels for each item/dimension. Indeed, the number of items and dimensions vary considerably (Table 1): some measures include one item per dimension (e.g. EQ-5D and HUI-3), while others include several items (e.g. SF-6D and AQoL-8D). Since each GPBM includes a different number of dimensions, and the level of the dimensions are different across descriptive systems, each GPBM defines a different number of health states. Due to the dominant position of the EQ-5D, and since it is the primary focus in this thesis, it will be described in more detail to exemplify how an individual's health state is defined. Other measures included in this thesis are described in the appendix.

The EQ-5D is the shortest GPBM and includes five items/dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. In the original version (EQ-5D-3L), which was developed more than 25 years ago, each item had three levels, thus it defined 243 health states (3^5) [53]. More recently, a five-level version, the EQ-5D-5L, was developed [54] to respond to concerns about the insensitivity of the EQ-5D-3L. In the EQ-5D-5L, two more response levels were added to each dimension to reduce potential ceiling effects and improve reliability and sensitivity [54,65]: the level 'slight problems' was added in between "no problems" and "moderate problems", and the option "severe problems" was added in between "moderate problems" and "unable to/extreme problems" (Box 1). When responding to the EQ-5D-5L, the health state is determined by taking one level from each dimension. That is, the best possible health state (or full health) is defined as a response of no problem (level 1) on every dimension (i.e. 11111), while the worst possible health state is described by unable to/extreme problems (level 5) on every dimension (i.e. 55555). When including every other health state combination between best and worst health states, the EQ-5D-5L defines a total of 3125 (or 5^5) health states. Thus, the more dimensions and levels included in a GPBM

Table 1. Descriptive systems of GPBMs

GPBM	(N) Dimensions	Items	Response levels	Health states defined	Relative use (%)	
					Study 1[a]	Study 2[b]
EQ-5D-5L/3L	(5) Mobility, self-care, usual activities, pain/discomfort, and anxiety/depression	5	5/3	3,125/243	63.2	77.0
SF-6D[c]	(6) Energy, mental health, pain, physical functioning, role limitation, and social functioning	11	4 to 6	18,000	8.8	11.5
HUI-2	(7) Sensation, mobility, emotion, cognition, self-care, pain, and fertility	7	3 to 5	24,000	4.6	5.3
HUI-3	(8) Vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain	8	5 to 6	972,000	9.8	
15D	(15) Mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort/symptoms, depression, distress, vitality, and sexual activity	15	5	31 billion	6.9	4.4
QWB-SA[d]	(4) Mobility, physical activity, social functioning; 68 symptoms/problems	71	2 to 3	945	2.4	1.8
AQoL[e]	(8) Coping, happiness, independent living, mental health, pain, relationship, self-worth, and senses	35	4 to 6	$2.37*10^{23}$	4.3	0.0

[a] The relative use of GPBMs was based on 1682 studies published between 2005 and 2010. Of these, 15% were primarily concerned with economic evaluations (for details, see Richardson et al. [18]).
[b]The relative use of GPBMs was based on 370 studies published in 2010 (for details, see Wisloff et al. [19]
[c]SF-6D derived from SF-36.
[d]The three multi-response items of the QWB-SA - mobility, social activity, and physical activity - define 47 health states, and the remaining symptom/problem groups define 898 health states.
[e]AQoL-8D has 35 items comprising of eight dimensions.

descriptive system, the more health states can be defined, e.g. 15D comprises 15 dimensions with 5 levels each, defining more than 30 (or 5^{15}) billion health states. After respondents describe their health state, the next step is to apply an algorithm that assigns a preference

Box 1: EQ-5D-5L descriptive system

Select the answer under each heading below that best describes your own health state *today*

a) Mobility

- I have no problem in walking about ☐
- I have slight problems in walking about ☐
- I have moderate problems in walking about ☐
- I have severe problems in walking about ☐
- I am unable to walk about ☐

b) Self-care

- I have no problems washing or dressing myself ☐
- I have slight problems washing or dressing myself ☐
- I have moderate problems washing or dressing myself ☐
- I have severe problems washing or dressing myself ☐
- I am unable to wash or dress myself. ☐

c) Usual activities *(e.g. work, study, housework, family or leisure activities)*

- I have no problems doing my usual activities ☐
- I have slight problems doing my usual activities ☐
- I have moderate problems doing my usual activities ☐
- I have severe problems doing my usual activities ☐
- I am unable to do my usual activities ☐

d) Pain/discomfort

- I have no pain or discomfort ☐
- I have slight pain or discomfort ☐
- I have moderate pain or discomfort ☐
- I have severe pain or discomfort ☐
- I have extreme pain or discomfort ☐

e) Anxiety/depression

- I am not anxious or depressed ☐
- I am slightly anxious or depressed ☐
- I am moderately anxious or depressed ☐
- I am severely anxious or depressed ☐
- I am extremely anxious or depressed ☐

weight, or an index value, to each health state. These algorithms are developed based on methods for measuring preferences on a 0-1 scale, where 0 equals being dead and 1 equals full health. However, negative values are also possible and indicate health states that are considered worse than being dead. There are four valuation techniques commonly referred to in the literature for valuing health states, namely the VAS, SG, TTO, and more recently, the DCE (Table 2).

Table 2. Valuation techniques of GPBMs

GPBM	Valuation technique	Forms of algorithm	Scoring formula	Minimum score
EQ-5D[a]	TTO, DCE,	Statistical	Additive	3L: -0.594[a] 5L: -0.281[b]
SF-6D	SG	Statistical	Additive	0.301
HUI-3	SG, VAS	MAU[c]	Multiplicative	-0.36
15D	VAS	MAU	Additive	0.00
QWB-SA	VAS	MAU	Additive	0.00
AQoL-8D[d]	TTO, VAS	Statistical and MAU	Multiplicative	-0.04

[a] The minimum score for the UK value set (for details, see Dolan [66]).
[b] The minimum score for the English value set (for details, see Devlin et al. [67])
[c] The MAU theory reduces the valuation task by making simplifying assumptions about the relationship between dimensions (for details, see Brazier et al. [1]).
[d]AQoL-8D employs both MAU theory and statistical modelling to estimate a function for valuing health states.

2.3.2 Valuation techniques

Visual analogue scale

VAS is a line (usually presented vertically) with well-defined endpoints, on which the value 0, located at the lower end, indicates the worst imaginable health or being dead, and 100, located at the upper end, indicates the best imaginable or full health (Figure 1). Since respondents are asked to judge, value, or feel where their health state is located on the scale, the VAS is sometimes referred to as a feeling thermometer. The VAS is considered to have interval properties, where the distance between intervals reflects a respondent's preference for the different health states being measured. Thus, the difference in health from 10 to 20 should be equal to the differences between 60 and 70.

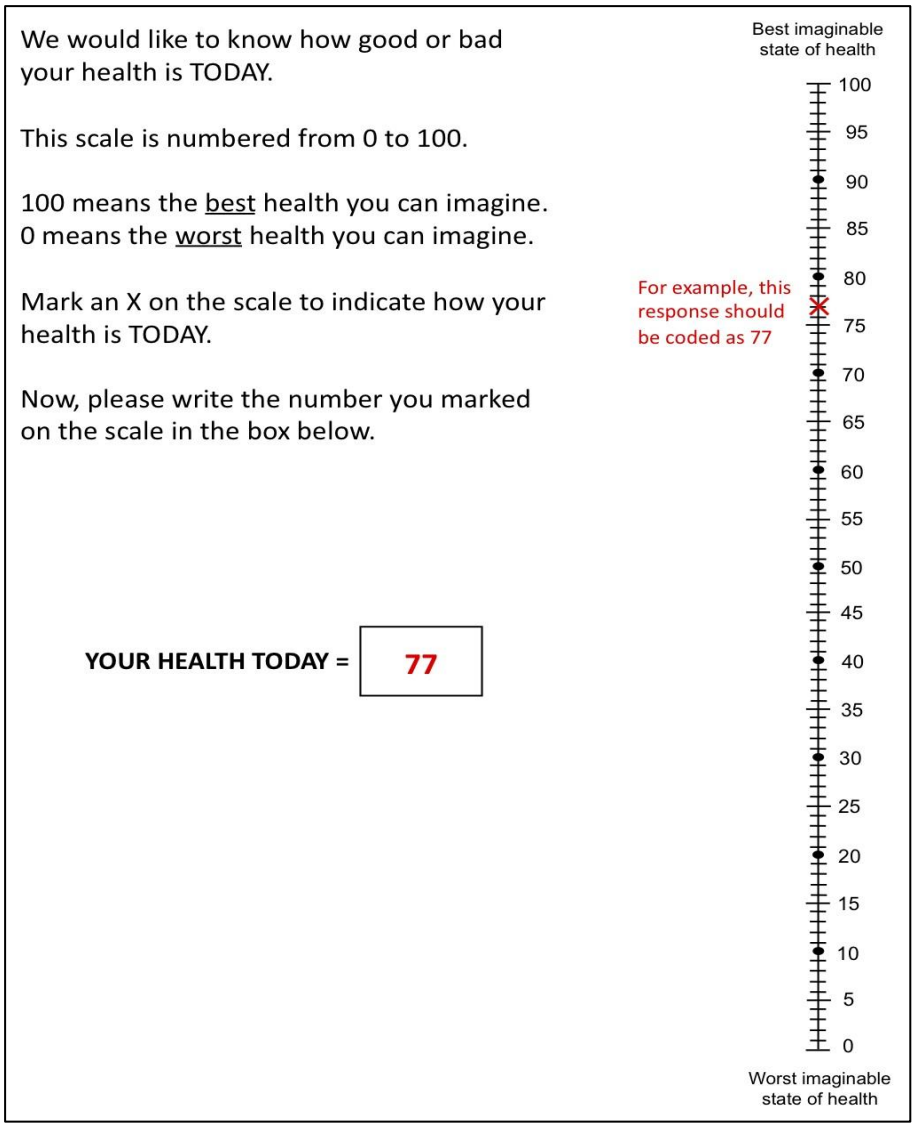

Figure 1. Example of a VAS

Standard gamble

SG is the classic method for measuring preferences, based directly on the axioms of the von Neumann and Morgan utility theory (86). The utility for a health state is the amount of risk, in terms of probability, the respondent is willing to accept for not being in the valued health state. The preference for the health state is the point where the respondent becomes indifferent between two treatment outcomes: one which involves uncertainty with two possible outcomes and one that has an intermediate certain outcome.

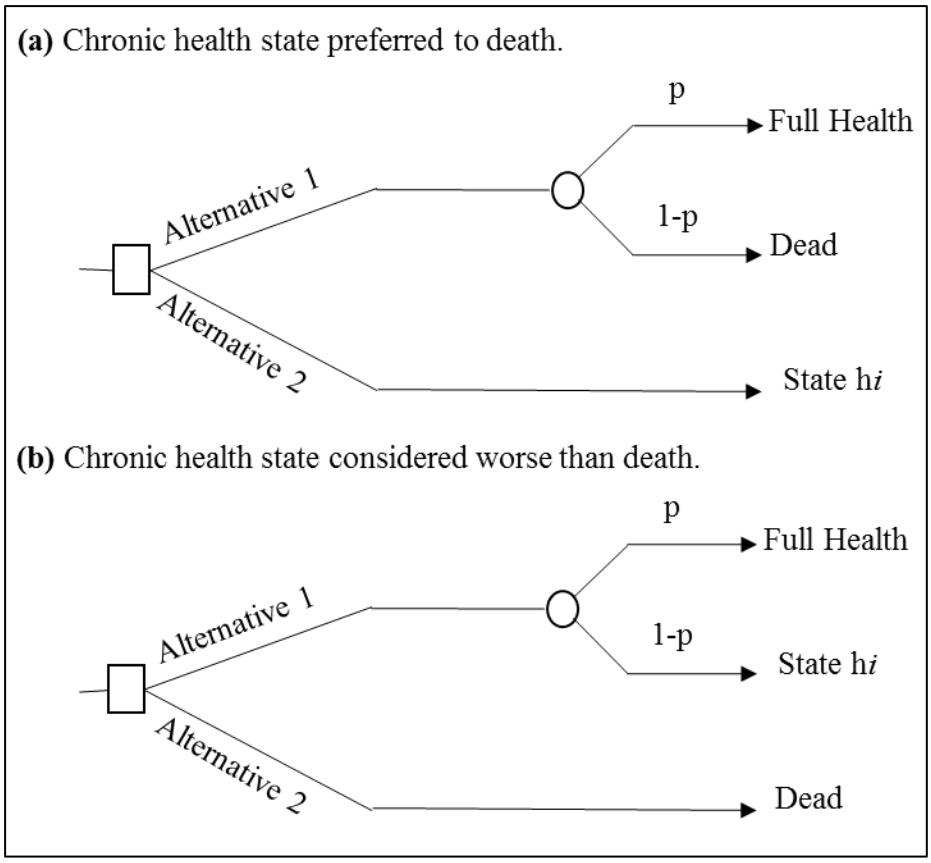

Figure 2. Valuation of health states with the SG

The basic format for the SG in the case of chronic health states that are preferred to being dead is illustrated in Figure 2a. Alternative 1 has two possible outcomes: either the patient returns to full health and lives for additional t years (with probability p) or the patient dies immediately (with probability 1-p). Alternative 2 is a certain outcome of chronic state h_i for life (t years). The probability p of successful treatment (returning to full health) varies until the individual is indifferent between the risky option (alternative 1) and the certain outcome (alternative 2), at which the value of h_i is equal to p. That is:

h_i=p* full health + (1-p) dead dead => h_i=p, where full health=1 and dead=0

In the case of chronic health states considered worse than being dead, the SG can be slightly modified by making the certain outcome (alternative 2) dead, and making the risky option

(alternative 1) a gamble between full health after treatment (probability p) or remaining in the chronic health state (h_i) for life (probability 1-p) (86). As before, probability p varies until the respondent becomes indifferent between the certain outcome of death and the risky option, at which the utility for h_i=-p/ (1-p) (Figure 2b). In the literature, the SG is often regarded as the preferred method due its theoretical basis and the fact that one of its features is decision-making under uncertainty [1]. Indeed, since medical decisions usually involve uncertainty, the SG is often labeled the gold standard due to its uncertain nature [7]. However, it has been argued that the type of uncertainty in the SG is not comparable with the various uncertainties in medical decisions, making this feature less relevant [68]. Furthermore, individuals often have a hard time understanding probabilities.

Time trade-off

Torrance et al. [69] introduced the TTO method to provide a simpler method than the SG. Both methods derive preferences implicitly based on the respondent's choices in given situations. However, while the SG is risk-sensitive due to uncertainty in the outcome, the TTO is riskless. The basic format for the TTO in the case of chronic health states that are preferred to being dead is illustrated in Figure 3a. The respondents are offered two alternatives: alternative 1 is health state i for time t (usually 10 years) followed by death; alternative 2 is healthy for time x ($x < t$) followed by death. If the respondent is willing to trade life expectancy, time x is varied until the respondent is indifferent between the two alternatives, at which point the preference value for chronic state i is: $h_i=x/t$.

For chronic health states considered worse than being dead, the TTO can be altered so that respondents can choose between immediate death (alternative 1), or health state h_i for a period of time (y), followed by x years in full health where $x+y=t$. By varying time (x) until the respondent is indifferent between the two alternatives, the value of h_i can then be given as: h_i $=-x/(t-x)$. Thus, the score for state h_i will be lower if more time in full health is needed to

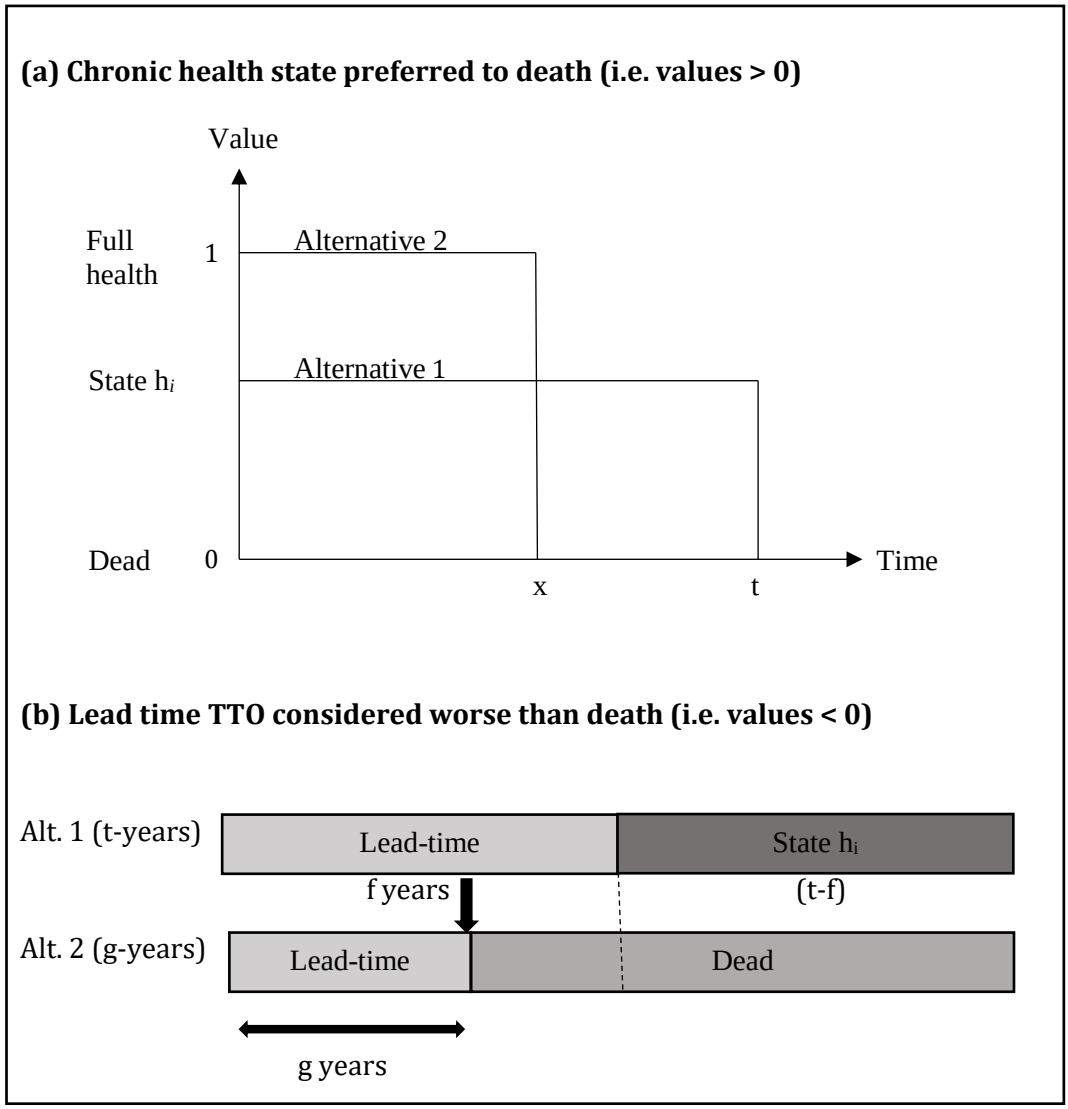

Figure 3. Valuation of health states with (a) conventional TTO and (b) lead-time TTO

compensate for the time spent in h_i. The formula translating TTO responses to health state values considered worse than being dead produces a scale that ranges from minus infinity to 1, where a greater weight is given to negative values, which has implications for economic evaluations [52,70,71]. While this has been resolved by assigning a preference value of -1 to the worst possible health state among those health states considered worse than being dead [66], this value is arbitrary and has no theoretical support [72].

A solution has recently been proposed to handle values for states considered worse than being dead, referred to as lead–time TTO. This method was introduced as an alternative to the conventional TTO and is applicable for health states considered either better or worse than being dead [70]. This approach involves adding additional time spent in full health before the period spent in the impaired health state (alternative 1), as well as to the period spent in full health or death (alternative 2) [1]. That is, alternative 1 is full health for f years (lead-time), then state h_i for t-f years; while alternative 2 is full health for g years (lead-time), where g is larger than f for health states considered to be better than being dead, and less than f for states considered to be worse than being dead. The latter is illustrated in Figure 2b. State h_i is then calculated as $h_i=(g-f)/(t-f)$. Health states considered better than being dead receive a positive value and states considered worse than being dead receive a negative value [73].

Since studies have shown the lead-time TTO exercise has severe framing effects, and it is clearly difficult for respondent to perform the task [70,71,74], a composite TTO was introduced as a compromise between the conventional TTO and lead-time TTO [75]. Thus, the composite TTO considers the conventional TTO for health states considered better than being dead and the lead-time TTO for states below zero. The EuroQol group adopted this approach for the valuation of the EQ-5D-5L, which improved the means of eliciting values worse than being dead and resolved the problem of assigning an arbitrary value to the worst possible health state for rescaling the conventional TTO [76]. While the TTO was developed to be a simpler alternative to SG, there is still a concern that it is cognitively demanding for some populations, leading to several inconsistencies and subsequent exclusions that limit the representativeness of the values produced [77,78].

Discrete choice experiment

Another method for eliciting preferences is the DCE, which has been promoted as a simpler method than the conventional iterative TTO task [79]. In the DCE method, respondents are

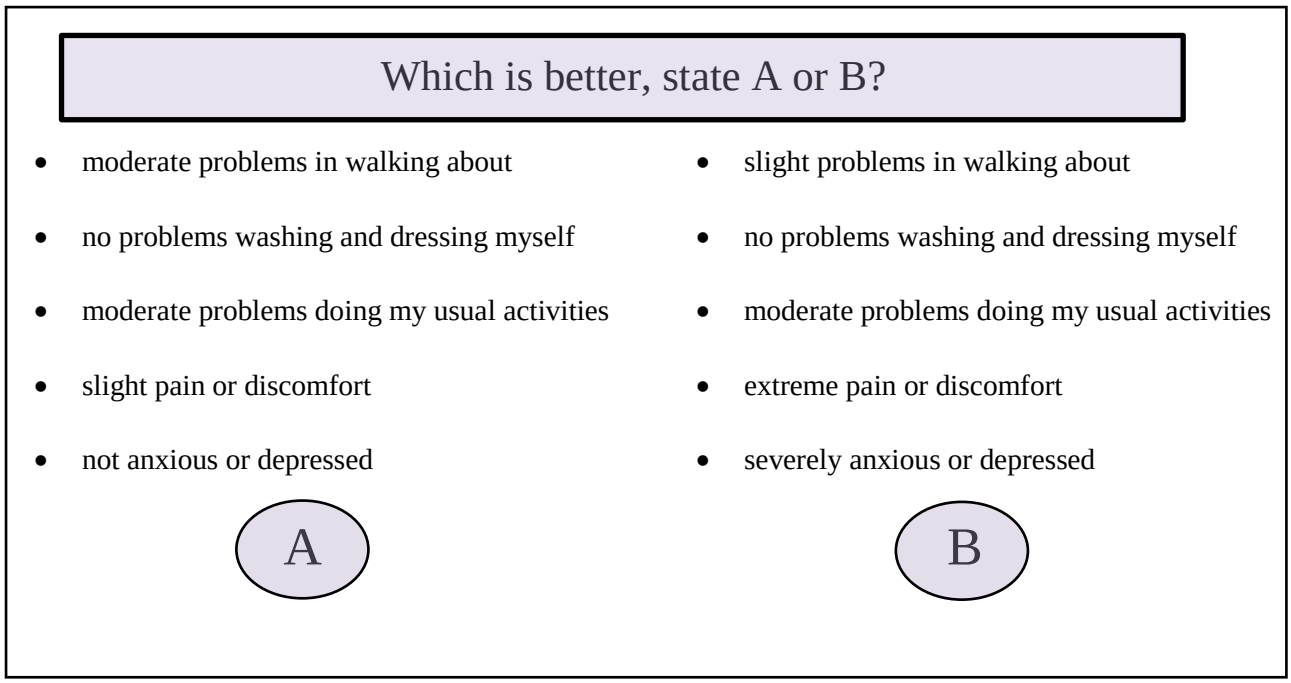

Figure 4. Example of two profiles from the EQ-5D-5L descriptive system provided in a DCE task

provided with two or more profiles, and they choose the most or least preferred, as exemplified in Figure 4. Different scenarios are constructed based on a descriptive system made up of levels of a limited number of important attributes [80].

Respondents simply indicate that option A is preferred to option B, without the iterative process used in the TTO to find the point of indifference between the two options [81]. The relative preferences of one health state over another are commonly provided by the conditional logit model [79]. Since the utility scale for DCE data from this model is not anchored to the 0-1 dead-healthy scale, it cannot be directly incorporated for calculating QALY. The EuroQoL valuation protocol includes DCE in addition to a TTO task and combines data from both techniques into a single modelling framework, referred to as the hybrid model [67,82]. The coefficient from both models are connected via a link function to account for the differences between the scales [83]. The hybrid model maximizes the use of

the available data from a valuation study using the EuroQoL valuation protocol [82]. However, there are promising approaches referred to as DCE_{TTO} that link health states to normal health and death within a DCE by including 'survival duration' as an attribute [67,81,84,85].

Comparison of valuation techniques

Different valuation techniques normally produce different values for the same health states. The majority of studies suggest that the VAS generally generates lower values than SG and TTO [86]. However, it has been shown that milder health states generate lower SG values than the VAS, with a crossover point at around 0.8 on a 0-1 dead-healthy scale [87]. While the relationship is less consistent for studies reporting VAS and TTO results, VAS usually generates lower values [86,88]. Studies have also indicated inconsistent results for the relationship between the SG and TTO. As the SG involves uncertainty, it has been suggested that it produces higher values than the TTO due to risk aversion. This is a key difference between the choice-based techniques. Nevertheless, one study suggested that the TTO produces higher values for milder health states, with a crossover point around 0.4 when compared to the VAS [87]. Additional sources of bias that may lead to differences in SG and TTO values include: probability weighting (upward bias in SG values), utility curvature (downward bias in TTO values), loss aversion (upward bias in both TTO and SG values), and scale compatibility (upward bias in TTO values and ambiguous bias in SG values, respectively) [1,89]. Probability weighting does not affect TTO values since these are elicited under the condition of certainty, while the SG is not affected by utility curvature since no restrictions are imposed on the utility function for the duration of the health state. Considering the relationship between the DCE and TTO, one study showed that the DCE assigns relatively higher values to milder health states and lower values to poorer health states [90].

2.3.3 Generic preference-based measures compared

The majority of studies comparing GPBMs suggest a low level of agreement [18]. It has generally been indicated that health state utility values are not equivalent across measures, and comparisons "warrant caution" [23,91]. While mean scores have been found to be similar, they often mask major differences across the distribution [14,23,92]. The differences across measures can be explained in part by differing descriptive systems, valuation techniques, and the model used to create the formula or algorithm used to extrapolate results. Studies have suggested that the dominant reason for these differences is the lack of overlap in the descriptive systems [16,93]. Similarly, the scale effect that arises from the use of different valuation techniques is also an important source of variation. One approach to increase comparability across GPBMs is to develop mapping algorithms that can predict health state utility values from one GPBM based on values from another GPBM.

2.4 Transformations

2.4.1 The concept of mapping

Mapping is conducted to link outcome data collected in clinical trials or observational studies to a GPBM to obtain utility values. Key clinical trials are often designed for purposes other than economic evaluations, hence a GPBM is not necessarily included as a PROM. In the scenario of missing utility data, mapping or "cross-walking" is one solution to enable cost-effectiveness analyses [94]. This approach enables the transformation of scores from a source measure, usually a condition-specific measure, into health state utilities, by applying a pre-existing mapping algorithm. Generally, mapping algorithms are developed by distributing both measures of interest to the same respondents, then applying statistical methods to predict health state utilities from scores on a source measure. Subsequently, the mapping algorithm can be applied to transform condition-specific data from clinical trials into health state utility

values [1,22]. Here I focus on mapping onto the EQ-5D for two reasons. First, it is the most widely applied GPBM in mapping studies, and in cost-utility analysis in general; second, the EQ-5D is of main interest in all three papers in this thesis.

2.4.2 The literature on mapping studies: the case of the EQ-5D

The practice of mapping onto the EQ-5D from other measures of health outcome is increasing in number, especially after the UK National Institute for Health and Care Excellence endorsed this practice when EQ-5D utilities are unavailable [22]. A literature search performed on 26 October 2017 in the EMBRACE and HERC databases identified a total of 150 studies that mapped onto the EQ-5D. For detailed information on the inclusion/exclusion criteria for the literature search, see Dakin [22]. Although the two databases searched identified mostly the same studies, 18 studies found in EMBRACE were not found in HERC. This is because these studies were published after the HERC database was last updated in May of 2016. Of the 150 studies identified, 141 studies mapped onto the EQ-5D-3L, while nine studies mapped onto the EQ-5D-5L [24,95-102]. Of the nine 5L mapping studies, eight mapped from DSMs, while one mapped from other GPBMs [24]. Five of the EQ-5D-5L mapping studies applied the interim cross-walk value set [103], and three applied directly-elicited value sets. Of the latter three studies, two applied the English and Dutch value sets [98,99], while one applied the Japanese value set [96]. The source measures in these studies included DSMs related to cancer and epilepsy. A recent review by Dakin et al. [104] supports the findings of the literature search in the current thesis. While the studies that mapped onto EQ-5D-3L did include mental health measures, none of those that mapped onto EQ-5D-5L used directly-elicited value sets.

2.5 Causal and effect indicators among health-related quality of life dimensions

HRQoL measures comprise items that relate to various aspects of symptoms and functioning. Previous research has attempted to classify the items included in these measures as causal or effect indicators of HRQoL [105]. Effect indicators (also referred to as reflective indicators) can be seen as manifestations of an underlying construct, in which indicators are assumed to be drawn from an infinite pool of homogeneous indicators representing that construct, making them largely interchangeable. Thus, the causal flow is from the construct to the indicators, implying that any change in the construct will have an effect on the indicators. Conversely, causal indicators (also referred to as formative indicators) drive a change in the construct. As exemplified by Bollen and Lennox [106], life stress can be indicated by observed variables like job loss, divorce, recent bodily injury, or death in the family. These indicators are clearly causal indicators since the causal flow is from the indicators to the construct i.e. a change in life stress does not necessarily imply that a simultaneous change will occur across all causal indicators.

There is evidence to suggest that symptoms have a strong causal component that drives a change in other items [107,108]. The research into the causal nature of various HRQoL items has been limited to the cancer-specific measure, the European Organization for Research and Treatment of Cancer Quality-of-Life questionnaire (EORTC QLQ-C30), which has been investigated in three studies. Fayers and colleagues found strong evidence that physiological symptom items (e.g. nausea, memory problems, shortness of breath) were causal indicators, while items such as poor concentration, irritability, and feeling tense were likely to be effect indicators [107]. Boehmer and Luszczynska [108] identified both causal indicators (symptoms like fatigue and pain) and effect indicators (e.g. physical, role, cognitive, social, and emotional functioning). They suggested that physical functioning and pain might be

intermediate indicators. Using eight EORTC QLQ-C30 items, Bollen et al. [109] concluded that symptom items (e.g. shortness of breath, problems sleeping, lack of appetite) should be treated as causal indicators, while global health status and QoL should be treated as effect indicators. So far, no studies have investigated the classification of causal and effect indicators among GPBMs.

2.6 Objectives

The general objective of this thesis was to provide a better understanding and knowledge of GPBMs commonly applied in economic evaluations. The research questions addressed in the three papers included in this thesis are:

Paper 1: To investigate the degree of non-linear relationships across the four most widely used GPBMs (EQ-5D-5L, SF-6D, HUI-3, and 15D). We also provided exchange rates (coefficients) between GPBMs that differ depending on which intervals of the scales are considered.

Paper 2 had several aims: First, to replace existing mapping algorithms between the depression-specific measures DASS-21 and K10, and the EQ-5D-5L, which were developed using an interim EQ-5D-5L cross-walk value set based on the EQ-5D-3L value set for the UK. Second, to investigate if the mapping algorithms differed across different, directly elicited, country-specific health state preferences, including four Western countries (England, the Netherlands, Spain, Canada), three Asian countries (China, Japan, Korea) and one South American country (Uruguay). Third, to investigate the relative merit of six regression models.

Paper 3: To develop a conceptual framework for causal and effect relationships among the five dimensions of the EQ-5D-5L based on theoretical models of HRQoL, and to test this framework using empirical data.

3 Materials and methods

3.1 Data

This thesis is based on a unique international dataset from the Multi Instrument Comparison (MIC) project, which is the world's largest survey comparing GPBMs. The project was established in response to the growing evidence showing that different GPBMs produced different values for the same respondents and measured different constructs, although all GPBMs purport to measure the same construct: health state utility. While Richardson et al. [18] identified 392 pair-wise comparisons of GPBMs, only four studies included five GPBMs. Thus, the lack of thorough comparisons and comparative data was the principal motivation for the MIC project. The main aim was to document differences and the extent of the problem using a large database. The MIC project is the first study identified in the literature to include all six GPBMs, as well as eight DSMs and three subjective well-being measures. The MIC project is also unique in that it includes respondents from six countries (i.e. Australia, Canada, Germany, Norway, the UK, and the US), comprising a total of seven disease groups (i.e. asthma, arthritis, cancer, depression, diabetes, hearing loss, and heart disease) and an undiagnosed healthy group. All respondents reported their health on all GPBMs and subjective well-being measures, while only respondents in each disease group reported their health on the DSM for that particular group. This allowed comparisons with the most widely used DSMs in the different chronic disease areas, as well as with well-being measures. The selection of DSMs was based on reviews of the literature and advice from researchers from the different areas [110].

A global survey company, CINT Pty Ltd, invited individuals registered in their database to participate in an online survey [110]. Respondents were initially asked to rate their overall health on a VAS of 0-100, where 0 represented the least desirable health you could imagine

Table 3. Respondents by disease group and country

Diseases	Australia	UK	USA	Canada	Norway	Germany	Total
Asthma	141	150	150	138	129	147	855
Cancer	154	137	148	138	80	115	772
Depression	146	158	168	145	140	160	917
Diabetes	168	161	168	144	143	140	924
Hearing loss	155	126	156	144	113	136	830
Arthritis	163	159	179	139	130	159	929
Heart diseases	149	167	170	154	151	152	943
Healthy group	265	298	321	328	288	260	1760
Total	1341	1356	1460	1330	1174	1269	**7933**

and 100 represented the best possible health (physical, mental, and social), and to indicate if they had any chronic diseases. Respondents were placed in the non-diagnosed healthy group if they reported no chronic disease and an overall health rating of at least 70 on the VAS. In each country, quotas were used to provide a demographically representative sample according to age, sex, and education. For each of the seven disease groups, a quota of 150 respondents was sought. To ensure the quality of the data, a series of editing criteria were used to eliminate unreliable respondents, e.g. those who completed the survey in less than 20 minutes (median was 40) and inconsistency in response to duplicated questions. Based on the eight edit criteria provided to eliminate unreliable answers, a total of 17% of respondents were excluded. Eventually, a total of 7933 respondents were included in the dataset. For further details on respondent recruitment, see Richardson et al. [110].

In Papers 1 and 3, the full sample (N=7933) was employed, while in Paper 2 only the individuals diagnosed with depression (N=917) were included. A summary of the study sample by disease group and country is shown in Table 3.

3.2 Health outcome measures

In all papers in this thesis, the most widely used GPBM, the EQ-5D-5L, has a central role. In Papers 1 and 2, the EQ-5D-5L utility index was applied, while in Paper 3 the focus is on the 5 dimensions of the EQ-5D-5L descriptive system.

GPBMs

In Paper 1, the EQ-5D-5L, SF-6D (derived from SF-36), HUI-3, and 15D were applied (see Table 1). The EQ-5D-5L utility index was calculated using the new English value set based on a representative sample of the English public (N=996) [67]. For the SF-6D utility index, a UK value set based on a representative sample of members of the UK general population was used (N=836) [55]. The HUI-3 utility index was calculated using a representative sample of adult Canadians (N=504) [57], and the 15D utility index used a value set based on five random samples of the Finnish general population (N=2500) [111]. An overview of valuation techniques is presented in Table 2. In Paper 2, in addition to the new English value set, other, directly-elicited, country-specific EQ-5D-5L value sets were applied, including the Netherlands, Spain, Canada, China, Japan, Korea, and Uruguay [67,83,112-117].

Disease-specific measures

The DASS-21 comprises 21 items, each with a 4-point severity scale (did not apply to me; applied to some degree; applied to a considerable degree; applied very much or most of the time) [118]. It comprises three 7-items subscales that measure core symptoms of depression, anxiety, and stress. Subscale scores range from 0 to 42, where lower values indicate fewer problems.

The K10 measures psychological distress and comprises 10 items on anxiety and depressive symptoms experienced in the last 4 weeks [119]. Each item has five response levels (all the time; a little of the time; some of the time; most of the time; all of the time), resulting in a total score range of 10 to 50, where lower values indicate fewer problems.

3.3 Analysis

3.3.1 Comparing GPBMs

Paper 1 examined non-linearity among four GPBMs (EQ-5D-5L, SF-6D, HUI-3, and 15D) across different severity levels using QRM. It also investigated the exchange rates for these GPBMs.

Testing non-linearity

QRMs were used to study the relationship between pairs of GPBMs. The strength of this approach is that it permits us to explore the entire conditional distribution by analyzing the effects of one GPBM (the source) at different levels of another GPBM (the target) [120]. Thus, unlike ordinary least squares (OLS) regression, which focuses on the conditional mean of the dependent variable, the QRM tests if the relationship between two GPBMs varies at different quantiles of the dependent variable. (For a theoretical background on QRM, see Koenker and Hallock [121]). Furthermore, in comparison to ordinary linear regression, the QRM is more robust to outliers and is semi-parametric, avoiding the assumptions about the parametric distribution of the error terms [122]. Thus, following Koenker and Bassett [123], the QRM can be expressed as:

$$Y_i = \beta_0^{(q)} + \beta_1^{(q)} X_i + \varepsilon_i \qquad (1.1)$$

where Y_i is an outcome variable (target instrument), X_i is the independent variable (source instrument), $\beta^{(q)}$ is the vector of parameters to be estimated for each quantile (q) under consideration, ε_i is error term, and $0<q<1$ indicates the proportion of the population with scores below the quantile specified. Formulation of QRM requires that the q^{th} quantile of the error term be zero; and hence $Quant^{(q)}(Y_i | X_i = \beta^{(q)} X_i)$. Thus, the quantile regression estimator for the q^{th} quantile, $0<q<1$, minimizes the objective function:

$$\min_{\beta \in \Re} \left[\sum_{i:Y_i > \beta X_i} q \,|\, Y_i - \beta^{(q)} X_i \,| + \sum_{i:Y_i < \beta X_i} (1-q) \,|\, Yi - \beta^{(q)} X_i \,| \right], \qquad (1.2)$$

where variables and parameters were defined as in Equation (1). The residuals are measured using a weighted sum of vertical distances (without squaring), where the weight is *1 – q* for points below the fitted line and *q* for points above the line. The ability to estimate parameters appropriate for the chosen quantiles other than the median is a unique feature of QRMs. In this thesis, a simultaneous QRM was applied to estimate the effect of the independent/source variable at nine different quantiles of the outcome/target variable; that is, the 10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th} (median), 60^{th}, 70^{th}, 80^{th}, and 90^{th} percentile. This allows us to test if the association between two instruments differs across severity levels.

Wald F-statistics were used to test for equality of coefficients across the quantile regression results. The degree of non-linearity between GPBMs was calculated by dividing the highest coefficient in each estimation by the lowest coefficient, referred to as the maximum degree of differences in coefficients (MDDC). To inquire into variations in the degree of non-linearity across disease groups, F-tests and MDDC were presented for each of the seven disease groups.

Exchange rates

The exchange rates (ERs) presented in this paper differ from those used in traditional mapping algorithms, which are usually derived from regression techniques (e.g. OLS). Instead, our ERs were based on individual-level data and a simple calculation relying on aggregate data where the utility scale of each GPBM was split into six utility intervals with 0.2 successive decrements in utility starting from perfect health at 1.00. That is, <0.2, [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1), 1. Scale-dependent ERs were developed, and defined as the change in utilities measured using GPBM *i* (ΔU_i) divided by the change in utilities measured using GPBM *j* (ΔU_j):

$$ER_{ij} = \Delta U_i / \Delta U_j \qquad (1.3)$$

This enabled comparisons at different intervals across the health state utility scale, which may give a truer picture of the real change in utility when transforming utility gains across instruments, especially changes at the more extreme parts of the utility scale.

A bootstrap method was chosen to calculate the 95% confidence intervals for each ER. The method randomly draws a 60% sample from the full sample to calculate the ER and replicates the procedure 1000 times.

3.3.2 Predicting EQ-5D-5L utilities

In Paper 2, the mapping of depression scales (measured by DASS-21 and K10) onto eight country-specific EQ-5D-5L value sets was examined using six different regression models. Mapping is feasible only when there is a conceptual overlap between the source and target instruments.

Conceptual overlap

Initially, the dimensional structure of the EQ-5D-5L, DASS-21, and K10 were investigated using Spearman's rank correlation and exploratory factor analysis to provide insights into the conceptual overlap of the target and source measures. This could also inform the choice of the method applied for mapping, i.e. between a direct or an indirect (response) mapping approach [124]. Results of exploratory factor analysis revealed that all items on the depression scales were loaded onto the 'depression', 'anxiety', or 'stress' factors. However, the only EQ-5D-5L dimension that overlaps in any substantial way with these factors was the anxiety/depression dimension that loaded on the depression factor. Thus, the probability of accurately predicting five response levels for all the dimensions of the EQ-5D was low. In general, response mapping depends on the correct predictions for each dimension of the EQ-5D to make an exact prediction of a health state. Consequently, response mapping can be severely penalized when an incorrect prediction is made [125]. Thus, a direct mapping technique was applied to predict the EQ-5D-5L utility index (the dependent variable) using the source instrument, either the DASS-21 subscale score or the K10 total score (the independent variable). In addition to the source instrument, age and sex were considered as covariates.

Regression models

Six alternative models were compared, including OLS, the generalized linear model, the MM-estimator, the censored least absolute deviations model, the FRM, and the BB regression model. For each model, a forward stepwise selection method was used for variable selection ($p<0.05$). Interaction and squared terms were considered only if the original variable was significant. FRM and BB regression provided optimal mapping functions for predicting EQ-5D-5L utility values from both the DASS-21 and K10. Therefore, a detailed description of these models is presented in the next section.

FRM involves a semi-parametric approach that was developed to address the modeling of empirically-bound dependent variables, such as proportions and percentages, that exhibit piling-up at one of the two corners [126]. The advantages of FRM are several: (a) it does not require any special correction of the values observed at the bounds, (b) it accounts for the non-linearity in the data, and (c) it allows for direct recovery of the regression function for the dependent variable given the set of predictors[1]. Following Papke and Wooldridge [126], the basic assumption underlying the FRM can be summarized as:

$$E(Y \mid X) = G(X\beta) \tag{2.1}$$

where G(·) is a known non-linear function satisfying $0 \leq G(\cdot) \leq 1$, X is a vector of independent variables, and β is a vector of parameters to be estimated. This is well defined if Y_i takes any value in the specified range including 0 and 1 with positive probability. Unlike other parametric methods, the important advantage of this semi-parametric FRM is that it does not make any distributional assumption about an underlying structure used to obtain Y_i. Several examples of non-linear functional forms are used for G (·). The logistic link function is the most widely applied functional form and is a natural choice for modelling bounded data, since it ensures that $0 < E(Y|X) < 1$. It must be directly estimated using a non-linear function instead of being first linearized [127]. It is defined as follows:

$$E(Y \mid X) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{2.2}$$

[1] In the FRM model, EQ-5D-5L utility values are linearly transformed onto a 0-1 scale by subtracting the minimum value from observed utilities of EQ-5D-5L and then dividing by the range.

The non-linear parameters of the model defined by Equation (2.1) may be estimated using a quasi-maximum likelihood method via the maximization of the Bernoulli log-likelihood function:

$$LL_i(\beta) = Y_i \log\left[G(X_i\beta)\right] + (1 - Y_i)\log\left[1 - G(X_i\beta)\right] \qquad (2.3)$$

which is well defined for 0<G(.)<1. The quasi-maximum likelihood estimator of β is consistent and asymptotically normal, regardless of the true distribution of the dependent variable, conditional on the predictors, provided that Equation 2.1 is correctly specified, which can actually be tested using the RESET test [127].

BB regression is a similar method for modelling bounded data. It involves a fully parametric approach that allows the dependent variable to be skewed and is capable of modeling bounded dependent variables restricted between 0 and 1. As this parametric model is not defined at the boundary values, the outcome values should be restricted to a 0-1 range, excluding 0 and 1. This can be achieved by linear transformation [Y(N-1)+0.5]/N following earlier literature [128,129], where N refers to sample size, and Y is the dependent variable. BB regression is flexible and allows a great variety of asymmetric forms. The most popular choice to estimate Equation 2.1 is the parametric beta distribution, see Khan and Morris [130] for applications of the beta binomial regression model. However, Khan and Morris used an inflated BB when mapping onto the EQ-5D, due to piling of responses at 1. This was not an issue in the MIC project depression sample (less than 2% reported the utility of 1). Following re-parameterization of Ferrari and Cribari-Neto [131], beta regression is a fully parametric approach, assuming that the dependent variable follows a beta distribution with density function:

$$f(Y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(1-\mu\phi)}Y^{\mu\phi-1}(1-Y)^{(1-\mu)\phi-1} \quad (2.4)$$

where r(.) denotes the gamma function 0<Y<1, and the parameter μ denotes the expected value of Y; that is, E(*Y*)=μ; and φ dispersion parameter. The parameter φ can be interpreted as a precision parameter, because for fixed μ, the greater the value of φ, the smaller the variance of the dependent variable (for detail see Ferrari and Cribari-Neto [131]). That is,

$$Var(Y)=\frac{Var(\mu)}{\phi+1}=\frac{\mu(1-\mu)}{\phi+1} \quad (2.5)$$

Again, using logit as a link function (Equation 2.2), the BB regression for Y_i (EQ-5D-5L) with X (DASS-21 subscales or K10 total scale) as independent predictor(s), and β as a vector of parameters is given as:

$$\mu_i=\frac{\exp(X\beta)}{1+\exp(X\beta)} \quad (2.6)$$

It is also possible to model the dispersion in terms of the independent predictors instead of assuming it is a nuisance parameter. However, since modeling the precision parameter separately barely adds to the prediction performance of the model, we considered Equation 2.1 and 2.6 alone, assuming φ is a nuisance parameter for the sake of parsimony.

Model performance

To assess the predictive performance of each model in Paper 2, common criteria in mapping studies were applied: mean absolute error (MAE) and root mean square error (RMSE)

[20,21]. Due to a different number of independent variables across models, the degree of freedom was adjusted for both MAE and RMSE. Since a wider scale length of the dependent variable produces larger error [132], adjustment for scale differences was performed to allow reasonable comparisons between datasets or models with different scales. Both MAE and RMSE were normalized to the range (defined as the difference between the maximum and the minimum values) of the measured data. For instance, when applying the English value set, the MAE and RMSE were divided by the range of the data (1.17) as estimated by OLS, but they were divided by 0.998 when estimated by BB regression (since BB requires the transformation of values to be restricted between 0 and 1). In comparison, when estimated by OLS and applying the Dutch and Japanese value sets, MAE and RMSE were divided by 1.41 and 0.88, respectively. Lastly, the performance of each model was also assessed by the square of the correlation coefficient between the observed and predicted values adjusted for the number of predictors in the model (adj. r^2) [133].

To investigate the generalizability of the preferred mapping algorithms, cross-validation was performed by splitting the existing data in two: estimation and validation samples via random selection procedures. In this study, the total sample was randomly divided into two equally-sized groups to evaluate the model fit in out-of-sample data. The model was fitted on the estimation sample, and the resulting parameters from the fitted model were then used to predict the EQ-5D-5L in the validation sample. This procedure was then repeated by reversing the validation and estimation sample. The average MAE, RMSE, and r^2 for both iterations were calculated to compare of the models' predictive performance. Lastly, the best-fitting model was estimated using the full sample (N=917).

3.3.3 Testing the relationship between EQ-5D-5L dimensions

In Paper 3, the causal relationship between EQ-5D-5L dimensions was explored by specifying a number of testable models that specified EQ-5D-5L items as causal indicators or effect indicators of HRQoL. Model 1 specified all five EQ-5D-5L items as effect indicators of the construct HRQoL, whereas Models 2, 3, and 4 are multiple-cause multiple-indicator models. Model 2 tested whether the symptom items pain/discomfort and anxiety/depression should be treated as causal indicators and whether the activity/participation items mobility, self-care, and usual activities should be treated as effect indicators. On the other hand, Model 3 treated the symptom items pain/discomfort, anxiety/depression, and mobility as causal indicators, and self-care and usual activities as effect indicators. Model 4 was a variant of Model 3, where mobility had an intermediate position between pain/discomfort and the construct. See Paper 3 in appendix for illustrations of the models. In addition, due to the uncertain nature of anxiety/depression and the investigation of reverse causality, alternative models were specified (see Paper 3 in appendix for description of these models).

Two model-testing approaches were selected to test the specified models: confirmatory tetrad analysis (CTA) and confirmatory factor analysis (CFA). However, due to the specification of mobility in Model 4, CTA is not appropriate; thus, Model 4 was tested by CFA only. CTA seeks to determine whether items of a latent variable should be treated as causal or effect indicators [134,135]. Unlike the general SEM, the CTA does not estimate parameters; it only tests model fit using Chi-square (χ^2). Furthermore, differently from traditionally nested models, a nested CTA can determine if two models are nested in terms of model-implied vanishing tetrads, not parameters. A bootstrap tetrad test was used to minimize the problem of non-normality [136]. For a more detailed description of CTA, see Paper 3 in appendix. Considering CFA, maximum likelihood estimation is considered robust when using non-

continuous data [137-139] or data that violate multivariate normality assumptions [140-142]. However, since maximum likelihood can be affected by deviation from normality [143], bootstrap standard errors (with 1000 bootstrap draws) were used [144].

Several fit indices were used to examine the model fit to data, including the comparative fit index, the Tucker-Lewis index, root-mean square error of approximation, the standardized root-mean square residual, the Akaike information criterion, and the sample-size adjusted Bayesian information criterion. A comparative fit index and Tucker-Lewis index value greater than 0.95, and a standardized root-mean square residual value less than 0.08, represent a well-fitting model [145]. Root-mean square error of approximation values less than 0.05 reflect a good fit [146], and values as high as 0.08 reflect an adequate fit [147]. Akaike information criterion and sample-size adjusted Bayesian information criterion are only meaningful when different models are compared, and models with the lowest values represent those with the best fit.

4 Results

4.1 Paper 1: Non-linearity across generic preference-based measures

The results of Paper 1 revealed differences across the EQ-5D-5L, SF-6D, HUI-3, and 15D. Ceiling effects (utility=1) ranged from 1.4% (SF-6D) to 19.3% (EQ-5D-5L). The EQ-5D-5L and HUI-3 scales allow for utilities below zero, and have larger proportions at the bottom end of the scales. Because of the different scale lengths, the differences between health state utilities varied depending on the range compared, with the largest differences observed at the 10^{th} percentile, which implies that the potential utility gain involved from a cure would differ a lot depending on the GPBM used. Mean utilities varied from 0.71 (SF-6D and HUI-3) to 0.85 (15D), while median utilities ranged between 0.70 (SF-6D) and 0.88 (15D). Mean and median utility for EQ-5D is 0.83 and 0.87, respectively

The key findings from Paper 1 were that non-linear relationships were evident across all four GPBMs. QRMs revealed that coefficients varied across all quantiles, implying that the strength of the relationship across GPBMs varies across the distribution of the target measure. For instance, when regressing any of the other GPBMs onto EQ-5D-5L, the effect is largest at the 0.1 quantile, then decreases, with the lowest effect at 0.9 quantile. Considering 15D as the source measure, the coefficient at the lower quantile of EQ-5D-5L was more than double (1.8, $p<0.001$) that of the upper quantile (0.7, $p<0.001$). The different effect of all GPBMs across the EQ-5D-5L distribution was supported by Wald F-tests, indicating a significant difference ($p<0.01$) between coefficients, which rejected the null hypothesis of the equality of coefficients across quantiles. This was also shown when regressing onto the other three GPBMs. For HUI-3 and 15D, a similar pattern of the coefficients across quantiles was observed, while for SF-6D the strongest effect was on the 0.6 quantile (15D) and 0.7 quantile (HUI-3 and EQ-5D-5L) and lowest at the 0.1 quantile. Except when regressing HUI-3 and 15D onto SF-6D in the depression group, all tests indicated that there was a significant difference ($p<0.01$) between coefficients.

The degree of non-linearity across the full study sample and in all seven disease groups was investigated by calculating the ratio between the highest and the lowest coefficients, referred to as the MDDC at each estimation. For instance, when regressing SF-6D onto EQ-5D, the highest coefficient was at the 0.1 quantile and the lowest at the 0.9 quantile (MDDC: 1.37/0.41=3.34). The degree of non-linearity was largest for EQ-5D-5L (MDDC: 2.49-3.34) and HUI-3 (MDDC: 2.05-3.48) and smallest for SF-6D (MDDC: 1.26-1.32). Similar results were found when looking into the seven disease groups.

The scale-dependent exchange rates further revealed the non-linear relationships. The ERs were calculated based on a change in mean utilities of the target measure at different utility intervals of that measure. For instance, the ERs between each utility interval on the 15D scale

indicate the value by which a 15D increment has to be multiplied to get the corresponding change in utility had the EQ-5D been applied. As an example, take the ERs between 15D and EQ-5D-5L. If 15D had been applied in a study where the patient group at baseline was in the 0.4-0.6 interval with a mean utility of 0.53, and after treatment they were in the 0.6-0.8 interval with a mean utility of 0.72, it would represent an increase of ΔU_{15D}=0.19 on the 15D utility scale. The corresponding increment on the EQ-5D-5L scale would be from 0.38 to 0.67 ($\Delta U_{EQ\text{-}5D}$ =0.29). Hence, the 15D increment has to be multiplied by an exchange rate of 1.51 ($\Delta U_{EQ\text{-}5D}$ / ΔU_{15D}) to make the utility increment comparable to the increment had the EQ-5D-5L been applied.

4.2 Paper 2: Mapping from disease-specific to generic measures

The results of Paper 2 showed that both mean EQ-5D-5L utility values and the range of these values varied depending on the choice of country-specific value sets. The mean EQ-5D-5L utility ranged from 0.59 in the Dutch value set to 0.83 in the Uruguayan value set, while the minimum utility value ranged from -0.41 in the Dutch value set to 0.12 in the Korean and Uruguayan value set. Such differences across value sets suggest that a separate mapping algorithm needs to be estimated.

Spearman's rank correlation indicated that the EQ-5D-5L anxiety/depression dimension produced the highest correlation with the source measures ($r \geq 0.50$), while the mobility dimension produced the lowest ($r \leq 0.25$). In addition, the EQ-5D-5L usual activities dimension correlated moderately (r=0.35 to 0.37) with DASS-anxiety, DASS-depression, and the K10 scale. Exploratory factor analysis revealed that only the EQ-5D-5L anxiety/depression dimension overlapped with the depression dimensions extracted for both the DASS-21 and the K10. The remaining four EQ-5D-5L items were mainly loaded on the fourth factor (i.e. physical functioning), with no high cross-loadings (>0.30).

In the evaluation of model performance, FRM performed best in the majority of goodness-of-fit measures, i.e. adjusted-r^2, normalized MAE and normalized RMSE, for both DASS-21 and K10. The only exception was for the Japanese value set, for which BB regression was the preferred model. Although median estimators (censored least absolute deviations and MM-estimator) generally performed best in terms of normalized MAE, they performed poorly on the other two measures.

When DASS-21 was the source measure, the best fitting regression results included depression subscale scores, anxiety subscale scores, and age as significant ($p<0.05$) predictors in all models. When K10 was the source measure, K10 total score and age were significant ($p<0.05$) predictors.

4.3 Paper 3: Causal links across health-related quality of life dimensions

Paper 3 explored the causal pattern among the five dimensions of EQ-5D-5L. Frequency distribution of EQ-5D-5L health states revealed that the three most prevalent health states (i.e. 11121, 11112, 11122) accounted for more than one third of the sample that was in a non-11111 health state. When including other health states with decrements in symptom dimensions only, almost half of the sample (i.e. 47%) was accounted for. In contrast, our findings revealed that decrements in activity/participation dimensions are rarely reported without any decrements in symptoms (i.e. 1.5% of the sample). Additionally, the relationship between summary scores of symptom dimensions and summary scores of activity/participation dimensions (see Figure 3 in the appendix of Paper 3 for illustration of the result), indicated that increasing symptoms are associated with more problems in activity/participation dimensions. However, it appears that problems with activity/participation lag behind problems in symptom dimensions.

The result of the two SEM approaches applied (CTA and CFA) indicated support for Model 3, which specifies mobility, pain/discomfort, and anxiety/depression as causal indicators, and self-care and usual activities as effect indicators (see Figure 4 in the appendix of Paper 3).

The results of the CTA for Model 1 (χ^2=1500.00, degree of freedom, df=15), Model 2 (χ^2=893.79, df=6), and Model 3 (χ^2=105.84, df=3) revealed highly significant χ^2 estimates (P<0.0001). Model 3 clearly produced the lowest χ^2 estimates, suggesting it was the best model. Although the significant χ^2 estimate indicated that the model was a poor fit to the data, it is common for χ^2 estimates to be significant in large samples [148]. A nested CTA test that compared Model 2 and Model 3 revealed a highly significant χ^2 difference (χ^2 diff=787.62, df=6, p<0.0001), favoring for the model with fewest vanishing tetrads (Model 3).

When looking at the results of CFA, the fit indices suggested a similar satisfactory fit to the data. For Model 3 and Model 4, the comparative fit index and Tucker-Lewis index were greater than 0.95, the root-mean square error of approximation was lower than 0.08, and the standardized root-mean square residual was 0.12 and 0.16, respectively. However, the Akaike information criterion and sample-size adjusted Bayesian information criterion for Model 3 (Akaike information criterion=20400.537; Bayesian information criterion=20434.746) was preferred when compared to Model 4 (Akaike information criterion=36580.861; Bayesian information criterion=36626.472).

An alternative model specifying the anxiety/depression dimension as an effect indicator along with self-care and usual activities did not produce a good fit, either with CTA (χ^2=927.93, df=6, p<0.0001) or CFA (comparative fit index=0.965; Tucker-Lewis index=0.922; root-mean square error of approximation=0.122; standardized root-mean square residual=0.026). Further models investigated other specifications of the interrelationships among the three

causal indicators (mobility, pain/discomfort, and anxiety/depression) in Model 4. All these models had a poor fit compared to the chosen model.

5 Discussion

5.1 Methodological issues

5.1.1 Study design

The MIC project is the largest international project designed to make comparisons and transformations across GPBMs, as well as commonly used DSMs [110]. There was a growing amount of literature on the dissimilarities and lack of agreement across the existing methods for measuring and valuing HRQoL. Thus, the motivation for the MIC project was to assess this discrepancy, and the project generally aims to contribute to the methodological development of the field of measuring HRQoL.

The MIC project has a cross-sectional design and includes respondents from six countries with a diverse range of health states that are associated with major chronic diseases, as well as a non-diagnosed healthy group. The aim of this thesis overlaps with that of the MIC project, i.e. to compare and transform measures (making the MIC data appropriate for the research questions posed in this thesis).

All the three papers in this thesis are exclusively based on data from the MIC project. While Papers 1 and 3 included the full study sample, which included seven disease groups and a non-diagnosed healthy group, Paper 2 included the depression subsample only. Mean age in the full sample was 51.5, and sex was more equally distributed (female=52.2%). Moreover, respondents in the depression subsample were younger (mean age is 42) and this group consisted of more women (65.9%). Previous studies have also shown that the prevalence of depression is commonly higher among women and younger individuals [25,149].

The MIC project survey was administered by an online survey company, CINT, that invited individuals registered in their database to participate. Initially, respondents were asked about diagnoses they had received regarding any of the seven diseases, and based on their reply they were placed in either the healthy group or the disease group. A target number of respondents were sought for each of the seven disease groups. Respondents would then proceed to answer the core questionnaire containing GPBMs and DSM(s) until the quota was reached for the group to which they were assigned. Due to this recruitment procedure, the issue of "response rate" becomes less relevant. However, 83% of responses were retained after the stringent criteria used in the editing process were applied to remove unreliable answers, which can be considered quite high.

Self-selection may be an issue, since participants choose to register in the CINT online survey database and subsequently the MIC project survey. Nevertheless, the high retention of responses after the stringent editing procedure, as well as the large sample size, lends support to the high quality of the data. Furthermore, there is no missing information on any of the variables used in this thesis. This is due, in part, to the fact that all participants were required to respond to each item when responding to any of the HRQoL measures.

5.1.2 Reliability and validity of HRQoL measures

The four GPBMs (EQ-5D, SF-6D, 15D, HUI-3) and the two DSMs (DASS-21 and K10) employed in this thesis have broad and extensive applications. The four GPBMs were selected because they were used in more than 98% of studies that measured and valued QALY gains [19], while the DASS-21 and K10 are among the most widely used mental health scales applied to identify emotional disturbance by assessing core symptoms of depression, anxiety, and stress. All measures have previously been validated in different studies, in different countries, and in different settings.

Reliability

The generic and disease-specific measures applied in this study have shown satisfactory reliability in previous studies, as well as in this thesis. Reliability refers to the overall consistency of a measure [1]. Three types of consistency should be considered: test-retest reliability, i.e. if the measure produces similar results under the same conditions over time; inter-rater reliability, i.e. if the measure produces similar results across different observers assessing the same person; and finally internal consistency, i.e. if the measure produces similar results across items of the measure. While previous studies have shown evidence of test-retest and inter-rater reliability of GPBMs [150,151], the data from the MIC project did not include measurement at different time points nor from different observers' perspectives. However, internal consistency, commonly measured by Cronbach's alpha, indicated the consistency of individuals' responses across the items of a measure. In Table 4, except for HUI-3, all measures had a Cronbach's alpha coefficient above 0.80, which is indicative of good internal consistency [152]. The lower alpha coefficient produced by HUI-3 may be due to the definition of health that is used in its development. Indeed, the HUI-3 does not focus on "beyond the skin aspects of health" and mainly assesses impairments of body functions and disability. This may lead to lower inter-item correlations, and consequently a lower Cronbach alpha [57].

Validity

A valid measurement scale measures what it is intended to measure, which cannot be established by high reliability alone [152]. Important aspects of validity are content and construct validity. Content validity is the extent to which a measure covers the health dimensions of interest and is sufficiently sensitive to change [152]. Convergent and discriminant validity are the two subtypes of validity that make up construct validity.

Table 4. Reliability and validity of HRQoL measures

	SF-36 subscales*										
									Cronbach's	Kruskal-Wallis[a]	
Measures	MH	RE	SF	VT	BP	RP	PF	GH	α	χ^2	*p-level*
EQ-5D-5L	0.57	0.48	0.65	0.62	0.71	0.55	0.70	0.63	0.81	68.209	0.0001
SF-6D[b]	0.71	0.70	0.80	0.76	0.73	0.71	0.67	0.69	0.84	47.696	0.0001
HUI-3	0.60	0.48	0.64	0.63	0.66	0.54	0.67	0.63	0.69	119.515	0.0001
15D	0.62	0.53	0.68	0.72	0.69	0.61	0.71	0.73	0.88	87.256	0.0001
DASS-D	0.73	0.40	0.55	0.55	0.24	0.27	0.27	0.41	0.92	14.767	0.0006
DASS-A	0.58	0.33	0.46	0.37	0.33	0.33	0.35	0.41	0.84	22.863	0.0001
DASS-S	0.60	0.33	0.42	0.40	0.24	0.25	0.23	0.35	0.86	12.979	0.0015
K10	0.81	0.45	0.61	0.59	0.33	0.36	0.35	0.47	0.92	18.366	0.0001

Note. Full sample applied above midline (N=7933); Depression subsample applied below midline (N=917). MH=mental health; RE=role emotional; SF=social function; VT=vitality; BP=bodily pain; RP=role physical; PF=physical functioning; GH=general health.
[a]Kruskal-Wallis H test statistics to test known-group validity across education levels (1=High school; 2=Diploma;3=University).
[b]Among the four GPBMs, the highest correlation coefficients for SF-6D should be expected, given it is a "short-form" of the full SF-36.
*All Pearson correlation coefficients are significant at $p<0.001$.

Convergent validity refers to the degree to which a measure correlates positively with a theoretically similar measure, while discriminant validity is the extent to which a measure does not correlate with scores on measures that are conceptually distinct [152].

In this thesis, content and convergent validity were assessed by investigating how the GPBMs and DSMs correlated with the eight health dimensions of the SF-36. The GPBMs produced moderate to large correlations across SF-36 subscales, while for DSMs, large correlations were produced with the SF-36 mental health dimension, and weaker correlations with other SF-36 dimensions e.g. physical functioning and bodily pain. The weaker correlations for the DSMs with the physical health dimensions of the SF-36 are indicative of discriminant validity. As seen in the last column of Table 4, discriminant validity was also indicated by the ability of the measures to discriminate between known groups (education levels).

Some studies have also revealed the validity of the GPBMs (for instance see Finch et al. [153] and Richardson et al. [17]). Although there is no gold standard that HRQoL measures can be compared with, the validity of measures applied in this thesis was acceptable, which was supported by evidence in the literature.

5.2 Discussion of results

The major aim of this thesis was to provide a better understanding of GPBMs used in economic evaluations to assign health state utilities, with particular emphasis on the most widely used measure, the EQ-5D. Paper 1 focused on the comparability across four GPBMs: the EQ-5D-5L, SF-6D, HUI-3, and 15D. Quantile regressions revealed that the EQ-5D-5L has strong non-linear relationships with all the other GPBMs, and this finding was consistent across the seven disease groups. For instance, the MDDC (i.e. the ratio between the highest and the lowest coefficients across quantiles) ranged from 2.05 to 3.39. This indicates that across the full sample and subsamples, the coefficient estimated at the 0.1 quantile was at least twice the size of the coefficient estimated at the 0.9 quantile. These findings support other studies that have indicated non-linear relationships across GPBMs [23,24,92]. The scaling effect could be a key factor in explaining the observed non-linear relationship [16]. For instance, a recent study by Whitehurst et al. [93] compared EQ-5D and SF-6D responses from seven patient datasets using published DCE-derived scoring algorithms against previous conventional EQ-5D-3L and SF-6D index scores, and found that SF-6D produced consistently lower values for severe health states, which is contrary to previous findings. This suggests that the 'floor effect', or the worst health states, in SF-6D compared to EQ-5D-3L can be explained by the technique used for valuation, while the remaining incommensurability is due to differences in the descriptive system [1]. Furthermore, we

confirmed earlier findings from Seymour et al. [92] that the effect of the EQ-5D differs at different parts of the SF-6D distribution. While the degree of non-linearity in the current paper was less for the SF-6D than for the EQ-5D-5L, the non-linearity was still replicated across all disease groups, with only one exception in the depression subsample, where linear relationships were indicated between the SF-6D and both the 15D and HUI-3. This implies that there may be a constant exchange rate between the SF-6D and 15D as well as the HUI-3, irrespective of severity levels.

Compared to the OLS coefficient, quantile regression revealed that the consequences of applying OLS regression when estimating mapping algorithms across all GPBMs would over-predict utility for respondents with poor health and under-predict utility for respondents with moderate to good health, except for the SF-6D where this tendency is reversed. This may be because SF-6D is more sensitive among respondents with better health (only 1.4% have utility=1), as well as the lesser preference weights attached to the more severe health states as compared to others. In general, the problem of overestimating utilities in respondents with poor health when mapping across GPBMs persists [24]. Our results further strengthen the claim that non-linear associations are important to take into account when comparing healthcare programs whose effectiveness have been measured by different GPBMs.

In Paper 2, the primary aim was to develop mapping algorithms to estimate EQ-5D-5L health state utility values from two widely used depression-specific measures. The findings showed that the mapping algorithms differed across country-specific value sets, which indicated that the strength of preferences for the different health dimensions in EQ-5D-5L is culture-dependent. Thus, country-specific mapping algorithms should be applied to estimate utilities for a particular country.

An earlier study mapped the DASS-21 and K10 onto EQ-5D-5L using the interim UK cross-walk value set [102]. Although both studies applied the same data, the results were not directly comparable due to the difference in value sets, regression models, and choice of covariates. In order to better compare the studies, the MAE and RMSE used therein were normalized, as they were in the current study. This produced a normalized MAE and normalized RMSE of 0.119 and 0.159, respectively, for the DASS-21, and 0.115 and 0.154, respectively, for the K10. In the previous study, the OLS model was preferred over the generalized linear model. In the current study, the FRM produced lower estimates for both the DASS-21 and the K10, indicating better predictive performance when the English value set was applied. Furthermore, among the eight country-specific value sets in the current study, prediction accuracy was best when using the Uruguayan value set for both the DASS-21 (normalized MAE=0.099; normalized RMSE=0.138) and the K10 (normalized MAE=0.097; normalized RMSE=0.138) and worst for Canada for both the DASS-21 (normalized MAE=0.132; normalized RMSE=0.175) and the K10 (normalized MAE=0.128; normalized RMSE=0.175).

The FRM showed the best fit in all cases except Japanese values, for which BB regression model was preferred. It is not surprising that different models may fit different country-specific value sets. Mapping is data-dependent, which is one of the key reasons why different econometric methods need to be tested and then based on goodness-of-fit estimates to identify the optimal one. The possible explanation for the choice of different model for the Japanese value set could be dissimilarities in the distribution of the preference pattern compared to other country-specific value sets. A previous study compared value sets from seven countries (four Western, two Asian, and one South American) and found that the Western value sets had more or less a similar pattern [154]. This resemblance in preference pattern could explain why the Western value sets fit the same model. On the other hand, this observation may

suggest that to derive an optimal mapping algorithm, different econometric techniques should be considered.

Paper 3 focused on the causal and effect nature of EQ-5D dimensions. The aim was to develop a conceptual framework based on theoretical models suggested by the International Classification of Functioning, Disability and Health and Wilson and Clearly [155,156]. Specifically, symptom/impairments dimensions are causal indicators, while functioning/activities or participation dimensions are effect indicators.

Investigation of commonly reported EQ-5D-5L health states in the current study revealed that decrements in symptom dimensions without decrements in activity/participation dimensions are common, while the opposite is rare. This finding suggest that problems with activity/participation are prevalence dependent on symptoms. Similar patterns of commonly reported EQ-5D-5L health states has been shown by others [157]. The suggestion that symptoms precedes problems in activity/participation dimensions was also supported by our result showing that increasing summary scores of symptom dimensions are associated with increasing summary scores of activity/participation dimensions, but that problems in the latter seem to lag behind.

The result of CTA and CFA suggested that self-care and usual activities acted as effect indicators of HRQoL; and mobility, pain/discomfort, and anxiety/depression appeared to be causal in nature, driving changes in self-care and usual activities. Previous research has suggested that mobility may have an intermediate role [108], and while our results indicated that this is plausible, the model specifying mobility as causal produced a better fit.

Studies have indicated that anxiety/depression and emotional functioning are effect indicators [108,158]. However, our results suggested that the anxiety/depression dimension is a causal indicator. There are reasons to believe that the role of anxiety/depression might vary

depending on the severity of these conditions. If the depression or anxiety is moderate or severe (levels 3-5), it could reflect more of a clinical symptom that may cause dysfunctions in self-care and usual activities that typically require treatment. If the condition is mild (level 2), it could be more subjective well-being, which may vary according to personality traits (e.g. optimist versus pessimist; level of neuroticism). Further investigation into the various disease groups might have indicated that the causal nature of anxiety/depression is disease-specific.

6 Policy implications and future research

6.1 Policy implications

The findings of this thesis have important implications for researchers and decision-makers involved in the economic evaluation of healthcare interventions and in setting priorities for the allocation of healthcare resources. Paper 1 showed that the result of a health intervention depends on the GPBMs used, implying that decision-makers would have a problem comparing QALY gains calculated with different GPBMs. We observed clear non-linear relationships, implying that the same exchange rate should not be applied across all levels of the utility scale.

Paper 2 indicated that mapping the DASS-21 or the K10 onto the EQ-5D-5L is feasible when applying any of the country-specific value sets. The best-performing regression models imply that researchers should account for non-linearity of the data when performing mapping, as well as consider the model's appropriateness for bounded data. Thus, in the absence of EQ-5D-5L utility, the preferred mapping model can adequately convert depression-specific scores into utility values. Due to the high prevalence and early onset of depression, enabling economic evaluation of clinical studies is very policy-relevant [159].

The findings of Paper 3 indicated that the EQ-5D-5L comprises both causal and effect indicators of HRQoL. Knowledge about the causal relationship among dimensions is important for researchers involved in developing new, and extending existing GPBMs (e.g. bolt-ons for EQ-5D). While adding effect indicators (i.e. social relationships) may make the EQ-5D more applicable across sectors (e.g. social care), adding a symptom indicator may make it more applicable to health care. However, focusing on effect indicators and the participation part of the health spectrum may be more in line with the aim of a generic measure, since it is not specific to particular health conditions. Nevertheless, sensitive GPBMs should probably comprise both types of dimensions. Whether causal or effect indicators should be emphasized is a question that is open for further research.

6.2 Future research

In the absence of external data, existing datasets are commonly split into an estimation dataset and a validation dataset. However, mapping algorithms should ideally be validated using an external dataset that has not been used for estimation. Future studies should investigate the merit of the mapping results produced in this thesis by applying external data.

Based on the WHO's definition of health, it has been suggested that health economic evaluations should put more emphasis on the mental and social dimensions of health. An interesting question for future studies is whether a GPBM like the EQ-5D should broaden its operationalization of the HRQoL concept in the direction of effect dimensions (e.g. social connections/network or general well-being) or in the direction of causal dimensions (e.g. tiredness). Furthermore, since anxiety and depression are highly prevalent disorders and the most commonly reported comorbidities that more often occur together than alone [160,161], another solution that may improve the sensitivity of the EQ-5D-5L in mental health would be to split anxiety and depression into two distinct dimensions in the descriptive system.

7 Conclusion

In conclusion, this thesis examined the degree of non-linear relationships across four GPBMs (EQ-5D-5L, SF-6D, HUI-3, and 15D) and then focused on the most widely used utility measure, the EQ-5D-5L, by i) mapping from depression-specific measures (DASS-21 and K10) onto EQ-5D-5L utilities; and ii) exploring the internal structure of the EQ-5D-5L descriptive system.

The clear non-linear relationships observed across GPBMs make it difficult to compare QALY gains from studies that have applied different measures. Exchange rates that are scale-dependent can convert a change in utility on a given measure into a corresponding utility change on another measure, which may enable a better comparison of the cost-effectiveness of competing health interventions that have been assessed by different GPBMs. Thus, accounting for non-linear relationships will increase the validity of such comparisons.

The preferred mapping algorithm between depression-specific measures and EQ-5D-5L utility values adequately predict mean health state utility values, which facilitates economic evaluations of mental health interventions. Since different EQ-5D-5L value sets produce different utility values, especially at the lower end, the country-specific mapping algorithm is a better option to reflect the preference in a particular country. Thus, in the absence of utility data, the mapping algorithms enable the conversion of DASS-21 or K10 scores to a generic outcome metric like QALYs.

A conceptual framework was developed based on theoretical models of HRQoL, which depicted the five dimensions in the EQ-5D-5L descriptive system as causal or effect indicators of HRQoL. Empirical testing of this framework supported that the EQ-5D-5L comprises causal indicators (mobility, pain/discomfort, anxiety/depression) and effect indicators (self-care and usual activities) of HRQoL.

8 References

1. Brazier, J., Ratcliffe, J., Salamon, J., & Tsuchiya, A. (2017). *Measuring and valuing health benefits for economic evaluation* (2ed.). New York: Oxford university press.
2. Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: the QALY and utilities. *British Medical Bulletin, 96*, 5-21, doi:10.1093/bmb/ldq033.
3. NICE (2013). Guide to the methods of technology appraisal 2013. Retrived from https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781. Accessed 15.08.2017.
4. NoMA (2018). Guidelines for the submission of documentation for single technology assessment (STA) of pharmaceuticals. Retrived from https://legemiddelverket.no/Documents/English/Price%20and%20reimbursement/Application%20for%20reimbursement/Guidelines_april_2018.pdf. Accessed 24.05.2018.
5. CADTH (2017). Guidelines for the economic evaluation of health technologies: Canada. Retrived from https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf. Accessed 24.05.2018.
6. PBAC (2016). Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee. Retrived from https://pbac.pbs.gov.au/. Accessed 24.05 2018.
7. Drummond, M. F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G. W. (2015). *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.
8. Weinstein, M. C., Torrance, G., & McGuire, A. (2009). QALYs: The Basics. *Value in Health, 12*, 5-9, doi:10.1111/j.1524-4733.2009.00515.x.
9. Deshpande, P. R., Rajan, S., Sudeepthi, B. L., & Abdul Nazir, C. P. (2011). Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Research, 2*(4), 137-144, doi:10.4103/2229-3485.86879.
10. Fayers, P. M., & Machin, D. (2015). *Quality of Life : the assessment, analysis and reporting of patient-reported outcomes*. Hoboken: Wiley.
11. de Jong, M. J., Huibregtse, R., Masclee, A. A. M., Jonkers, D. M. A. E., & Pierik, M. J. (2018). Patient-reported outcome measures for use in clinical trials and clinical practice in inflammatory bowel diseases: a systematic review. *Clinical Gastroenterology and Hepatology, 16*(5), 648-663.e643, doi:https://doi.org/10.1016/j.cgh.2017.10.019.
12. Ahmed, S., Berzon, R. A., Revicki, D. A., Lenderking, W. R., Moinpour, C. M., Basch, E., et al. (2012). The use of patient-reported outcomes (PRO) within comparative effectiveness research implications for clinical practice and health care policy. *Medical Care, 50*(12), 1060-1070, doi:10.1097/MLR.0b013e318268aaff.
13. Wiklund, I. (2004). Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. *Fundamental & Clinical Pharmacology, 18*(3), 351-363, doi:10.1111/j.1472-8206.2004.00234.x.
14. Brazier, J., Ara, R., Rowen, D., & Chevrou-Severac, H. (2017). A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics, 35*(Suppl 1), 21-31, doi:10.1007/s40273-017-0545-x.
15. Torrance, G. W. (1987). Utility approach to measuring health-related quality of life. *Journal of Chronic Diseases, 40*(6), 593-600, doi:https://doi.org/10.1016/0021-9681(87)90019-1.
16. Richardson, J., Iezzi, A., & Khan, M. A. (2015). Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. *Quality of Life Research, 24*(8), 2045-2053, doi:10.1007/s11136-015-0926-6.
17. Richardson, J., Khan, M. A., Iezzi, A., & Maxwell, A. (2015). Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Medical Decision Making, 35*(3), 276-291, doi:10.1177/0272989x14543107.
18. Richardson, J., McKie, J., & Bariola, E. (2014). Multi attribute utility instruments and their use In C. AJ (Ed.), *Encyclopedia of health economics* (pp. 341-357). San Diego: Elsevier Science.

19. Wisloff, T., Hagen, G., Hamidi, V., Movik, E., Klemp, M., & Olsen, J. A. (2014). Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. *Pharmacoeconomics, 32*(4), 367-375, doi:10.1007/s40273-014-0136-z.
20. Brazier, J. E., Yang, Y., Tsuchiya, A., & Rowen, D. L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European Journal of Health Economics, 11*(2), 215-225, doi:10.1007/s10198-009-0168-z.
21. Longworth, L., & Rowen, D. (2013). Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. *Value in Health, 16*(1), 202-210, doi:https://doi.org/10.1016/j.jval.2012.10.010.
22. Dakin, H. (2013). Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health and Quality of Life Outcomes, 11*, 151, doi:10.1186/1477-7525-11-151.
23. Fryback, D. G., Palta, M., Cherepanov, D., Bolt, D., & Kim, J. S. (2010). Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Medical Decision Making, 30*(1), 5-15, doi:10.1177/0272989x09347016.
24. Chen, G., Khan, M. A., Iezzi, A., Ratcliffe, J., & Richardson, J. (2016). Mapping between 6 multiattribute utility instruments. *Medical Decision Making, 36*(2), 160-175.
25. WHO (2017). Depression and other common mental disorders: global health estimates. Retrived from http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;jsessionid=694BFC0014BC54D45E1D54BBFD0A9A4F?sequence=1. Accessed 27.04.2018.
26. WHO (2017). Mental disorders. http://www.who.int/mediacentre/factsheets/fs396/en/. Accessed 27.04 2017.
27. Barofsky, I. (2012). Can quality or quality-of-life be defined? *Quality of Life Research, 21*(4), 625-631, doi:10.1007/s11136-011-9961-0.
28. Karimi, M., & Brazier, J. (2016). Health, health-related quality of life, and quality of life: what is the difference? *Pharmacoeconomics, 34*(7), 645-649, doi:10.1007/s40273-016-0389-9.
29. WHO (2006). Constitution of the World health organization. Retrived from http://www.who.int/governance/eb/who_constitution_en.pdf. Accessed 27.04.2018.
30. Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Toward an operational definition of health. *Journal of Health and Social Behavior, 14*(1), 6-23, doi:10.2307/2136932.
31. WHO (1986). Ottawa charter for health promotion. Ottawa, Ontario, Canada.
32. Tinetti, M. E., & Fried, T. (2004). The end of the disease era. *The American journal of medicine, 116*(3), 179-185.
33. Patrick, D. L., & Erickson, P. (1993). Assessing health-related quality of life for clinical decision-making. In S. R. Walker, & R. M. Rosser (Eds.), *Quality of Life Assessment: Key Issues in the 1990s* (pp. 11-63). Dordrecht: Springer Netherlands.
34. WHO (2016). Introducing Whoqol: Measuring Quality of Life. http://www.whoint/healthinfo/survey/whoqol-qualityoflife/en/.
35. Healthy People 2020. (2010). Foundation Health Measure Report: Health-Related. Quality of Life and Well-Being. Retrived from https://www.healthypeople.gov/sites/default/files/HRQoLWBFullReport.pdf.
36. Ferrans, C. E. P. D., Lipscomb, J., Gotay, C. C., & Snyder, C. (2004). Definitions and conceptual models of quality of life. In C. C. Gotay, C. Snyder, & J. Lipscomb (Eds.), *Outcomes Assessment in Cancer* (pp. 14-30). Cambridge: Cambridge University Press.
37. Sajid, M. S., Tonsi, A., & Baig, M. K. (2008). Health-related quality of life measurement. *International journal of health care quality assurance, 21*(4), 365-373, doi:10.1108/09526860810880162.
38. Hays, R. D., & Reeve, B. B. (2008). Measurement and modeling of health-related quality of life A2 - Heggenhougen, Harald Kristian (Kris). In *International Encyclopedia of Public Health* (pp. 241-252). Oxford: Academic Press.
39. Gill, T. M., & Feinstein, A. R. (1994). A critical appraisal of the quality of quality-of-life measurements. *JAMA, 272*(8), 619-626, doi:10.1001/jama.1994.03520080061045.

40. Osoba, D. (1994). Lessons learned from measuring health-related quality of life in oncology. *Journal of clinical oncology, 12*(3), 608-616, doi:10.1200/jco.1994.12.3.608.
41. Cella, D. F. (1995). Measuring quality of life in palliative care. *Seminars in oncology, 22*(2 Suppl 3), 73-81.
42. Leidy, N. K., Revicki, D. A., & Genesté, B. (1999). Recommendations for evaluating the validity of quality of life claims for labeling and promotion. *Value in Health, 2*(2), 113-127, doi:https://doi.org/10.1046/j.1524-4733.1999.02210.x.
43. Ebrahim, S. (1995). Clinical and public health perspectives and applications of health-related quality of life measurement. *Social Science & Medicine, 41*(10), 1383-1394.
44. Lin, X.-J., Lin, I. M., & Fan, S.-Y. (2013). Methodological issues in measuring health-related quality of life. *Tzu Chi Medical Journal, 25*(1), 8-12, doi:https://doi.org/10.1016/j.tcmj.2012.09.002.
45. Ferrans, C. E. (2007). Differences in what quality-of-life instruments measure. *Journal of the National Cancer Institute*(37), 22-26, doi:10.1093/jncimonographs/lgm008.
46. Fayers, P. M., & Machin, D. (2007). Introduction. In *Quality of Life* (pp. 1-30): John Wiley & Sons, Ltd.
47. Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health technology assessment, 2*(14), i-iv, 1-74.
48. Lenert, L., & Kaplan, R. M. (2000). Validity and interpretation of preference-based measures of health-related quality of life. *Med Care, 38*(9 Suppl), Ii138-150.
49. Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care, 30*(6), 473-483.
50. Simmons, C., & Lehmann, P. (2013). *Tools for strengths-based assessment and evaluation*. New York: Springer.
51. Lins, L., & Carvalho, F. M. (2016). SF-36 total score as a single measure of health-related quality of life: Scoping review. *SAGE Open Medicine, 4*, 1-12, doi:10.1177/2050312116671725.
52. Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal: A review. *Journal of Health Economics, 5*(1), 1-30, doi:https://doi.org/10.1016/0167-6296(86)90020-2.
53. Brooks, R. (1996). EuroQol: the current state of play. *Health Policy, 37*(1), 53-72, doi:https://doi.org/10.1016/0168-8510(96)00822-6.
54. Herdman, M., Gudex, C., Lloyd, A., Janssen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research, 20*(10), 1727-1736, doi:10.1007/s11136-011-9903-x.
55. Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics, 21*(2), 271-292, doi:https://doi.org/10.1016/S0167-6296(01)00130-8.
56. Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care, 42*(9), 851-859.
57. Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., et al. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care, 40*(2), 113-128.
58. Torrance, G. W., Feeny, D. H., Furlong, W. J., Barr, R. D., Zhang, Y., & Wang, Q. (1996). Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care, 34*(7), 702-722.
59. Sintonen, H. (2001). The 15D instrument of health-related quality of life: properties and applications. *Annals of Medicine, 33*(5), 328-336, doi:10.3109/07853890109002086.
60. Kaplan, R. M., & Anderson, J. P. (1988). A general health policy model: update and applications. *Health Service Research, 23*(2), 203-235.
61. Richardson, J., Sinha, K., Iezzi, A., & Khan, M. A. (2014). Modelling utility weights for the Assessment of Quality of Life (AQoL)-8D. *Quality of Life Research, 23*(8), 2395-2404, doi:10.1007/s11136-014-0686-8.
62. Hawthorne, G., Richardson, J., & Osborne, R. (1999). The assessment of quality of life (AQoL) instrument: a psychometric measure of health-related quality of life. *Quality of Life Research, 8*(3), 209-224, doi:10.1023/A:1008815005736.

63. Richardson, J. R., Peacock, S. J., Hawthorne, G., Iezzi, A., Elsworth, G., & Day, N. A. (2012). Construction of the descriptive system for the assessment of quality of life AQoL-6D utility instrument. *Health and Quality of Life Outcomes, 10*(1), 38, doi:10.1186/1477-7525-10-38.
64. Misajon, R., Hawthorne, G., Richardson, J., Barton, J., Peacock, S., Iezzi, A., et al. (2005). Vision and quality of life: the development of a utility measure. *Investigative Ophthalmology & Visual Science, 46*(11), 4007-4015, doi:10.1167/iovs.04-1389.
65. van Reenen, M., & Janssen, B. (2015). EQ-5D-5L User Guide: Basic Information on how to use the EQ-5D-5L Instrument. Retrived from https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf. Accessed 24.05.2018.
66. Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care, 35*(11), 1095-1108.
67. Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics, 27*(1), 7-22, doi:10.1002/hec.3564.
68. Richardson, J. (1994). Cost utility analysis: What should be measured? *Social Science & Medicine, 39*(1), 7-21, doi:https://doi.org/10.1016/0277-9536(94)90162-7.
69. Torrance, G. W., Thomas, W. H., & Sackett, D. L. (1972). A utility maximization model for evaluation of health care programs. *Health Service Research, 7*(2), 118-133.
70. Robinson, A., & Spencer, A. (2006). Exploring challenges to TTO utilities: valuing states worse than dead. *Health Economics, 15*(4), 393-402, doi:10.1002/hec.1069.
71. Devlin, N., Buckingham, K., Shah, K., Tsuchiya, A., Tilling, C., Wilkinson, G., et al. (2013). A comparison of alternative variants of the lead and lag time TTO. *Health Economics, 22*(5), 517-532, doi:10.1002/hec.2819.
72. Rowen, D., Brazier, J., Young, T., Gaugris, S., Craig, B. M., King, M. T., et al. (2011). Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value in Health, 14*, doi:10.1016/j.jval.2011.01.004.
73. Devlin, N. J., Tsuchiya, A., Buckingham, K., & Tilling, C. (2011). A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. *Health Economics, 20*(3), 348-361, doi:10.1002/hec.1596.
74. Augustovski, F., Rey-Ares, L., Irazola, V., Oppe, M., & Devlin, N. J. (2013). Lead versus lag-time trade-off variants: does it make any difference? *The European Journal of Health Economics, 14*(Suppl 1), 25-31, doi:10.1007/s10198-013-0505-0.
75. Janssen, B. M. F., Oppe, M., Versteegh, M. M., & Stolk, E. A. (2013). Introducing the composite time trade-off: a test of feasibility and face validity. *The European Journal of Health Economics, 14*(Suppl 1), 5-13, doi:10.1007/s10198-013-0503-2.
76. Oppe, M., Devlin, N. J., van Hout, B., Krabbe, P. F. M., & de Charro, F. (2014). A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health, 17*(4), 445-453, doi:https://doi.org/10.1016/j.jval.2014.04.002.
77. Norman, R., Cronin, P., & Viney, R. (2013). A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied Health Economics and Health Policy, 11*(3), 287-298, doi:10.1007/s40258-013-0035-z.
78. Craig, B. M., Busschbach, J. J., & Salomon, J. A. (2009). Keep it simple: ranking health states yields values similar to cardinal measurement approaches. *Journal of Clinical Epidemiology, 62*(3), 296-305, doi:10.1016/j.jclinepi.2008.07.002.
79. Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate Data. *Journal of Marketing Research, 20*(4), 350-367, doi:10.2307/3151440.
80. Louviere, J., Hensher, D., & Swait, J. (2000). *Stated choice methods: analysis and application* (Vol. 17). New York: Cambridge University Press.
81. Bansback, N., Brazier, J., Tsuchiya, A., & Anis, A. (2012). Using a discrete choice experiment to estimate health state utility values. *Journal of Health Economics, 31*(1), 306-318, doi:http://dx.doi.org/10.1016/j.jhealeco.2011.11.004.
82. Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goñi, J. M., & Luo, N. (2016). EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *Pharmacoeconomics, 34*(10), 993-1004, doi:10.1007/s40273-016-0404-1.

83. Ramos-Goni, J. M., Pinto-Prades, J. L., Oppe, M., Cabases, J. M., Serrano-Aguilar, P., & Rivero-Arias, O. (2017). Valuation and modeling of EQ-5D-5L health states using a hybrid approach. *Medical Care, 55*(7), 51-58, doi:10.1097/MLR.0000000000000283.
84. Bansback, N., Hole, A. R., Mulhern, B., & Tsuchiya, A. (2014). Testing a discrete choice experiment including duration to value health states for large descriptive systems: Addressing design and sampling issues. *Social Science & Medicine, 114*, 38-48, doi:https://doi.org/10.1016/j.socscimed.2014.05.026.
85. Brazier, J., Rowen, D., Yang, Y., & Tsuchiya, A. (2012). Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. *The European journal of health economic, 13*(5), 575-587, doi:10.1007/s10198-011-0352-9.
86. Green, C., Brazier, J., & Deverill, M. (2000). Valuing health-related quality of life. A review of health state valuation techniques. *Pharmacoeconomics, 17*(2), 151-165.
87. Dolan, P., & Sutton, M. (1997). Mapping visual analogue scale health state valuations onto standard gamble and time trade-off values. *Social Science & Medicine, 44*(10), 1519-1530.
88. Badia, X., Herdman, M., Roset Dipstat, M., & Ohinmaa, A. (2001). Feasibility and validity of the VAS and TTO for eliciting general population values for temporary health states: a comparative study. *Health Services and Outcomes Research Methodology, 2*(1), 51-65, doi:10.1023/a:1011480201653.
89. Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics, 11*(5), 447-456, doi:10.1002/hec.688.
90. Viney, R., Norman, R., Brazier, J., Cronin, P., King, M. T., Ratcliffe, J., et al. (2014). An Australian discrete choice experiment to value EQ-5D health states. *Health Economics, 23*(6), 729-742, doi:10.1002/hec.2953.
91. Hawthorne, G., Richardson, J., & Day, N. A. (2001). A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Ann Med, 33*(5), 358-370.
92. Seymour, J., McNamee, P., Scott, A., & Tinelli, M. (2010). Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. *Health Economics, 19*(6), 683-696, doi:10.1002/hec.1505.
93. Whitehurst, D. G., Norman, R., Brazier, J. E., & Viney, R. (2014). Comparison of contemporaneous EQ-5D and SF-6D responses using scoring algorithms derived from similar valuation exercises. *Value in Health, 17*(5), 570-577, doi:10.1016/j.jval.2014.03.1720.
94. Longworth, L., & Rowen, D. (2013). Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. *Value Health, 16*(1), 202-210, doi:10.1016/j.jval.2012.10.010.
95. Chen, G., Iezzi, A., McKie, J., Khan, M. A., & Richardson, J. (2015). Diabetes and quality of life: Comparing results from utility instruments and Diabetes-39. *Diabetes Research and Clinical Practice, 109*(2), 326-333, doi:10.1016/j.diabres.2015.05.011.
96. Cheung, Y. B., Luo, N., Ng, R., & Lee, C. F. (2014). Mapping the Functional Assessment of Cancer Therapy-Breast (FACT-B) to the 5-level EuroQoL group's 5-dimension questionnaire (EQ-5D-5L) utility index in a multi-ethnic Asian population. *Health and Quality of Life Outcomes, 12*(1), 180, doi:10.1186/s12955-014-0180-6.
97. Chen, G., McKie, J., Khan, M. A., & Richardson, J. R. (2015). Deriving health utilities from the MacNew Heart Disease Quality of Life Questionnaire. *European journal of cardiovascular nursin, 14*(5), 405-415, doi:10.1177/1474515114536096.
98. Meregaglia, M., Borsoi, L., Cairns, J., & Tarricone, R. (2017). Mapping health-related quality of life scores from FACT-G, FAACT, and FACIT-F onto preference-based EQ-5D-5L utilities in non-small cell lung cancer cachexia. *The European Journal of Health Economics*, doi:10.1007/s10198-017-0930-6.
99. Wijnen, B. F. M., Mosweu, I., Majoie, M., Ridsdale, L., de Kinderen, R. J. A., Evers, S., et al. (2018). A comparison of the responsiveness of EQ-5D-5L and the QOLIE-31P and mapping of QOLIE-31P to EQ-5D-5L in epilepsy. *The European Journal of Health Economics, 19*(6), 861-870, doi:10.1007/s10198-017-0928-0.
100. Kaambwa, B., Chen, G., Ratcliffe, J., Iezzi, A., Maxwell, A., & Richardson, J. (2017). Mapping between the Sydney asthma quality of life questionnaire (AQLQ-S) and five multi-attribute

utility instruments (MAUIs). *Pharmacoeconomics, 35*(1), 111-124, doi:10.1007/s40273-016-0446-4.
101. Khan, I., Morris, S., Pashayan, N., Matata, B., Bashir, Z., & Maguirre, J. (2016). Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. *Health and Quality of Life Outcomes, 14*(1), 60, doi:10.1186/s12955-016-0455-1.
102. Mihalopoulos, C., Chen, G., Iezzi, A., Khan, M. A., & Richardson, J. (2014). Assessing outcomes for cost-utility analysis in depression: Comparison of five multi-attribute utility instruments with two depression-specific outcome measures. *British Journal of Psychiatry, 205*(5), 390-397.
103. van Hout, B., Janssen, M. F., Feng, Y. S., Kohlmann, T., Busschbach, J., Golicki, D., et al. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health, 15*(5), 708-715, doi:10.1016/j.jval.2012.02.008.
104. Dakin, H., Abel, L., Burns, R., & Yang, Y. (2018). Review and critical appraisal of studies mapping from quality of life or clinical measures to EQ-5D: an online database and application of the MAPS statement. *Health and Quality of Life Outcomes, 16*(1), 31, doi:10.1186/s12955-018-0857-3.
105. Costa, D. S. (2015). Reflective, causal, and composite indicators of quality of life: A conceptual or an empirical distinction? *Quality of Life Research, 24*(9), 2057-2065, doi:10.1007/s11136-015-0954-2.
106. Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314, doi:10.1037/0033-2909.110.2.305.
107. Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research, 6*(5), 393-406.
108. Boehmer, S., & Luszczynska, A. (2006). Two kinds of items in quality of life instruments: 'indicator and causal variables' in the EORTC QLQ-C30. *Quality of Life Research, 15*(1), 131-141, doi:10.1007/s11136-005-8290-6.
109. Bollen, K. A., Lennox, R. D., & Dahly, D. L. (2009). Practical application of the vanishing tetrad test for causal indicator measurement models: an example from health-related quality of life. *Statistics in medicine, 28*(10), 1524-1536, doi:10.1002/sim.3560.
110. Richardson, J., Kahn, M., Lezzi, A., & Maxwell, A. (2012). Cross-national comparison of twelve quality of life instruments: MIC Paper 1: Background, questions, instruments. Research paper 76. Melbourne, Australia: Monash University.
111. Sintonen, H., & Pekurinen, M. (1993). A fifteen-dimensional measure of health-related quality of life (15D) and its applications. In S. Walker, & R. Rosser (Eds.), *Quality of Life Assessment: Key Issues in the 1990s* (pp. 185-195): Springer Netherlands.
112. Shiroiwa, T., Ikeda, S., Noto, S., Igarashi, A., Fukuda, T., Saito, S., et al. (2016). Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value in Health, 19*(5), 648-654, doi:10.1016/j.jval.2016.03.1834.
113. Augustovski, F., Rey-Ares, L., Irazola, V., Garay, O. U., Gianneo, O., Fernandez, G., et al. (2016). An EQ-5D-5L value set based on Uruguayan population preferences. *Quality of Life Research, 25*(2), 323-333, doi:10.1007/s11136-015-1086-4.
114. Kim, S. H., Ahn, J., Ock, M., Shin, S., Park, J., Luo, N., et al. (2016). The EQ-5D-5L valuation study in Korea. *Quality of Life Research, 25*(7), 1845-1852, doi:10.1007/s11136-015-1205-2.
115. Luo, N., Liu, G., Li, M., Guan, H., Jin, X., & Rand-Hendriksen, K. (2017). Estimating an EQ-5D-5L value set for China. *Value in Health, 20*(4), 662-669, doi:10.1016/j.jval.2016.11.016.
116. M. Versteegh, M., M. Vermeulen, K., M. A. A. Evers, S., de Wit, G. A., Prenger, R., & A. Stolk, E. (2016). Dutch tariff for the five-level version of EQ-5D. *Value in Health, 19*(4), 343-352, doi:https://doi.org/10.1016/j.jval.2016.01.003.
117. Xie, F., Pullenayegum, E., Gaebel, K., Bansback, N., Bryan, S., Ohinmaa, A., et al. (2016). A time trade-off-derived value set of the EQ-5D-5L for Canada. *Medical Care, 54*(1), 98-105, doi:10.1097/mlr.0000000000000447.
118. Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the depression anxiety stress scales* (2nd ed.). Sydney: Psychology Foundation.

119. Kessler, R. C., Barker, P. R., Colpe, L. J., & et al. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry, 60*(2), 184-189, doi:10.1001/archpsyc.60.2.184.
120. Koenker, R. (2005). *Quantile regression* (Econometric Society Monographs). Cambridge: Cambridge University Press.
121. Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143-156, doi:doi: 10.1257/jep.15.4.143.
122. Chay, K. Y., & Powell, J. L. (2001). Semiparametric censored regression models. *The Journal of Economic Perspectives, 15*(4), 29-42.
123. Koenker, R. W., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50.
124. Oppe, M., Devlin, N., & Black, N. (2011). Comparison of the underlying constructs of the EQ-5D and Oxford Hip Score: implications for mapping. *Value in Health, 14*(6), 884-891, doi:10.1016/j.jval.2011.03.003.
125. Gray, A. M., Rivero-Arias, O., & Clarke, P. M. (2006). Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making, 26*(1), 18-29, doi:10.1177/0272989x05284108.
126. Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619-632, doi:Doi 10.1002/(Sici)1099-1255(199611)11:6<619::Aid-Jae418>3.0.Co;2-1.
127. J.S. Ramalho, J., & da Silva, J. V. (2009). A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms. *Quantitative Finance, 9*(5), 621-636, doi:10.1080/14697680802448777.
128. Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11*(1), 54-71, doi:10.1037/1082-989x.11.1.54.
129. Hunger, M., Baumert, J., & Holle, R. (2011). Analysis of SF-6D index data: is beta regression appropriate? *Value in Health, 14*(5), 759-767, doi:10.1016/j.jval.2010.12.009.
130. Khan, I., & Morris, S. (2014). A non-linear beta-binomial regression model for mapping EORTC QLQ- C30 to the EQ-5D-3L in lung cancer patients: a comparison with existing approaches. *Health and Quality of Life Outcomes, 12*, 163, doi:10.1186/s12955-014-0163-7.
131. Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799-815, doi:10.1080/0266476042000214501.
132. Versteegh, M. M., Leunis, A., Luime, J. J., Boggild, M., Uyl-de Groot, C. A., & Stolk, E. A. (2012). Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. *Medical Decision Making, 32*(4), 554-568, doi:10.1177/0272989x11427761.
133. Sullivan, P. W., & Ghushchyan, V. (2006). Mapping the EQ-5D index from the SF-12: US general population preferences in a nationally representative sample. *Medical Decision Making, 26*(4), 401-409, doi:10.1177/0272989X06290496.
134. Bollen, K. A., & Ting, K. F. (2000). A tetrad test for causal indicators. *Psychological methods, 5*(1), 3-22.
135. Bollen, K. A., & Ting, K. F. (1993). Confirmatory tetrad analysis. In P. Marsden (Ed.), *Sociological methodology* (pp. 147-1750). Washington, DC: American Socio-logical Association.
136. Johnson, T. R., & Bodner, T. E. (2007). A note on the use of bootstrap tetrad tests for covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(1), 113-124, doi:10.1080/10705510709336739.
137. Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika, 57*(1), 89-105, doi:10.1007/BF02294660.
138. Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters, 9*(1), 91-97, doi:http://dx.doi.org/10.1016/0167-7152(90)90100-L.
139. Lee, S.-Y., & Shi, J.-Q. (2001). Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics, 57*(3), 787-794, doi:10.1111/j.0006-341X.2001.00787.x.

140. Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189, doi:10.1111/j.2044-8317.1985.tb00832.x.
141. Hu, L.-t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362, doi:10.1037/0033-2909.112.2.351.
142. Chou, C.-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks, CA, US: Sage Publications, Inc.
143. Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling, 8*(3), 353-377, doi:10.1207/S15328007SEM0803_2.
144. Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research, 21*(2), 205-229, doi:10.1177/0049124192021002004.
145. Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55, doi:10.1080/10705519909540118.
146. MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods, 1*(2), 130-149, doi:10.1037/1082-989X.1.2.130.
147. Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258, doi:doi:10.1177/0049124192021002005.
148. Byrne, B. (2012). *Structural Equation Modeling with Mplus*. New York: Routledge.
149. Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., et al. (2018). Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry, 75*(4), 336-346, doi:10.1001/jamapsychiatry.2017.4602.
150. Richardson, J., Iezzi, A., Khan, M. A., & Maxwell, A. (2014). Validity and reliability of the Assessment of Quality of Life (AQoL)-8D multi-attribute utility instrument. *Patient, 7*(1), 85-96, doi:10.1007/s40271-013-0036-x.
151. Brazier, J., Connell, J., Papaioannou, D., Mukuria, C., Mulhern, B., Peasgood, T., et al. (2014). A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health technology assessment, 18*(34), doi:10.3310/hta18340.
152. Price, P. C., Jhangiani, R., & BCcampus (2013). *Research methods in psychology: core concepts and skills*: Flat World Knowledge.
153. Finch, A. P., Brazier, J. E., & Mukuria, C. (2018). What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *The European Journal of Health Economics, 19*(4), 557-570, doi:10.1007/s10198-017-0902-x.
154. Olsen, J. A., Lamu, A. N., & Cairns, J. (2018). In search of a common currency: A comparison of seven EQ-5D-5L value sets. *Health Economics, 27*(1), 39-49.
155. WHO (2001). International classification of functioning, disability and health (ICF). (Vol. 2016). Geneva: World Health Organization.
156. Wilson, I. B., & Cleary, P. D. (1995). Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA, 273*(1), 59-65.
157. Feng, Y., Devlin, N., & Herdman, M. (2015). Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? *Health and Quality of Life Outcomes, 13*, 171, doi:10.1186/s12955-015-0356-8.
158. Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6*(2), 139-150.
159. Kessler, R. C. (2012). The costs of depression. The *psychiatric Clinics of North America, 35*(1), 1-14, doi:10.1016/j.psc.2011.11.005.
160. Aina, Y., & Susman, J. L. (2006). Understanding comorbidity with depression and anxiety disorders. *The Journal of the American Osteopathic Association, 106*(5 Suppl 2), S9-14.

161. Hirschfeld, R. M. A. (2001). The comorbidity of major depression and anxiety disorders: recognition and management in primary care. *Primary Care Companion to The Journal of Clinical Psychiatry, 3*(6), 244-254.

9 Paper 1-3

10 Appendices

Appendix 1: Generic preference-based measures applied

SF-6D classification system (derived from SF-36)

Physical Functioning

- ❒ My health does not limit me in vigorous activities
- ❒ My health limits me a little in vigorous activities
- ❒ My health limits me a little in moderate activities
- ❒ My health limits me a lot in moderate activities
- ❒ My health limits me a little in bathing and dressing
- ❒ My health limits me a lot in bathing and dressing

Role limitations

- ❒ I have no problems with my work or other regular daily activities as a result of my physical health or any emotional problems
- ❒ I am limited in the kind of work or other activities as a result of my physical health
- ❒ I accomplish less than I would like as a result of emotional problems
- ❒ I am limited in the kind of work or other activities as a result of my physical health and accomplish less than I would like as a result of emotional problems

Social functioning

- ❒ My health limits my social activities none of the time
- ❒ My health limits my social activities a little of the time
- ❒ My health limits my social activities some of the time
- ❒ My health limits my social activities most of the time
- ❒ My health limits my social activities all of the time

Pain

- ❒ I have no pain
- ❒ I have pain but it does not interfere with my normal work (both outside the home and housework)
- ❒ I have pain that interferes with my normal work (both outside the home and housework) a little bit
- ❒ I have pain that interferes with my normal work (both outside the home and housework) moderately
- ❒ I have pain that interferes with my normal work (both outside the home and housework) quite a bit
- ❒ I have pain that interferes with my normal work (both outside the home and housework) extremely

Mental health

- ❒ I feel tense or downhearted and low none of the time
- ❒ I feel tense or downhearted and low a little of the time
- ❒ I feel tense or downhearted and low some of the time
- ❒ I feel tense or downhearted and low most of the time
- ❒ I feel tense or downhearted and low all of the time

Vitality

- ❒ I have a lot of energy all of the time
- ❒ I have a lot of energy most of the time
- ❒ I have a lot of energy some of the time
- ❒ I have a lot of energy a little of the time
- ❒ I have a lot of energy none of the time

HUI-3 classification system (Please select the answer for each attribute that is correct for you)

Vision

- ❒ Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses.
- ❒ Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses.
- ❒ Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses.
- ❒ Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses.
- ❒ Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses.

- ❒ Unable to see at all.

Hearing

- ❒ Able to hear what is said in a group conversation with at least three other people, without a hearing aid.
- ❒ Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but requires a hearing aid to hear what is said in a group conversation with at least three other people.
- ❒ Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people, with a hearing aid.
- ❒ Able to hear what is said in a conversation with one other person in a quiet room, without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
- ❒ Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid.
- ❒ Unable to hear at all.

Speech

- ❒ Able to be understood completely when speaking with strangers or people who know me well.
- ❒ Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well.
- ❒ Able to be understood partially when speaking with strangers or people who know me well.
- ❒ Unable to be understood when speaking with strangers but able to be understood partially by people who know me well.
- ❒ Unable to be understood when speaking to other people (or unable to speak at all).

Ambulation

- ❒ Able to walk around the neighbourhood without difficulty, and without walking equipment.
- ❒ Able to walk around the neighbourhood with difficulty, but does not require walking equipment or the help of another person.
- ❒ Able to walk around the neighbourhood with walking equipment, but without the help of another person.
- ❒ Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighbourhood.
- ❒ Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighbourhood.
- ❒ Cannot walk at all.

Dexterity

- ❒ Full use of two hands and ten fingers.
- ❒ Limitations in the use of hands or fingers, but does not require special tools or help of another person.
- ❒ Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person).
- ❒ Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with the use of special tools).
- ❒ Limitations in the use of hands or fingers, requires the help of another person for most tasks (not independent even with the use of special tools).
- ❒ Limitations in the use of hands or finders, requires the help of another person for all tasks (not independent even with the use of special tools).

Emotion

- ❒ Happy and interested in life.
- ❒ Somewhat happy.
- ❒ Somewhat unhappy.
- ❒ Very unhappy.
- ❒ So unhappy that life is not worthwhile.

Cognition

- ❒ Able to remember most things, think clearly and solve day to day problems.
- ❒ Able to remember most things, but have a little difficulty when trying to think and solve day to day problems.
- ❒ Somewhat forgetful, but able to think clearly and solve day to day problems.
- ❒ Somewhat forgetful, and have a little difficulty when trying to think or solve day to day problems.
- ❒ Very forgetful, and have great difficulty when trying to think or solve day to day problems.
- ❒ Unable to remember anything at all, and unable to think or solve day to day problems.

Pain

- ❒ Free of pain and discomfort.
- ❒ Mild to moderate pain that prevents no activities.
- ❒ Moderate pain that prevents a few activities.

- ❒ Moderate to severe pain that prevents some activities.
- ❒ Severe pain that prevents most activities.

15D (Select the answer which best describes *your present health status*)

Mobility
- ❒ I am able to walk normally (without difficulty) indoors, outdoors and on stairs.
- ❒ I am able to walk without difficulty indoors, but outdoors and/or on stairs I have slight difficulties.
- ❒ I am able to walk without help indoors (with or without an appliance), but outdoors and/or on stairs only with considerable difficulty or with help from others.
- ❒ I am able to walk indoors only with help from others.
- ❒ I am completely bed-ridden and unable to move about.

Vision
- ❒ I see normally, i.e. I can read newspapers and TV text without difficulty (with or without glasses).
- ❒ I can read papers and/or TV text with slight difficulty (with or without glasses).
- ❒ I can read papers and/or TV text with considerable difficulty (with or without glasses).
- ❒ I cannot read papers or TV text either with glasses or without, but I can see enough to walk about without guidance.
- ❒ I cannot see enough to walk about without a guide, i.e. I am almost or completely blind.

Hearing
- ❒ I can hear normally, i.e. normal speech (with or without a hearing aid).
- ❒ I hear normal speech with a little difficulty.
- ❒ I hear normal speech with considerable difficulty; in conversation I need voices to be louder than normal.
- ❒ I hear even loud voices poorly; I am almost deaf.
- ❒ I am completely deaf.

Breathing
- ❒ I am able to breathe normally, i.e. with no shortness of breath or other breathing difficulty.
- ❒ I have shortness of breath during heavy work or sports, or when walking briskly on flat ground or slightly uphill.
- ❒ I have shortness of breath when walking on flat ground at the same speed as others my age.
- ❒ I get shortness of breath even after light activity, e.g. washing or dressing myself.
- ❒ I have breathing difficulties almost all the time, even when resting.

Sleeping
- ❒ I am able to sleep normally, i.e. I have no problems with sleeping
- ❒ I have slight problems with sleeping, e.g. difficulty in falling asleep, or sometimes waking at night.
- ❒ I have moderate problems with sleeping, e.g. disturbed sleep, or feeling I have not slept enough.
- ❒ I have great problems with sleeping, e.g. having to use sleeping pills often or routinely, or usually waking at night and/or too early in the morning.
- ❒ I suffer severe sleeplessness, e.g. sleep is almost impossible even with full use of sleeping pills, or staying awake most of the night.

Eating
- ❒ I am able to eat normally, i.e. with no help from others.
- ❒ I am able to eat by myself with minor difficulty (e.g. slowly, clumsily, shakily, or with special appliances).
- ❒ I need some help from another person in eating.
- ❒ I am unable to eat by myself at all, so I must be fed by another person.
- ❒ I am unable to eat at all, so I am fed either by tube or intravenously

Speach
- ❒ I am able to speak normally, i.e. clearly, audibly and fluently.
- ❒ I have slight speech difficulties, e.g. occasional fumbling for words, mumbling, or changes of pitch.
- ❒ I can make myself understood, but my speech is e.g. disjointed, faltering, stuttering or stammering.
- ❒ Most people have great difficulty understanding my speech.
- ❒ I can only make myself understood by gestures.

Elimination
- ❒ My bladder and bowel work normally and without problems.
- ❒ I have slight problems with my bladder and/or bowel function, e.g. difficulties with urination, or loose or hard bowels.
- ❒ I have marked problems with my bladder and/or bowel function, e.g. occasional 'accidents', or severe constipation or diarrhea.

- ❒ I have serious problems with my bladder and/or bowel function, e.g. routine 'accidents', or need of catheterization or enemas.
- ❒ I have no control over my bladder and/or bowel function.

Usual Activities

- ❒ I am able to perform my usual activities (e.g. employment, studying, housework, free-time activities) without difficulty.
- ❒ I am able to perform my usual activities slightly less effectively or with minor difficulty.
- ❒ I am able to perform my usual activities much less effectively, with considerable difficulty, or not completely.
- ❒ I can only manage a small proportion of my previously usual activities.
- ❒ I am unable to manage any of my previously usual activities.

Mental Function

- ❒ I am able to think clearly and logically, and my memory functions well
- ❒ I have slight difficulties in thinking clearly and logically, or my memory sometimes fails me.
- ❒ I have marked difficulties in thinking clearly and logically, or my memory is somewhat impaired.
- ❒ I have great difficulties in thinking clearly and logically, or my memory is seriously impaired.
- ❒ I am permanently confused and disoriented in place and time.

Discomfort and Symtoms

- ❒ I have no physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
- ❒ I have mild physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc
- ❒ I have marked physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.
- ❒ I have severe physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc
- ❒ I have unbearable physical discomfort or symptoms, e.g. pain, ache, nausea, itching etc.

Depression

- ❒ I do not feel at all sad, melancholic or depressed.
- ❒ I feel slightly sad, melancholic or depressed.
- ❒ I feel moderately sad, melancholic or depressed.
- ❒ I feel very sad, melancholic or depressed.
- ❒ I feel extremely sad, melancholic or depressed.

Distress

- ❒ I do not feel at all anxious, stressed or nervous.
- ❒ I feel slightly anxious, stressed or nervous.
- ❒ I feel moderately anxious, stressed or nervous.
- ❒ I feel very anxious, stressed or nervous.
- ❒ I feel extremely anxious, stressed or nervous.

Vitality

- ❒ I feel healthy and energetic.
- ❒ I feel slightly weary, tired or feeble.
- ❒ I feel moderately weary, tired or feeble.
- ❒ I feel very weary, tired or feeble, almost exhausted.
- ❒ I feel extremely weary, tired or feeble, totally exhausted.

Sexual Activity

- ❒ My state of health has no adverse effect on my sexual activity.
- ❒ My state of health has a slight effect on my sexual activity.
- ❒ My state of health has a considerable effect on my sexual activity.
- ❒ My state of health makes sexual activity almost impossible.
- ❒ My state of health makes sexual activity impossible.

Appendix 2: **Disease-specific instruments applied**

DASS-21

Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you **over the past week**.

The rating scale is as follows:
0 Did not apply to me at all
1 Applied to me to some degree, or some of the time
2 Applied to me to a considerable degree or a good part of time
3 Applied to me very much or most of the time

1.	I found it hard to wind down	0	1	2	3
2.	I was aware of dryness of my mouth	0	1	2	3
3.	couldn't seem to experience any positive feeling at all	0	1	2	3
4.	I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion)	0	1	2	3
5.	I found it difficult to work up the initiative to do things	0	1	2	3
6.	I tended to over-react to situations	0	1	2	3
7.	I experienced trembling (eg, in the hands)	0	1	2	3
8.	I felt that I was using a lot of nervous energy	0	1	2	3
9.	I was worried about situations in which I might panic and make a fool of myself	0	1	2	3
10.	I felt that I had nothing to look forward to	0	1	2	3
11.	I found myself getting agitated	0	1	2	3
12.	I found it difficult to relax	0	1	2	3
13.	I felt down-hearted and blue	0	1	2	3
14.	I was intolerant of anything that kept me from getting on with what I was doing	0	1	2	3
15.	I felt I was close to panic	0	1	2	3
16.	I was unable to become enthusiastic about anything	0	1	2	3
17.	I felt I wasn't worth much as a person	0	1	2	3
18.	I felt that I was rather touchy	0	1	2	3
19.	I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat)	0	1	2	3
20.	I felt scared without any good reason	0	1	2	3
21.	I felt that life was meaningless	0	1	2	3

K10

These questions deal with the **past 4 weeks**.

	All of the time	Most of the time	Some of the time	A little of the time	None of the time
1. In the past 4 weeks, about how often did you feel tired for no good reason?					
2. In the past 4 weeks, about how often did you feel nervous?					
3. In the past 4 weeks, about how often did you feel so nervous that nothing could calm you down?					
4. In the past 4 weeks, about how often did you feel hopeless?					
5. In the past 4 weeks, about how often did you feel restless or fidgety?					
6. In the past 4 weeks, about how often did you feel so restless that you could not sit still?					
7. In the past 4 weeks, about how often did you feel depressed?					
8. In the past 4 weeks, about how often did you feel that everything was an effort?					
9. In the past 4 weeks, about how often did you feel so sad that nothing could cheer you up?					
10. In the past 4 weeks, about how often did you feel worthless?					