Faculty of health sciences / Department of community medicine

Mapping the EORTC QLQ-C30 to four preference-based measures (EQ-5D, SF-6D, 15D and HUI3).

Martin Jack Mwamba

HEL-3950 Master’s thesis in Public Health
December 2017

Supervisor: PhD Cand. Admassu Nadew Lamu
Co-supervisor: Prof. Jan Abel Olsen

Martin Jack Mwamba

2017

Mapping the EORTC QLQ-C30 to four preference-based measures (EQ-5D, SF-6D, 15D and HUI3).

Martin Jack Mwamba

https://www.ub.uit.no/iportal/ldap-login

University of Tromsø

ABSTRACT

Background: Cost utility analysis evaluates health care interventions in terms of their cost per quality adjusted life year (QALY) gained. The EQ-5D, SF-6D, 15D and HUI3 are the most common health state utility (HSU) instruments used to put the 'quality adjustment weight' into the QALY. However, HSU instruments are not always available or appropriate for every health conditions. For measuring the general cancer quality of life, clinicians and researchers prefer to use the European organization for research and treatment quality of life questionnaire core 30 (EORTC QLQ-C30). But the EORTC QLQ-C30 is not 'preference-based' and thus cannot be used to derive the 'quality adjustment weight' for use in QALYs. Mapping algorithms have been developed to predict health state values from EORTC QLQ-C30 but there is considerable uncertainty as to which HSU instrument best fits EORTC QLQ-C30.

Objectives: To estimate mapping models that predict utilities for four HSU instruments (EQ-5D, SF-6D, 15D and HUI3) based on EORTC QLQ-C30 using two regression techniques (OLS and GLM).

Methods: Data used for the study was obtained from the multi-instrument comparison (MIC) survey. The study focused on 772 respondents (cancer patients) who completed the questionnaires for EORTC QLQ-C30, EQ-5D, SF-6D, 15D and HUI3. Mapping algorithms were fitted to predict health state values for EQ-5D, SF-6D, 15D and HUI3 from the scales/items of EORTC QLQ-C30 using ordinary least square (OLS) methods and generalized linear models (GLM). Model predictive ability was compared by normalized mean absolute error (%MAE) and root mean squared error (%RMSE) even though the R^2, MAE and RMSE were reported.

Results: The OLS model generated identical mean utility values to the observed values for EQ-5D, SF-6D and 15D compared to only 15D for the GLM model. Explanatory powers were relatively high for all four HSU instruments with the R^2 ranging from 0.601 (HUI3 using GLM) to 0.762 (15D using OLS). The lowest %MAE was generated by the EQ-5D algorithm (6.4%) using OLS and the highest %MAE was for HUI3 (11.9%) using GLM. Algorithm mapping onto EQ-5D had the lowest %RMSE (9.3%) using OLS and the highest %RMSE was for HUI3 (15.1%) using GLM.

Conclusion: The mapping algorithms presented in the study prove that the scores of EORTC QLQ-C30 can be mapped onto any of the four HSU instruments without significantly compromising the results of the intended CUA.

ACKNOWLEDGEMENT

I would like to thank my main supervisor, Admassu Nadew Lamu, research fellow and Ph.D. candidate at ISM (Institutt for samfunnsmedisin) for his patience, understanding and expertise. I will forever be grateful for his valuable input and guidance throughout the period of the thesis writing. This master thesis would not be a success without the involvement of Prof. Jan Abel Olsen at the department of community medicine at university of Tromsø, who acted as my co-supervisor. Not only did he provide the necessary advice, but he also introduced me to the Multi-instrument comparison study (MIC) from where the data used in the thesis came from. I ‘am deeply grateful for the help and support he rendered throughout the writing process.

I would also like to express my sincere gratitude to the support staff at ISM: the program coordinator Tor Gisle Lorentzen and Prof. Bjarne Koster Jacobsen. Thank you for coordinating the MPH program so very well. The class of 2015, masters of public health deserves a very special thank you. They made my studying at UiT worthwhile. Thankyou class 2015 for the time we spent together.

Finally, I would like to thank my family, especially my wife Miyanda and my two children, Chileshe and Chimuka. Their support during the whole two years of studying at UiT was unwavering and helped me reach the finishing line.

Martin Jack Mwamba

Tromsø, December 2017

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

CUA - Cost utility analyses

DSI – Disease specific measures

DCE - Discreet choice experiment

EQ-5D - Euro-QoL-5-dimesions

EORTC QLQ-C3 - European organization for research and treatment quality of life questionnaire core 30 for cancer

HRQoL - Health related quality of life

HSUI – Health state utility instruments

HUI - Health utility index

ISM - Institutt for samfunnsmedisin (Norwegian) (Department for community medicine - English)

MAE – Mean absolute errors

MAUT – Multi-attribute utility theory

MCID - Minimal clinically important difference

MIC – Multi-instrument comparison

QoL - Quality of life

QALY - Quality adjusted life years

RMSE – Root mean standard error

SF-6D - Short form-6- dimensions

SG - Standard gamble

TTO - Time trade off

VAS – Visual analogy scale

1. Introduction

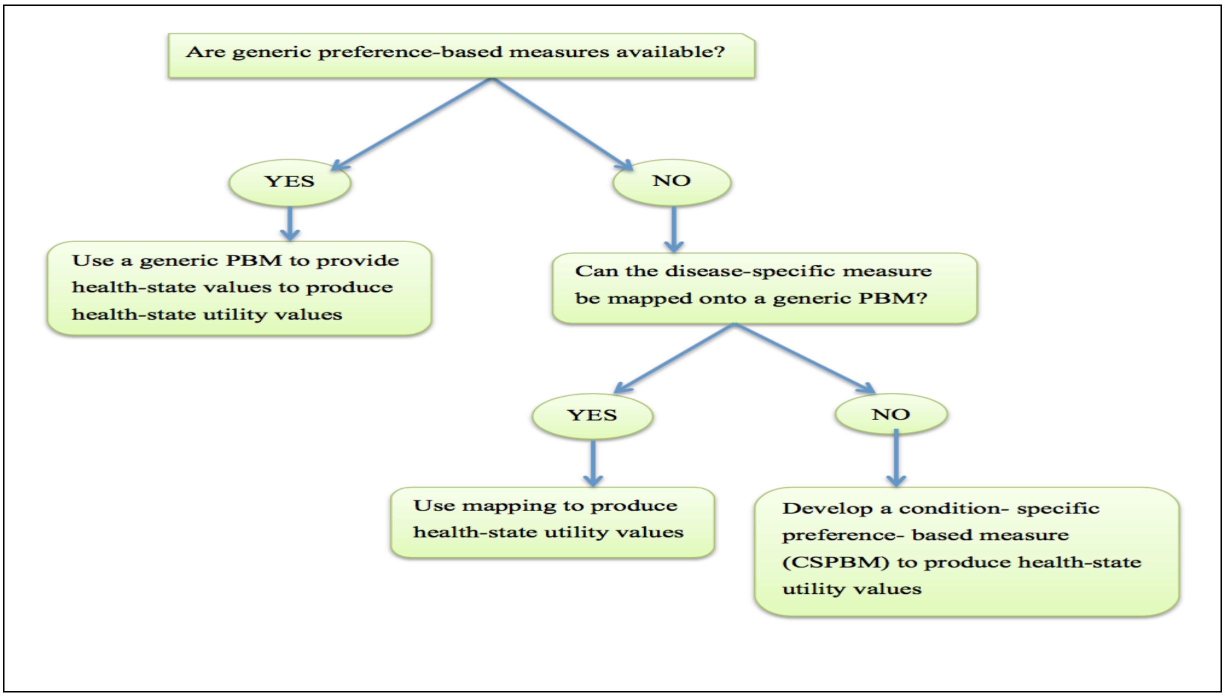
The recent advancements in cancer treatment and improved living conditions in terms of housing, hygiene and food have led to increased survival time and improved health related quality of life (HRQoL) for patients with cancer [1]. For the health sector, these advancements translate into increased demand for health care services with resultant increase in health care budgets. The estimates of costs related to cancer vary depending on the scope of analysis (patient, hospital or national level), cancer type (one or several cancers) and data sources (cancer registry, insurance claims, medical records). However, experts agree that these costs are enormous and are expected to rise further with increased cancer incidence in an aging population [2]. For instance, in the Nordic countries (Norway, Sweden, Finland, Denmark and Iceland), the annual treatment costs associated with cancer, including hospital costs and costs for prescription drugs, was estimated at 3 billion Euros in 2007 [3]. In the United States, the national cost of cancer care in 2010 was estimated to be $124.57 billion and the projected costs in 2020 are estimated to be $157.77 billion [2].

Given that health care resources are limited, health economists employ cost utility analysis as a means to inform the allocation of health care resources. Since alternative allocations of health care resources produce different outcomes, there is need to use a measure of health outcome that compares across different health care areas. Quality adjusted life year (QALY), which is a product of life years and health state utility is used as a generic measure of health outcome for this purpose. QALYs provides a way of comparing competing health care programs. Utilities for calculating QALYs are obtained from health state utility (HSU) instruments such as the EQ-5D, SF-6D, 15D and HUI3.

On the other hand, it is through clinical trials that the effectiveness of health care interventions can be evaluated. But then most clinical trials include disease specific instruments (DSI) and not HSU instruments. Therefore, this creates a mismatch between the information required for economic evaluation of health care interventions and the information generated in the clinical trials [4].

Furthermore, when it comes to measuring the HRQoL for most medical conditions and for cancer in particular, HSU instruments have been found to be insensitive to small but clinically important changes in HRQoL [5]. HSU instruments have also been found to be insensitive to the effects of cancer treatments on specific cancer related symptoms and side effects [5]. Therefore, though DSI provide valuable evidence on the effectiveness of an intervention, they cannot be used to calculate QALYs for use in economic evaluation of health care interventions. As shown in Figure 1, in the absence of HSU instruments, one solution is to map from the DSI on to HSU instruments using regression techniques [6-8].

Figure 1. Simple flow chart for deciding when to perform mapping

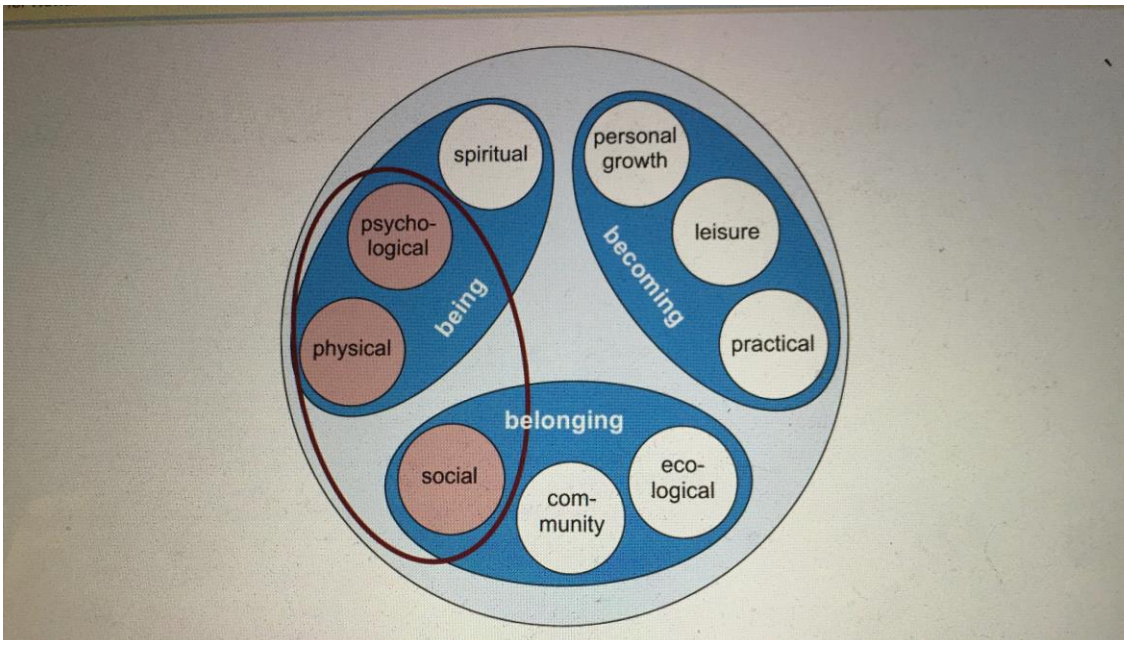

Are generic preference-based measures available?
YES
NO
Use a generic PBM to provide health-state values to produce health-state utility values
Can the disease-specific measure be mapped onto a generic PBM?
YES
NO
Use mapping to produce health-state utility values
Develop a condition- specific preference- based measure (CSPBM) to produce health-state utility values

The objective of this study is to map the scores of the EORTC QLQ-C30 to the four most widely used HSU instruments (EQ-5D, SF-6D, 15D and HUI3) and compare and establish which among them best fits the EORTC QLQ-C30 for estimating health state values. To arrive at this objective, the study is divided as follows: Chapter 2 discusses the theoretical background on HRQoL with particular attention given to the EORTC QLQ-C30 and four HSU instruments (EQ5D, SF-6D, 15D and HUI3). Chapter 3 presents the data, instruments and the analyses used. Chapter 4 (results section) presents the findings and Chapter 5 discusses these findings.

2. Theoretical Background

The terms health-related quality of life (HRQoL) and quality of life (QoL) are usually used interchangeably to refer to the same concept even though the two differ in some aspects. The general term QoL differs from HRQoL in that QoL is an all-inclusive concept incorporating all factors that impact upon an individual's life such as biological, physical, emotional, social, economic, cultural and spiritual aspects [9]. On the other hand, HRQoL is a subset relating only to the health domains of QoL (Figure 2).

Figure 2. Health-related quality of life as a component of quality of life



Source: Renwick *et. al*. (1996)

Considering that HRQoL is essentially a multidimensional phenomenon, most institutions including the WHO have adopted a multi-dimensional approach to measuring HRQoL of individuals or populations. The multi-dimensional concept encompasses mostly the physical, emotional and social components associated with illness and treatment. HRQoL is also considered a subjective matter and therefore individuals should assess how these components are affected by illness and treatment [10].

2.1. Classification of health-related quality of life instruments

HRQoL instruments measure quality of life relative to the health or disease status of individuals or populations. Several multi-attribute instruments (MAI) have been developed and used by clinicians, researchers and policy makers to measure HRQoL as it is affected by disease and treatment [11]. They can be generalized into three categories: disease specific, generic and health state utility instruments (Table 1)

Table 1. Classification system of health-related quality of life measures

Type	Purpose
Disease specific instruments	Includes aspects of health that are relevant to particular health problems and may measure several health domains
Generic instruments	Can be used across different patient populations and usually measures several health domains
Health state utility instruments	Developed for economic evaluation and incorporates preferences for health states

Source: Garratt *et. al.*, (2002)

2.1.1 Disease specific instruments

DSI are narrower in design in that they are meant to measure the HRQoL of a particular disease or condition. They focus on special areas of primary interest, where the measure maybe specific to a disease such as cancer or heart disease [12]. Their narrow range of applicability allows them to be useful for measuring small but clinically important changes that may be of relevance to clinicians [13]. For this reason, DSI are mostly used in clinical trials to assess within subject change in health status over a period of time. Despite being the most used measures in clinical trials and the most likely to capture the impact of disease or treatment on the HRQoL of patients with particular conditions, DSI are not utility-based and therefore cannot be used to estimate QALYs. Some examples of DSI are the European organization for research and treatment quality of life questionnaire core 30 (EORTC QLQ-C30–C30) for cancer, Diabetes-39 (D-39) for diabetes, Asthma quality of Life questionnaire (AQLQ-C30-Sydney) for asthma, IBS-QOL for irritable bowel syndrome, K10 and DASS 21 for depression, Macnew heart QoL and heart specific activity scale (SAS) for heart diseases and Arthritis impact measurement scale (AIMS2-SF) for Arthritis.

2.1.2 Generic instruments

Generic instruments provide a broad assessment of HRQoL. They incorporate domains that are health-related and thus influenced by disease, injury, treatment or health policy. Such domains include the duration of life, functional states, impairments, perceptions and social opportunity [13]. Generally, they are applicable to all types of patients irrespective of their condition or treatment because they have a standard unit of measure. They are preferred for their ability to capture a comprehensive picture of HRQoL across all patient populations and hence are used to evaluate treatments, allocate resources, or compare disease burden between patient groups. However, they are not utility-based and thus may not be used to calculate QALYs. Moreover, they may not cover all dimensions of relevance to some medical conditions as their focus is general rather than specific, and also they may not be appropriate for all conditions [14]. They include SF-36, the Sickness impact profile and the Nottingham health profile.

2.1.3 Health state utility instruments

HSU instruments are a specialized type of generic instruments that measure the patient’s utility or preference for a particular health state [15]. HSU instruments are composed of two parts: a health classification or descriptive system that defines health states and a valuation system or algorithm that converts the attribute responses into an index value or utility. Similar to generic measures, HSU instruments such as the EQ-5D, SF-6D, 15D and HUI3 can be applied to patients regardless of their condition or treatment and thus useful for comparing outcomes across patient groups. However, unlike generic measures, HSU instruments can assign a single numerical value to any health state that is defined by the descriptive system. These index values are used to estimate QALYs for use in economic evaluation of health care programs [16].

2.1.3.1 Health state valuation methods: Preference elicitation techniques

The concept of preference estimation is rooted in the economics of decision theory and is used to explain preference relations. The terms utility and value are often used interchangeably with the term preference [17]. However, preference is a general term that describes the desirability of a set of outcomes whilst utility and values are different types of preferences that depend on the method used to estimate the preference weights [15].

There are two methods of preference estimation: choice methods and rating scale methods. Examples of choice methods include the time trade off (TTO), standard gamble (SG) and the discreet choice experiment (DCE). The most commonly used rating scale method is the visual analogy scale (VAS). These methods of preference estimation can be applied singly or in combination. For example, the SF-6D uses the SG whilst the EQ-5D-5L combines the TTO and DCE.

The TTO measure values based on conditions of certainty since the alternatives presented to the respondents have outcomes that are known with certainty. The SG on the other hand measures utilities under conditions of uncertainty that satisfy certain axioms of expected utility theory. DCE are choice based methods that allow respondents to choose between scenarios that describe a health state by different levels of attributes of that health state [18]. Like TTO, DCE measure values based on conditions of certainty.

Using the narrow definition from expected utility theory, it can be argued that only preference weights developed from SG produce utilities whilst the rest produce values [15]. Based on this argument, the SG has been taken to be the “gold standard” in health state utility measurements. In the broader sense, utility can be measured under conditions of both uncertainty and certainty using SG and TTO or DCE respectively [19].

Utility can be referred to as a way of valuing HRQL and represents an individual's relative satisfaction with a health state [18]. Health state utilities are measured on a cardinal scale of 0 – 1, where 0 indicates death and 1 indicates full health. Anchoring the utility measurement on an interval scale of 0 and 1, allows for the same change to mean the same irrespective of the part of the scale being considered (e.g. a change in health from 0.1 to 0.2 is equivalent to a change from 0.6 to 0.7). States worse than death can also be accounted for, with such states taking a negative value [20].

2.1.3.2 Scoring Approach and forms of algorithm

The valuation systems for the EQ-5D and SF-6D apply statistical modeling or regression analysis. Econometric approaches using additive functional forms are used to estimate overall health index values for these instruments. Contrary, the valuation systems for the 15D and HUI3 are based on the application of the multi-attribute utility theory (MAUT). The MAUT is an extension of the von Neumann-Morgenstern theory that considers utility functions with more than one attribute. In order to produce the overall health index, each attribute contributes a single attribute utility function [11, 21]. There are three possible ways in which these single attribute utility functions could be combined to form the overall health index. Depending on the type of preference relation among these attributes, the combination could be additive, multiplicative or multi-linear. The additive form allows for no preference interactions among the attributes whilst the multiplicative form allows for one type of preference interaction among the attributes [9]. On the other hand, the multi-linear form allows for several types of preference interactions among the attributes. The 15D is based on the additive functional form whilst the HUI3 is based on the multiplicative functional form. For 15D, population preferences are elicited with rating scales (VAS) and for HUI, population preferences are elicited with VAS and SG [9, 10]. A summary of the properties of the HSU instruments included in the study is presented in Table 2.

Table 2. Properties of health state utility instruments included in the study

	EQ-5D-5l	**SF-6D**	**15D**	**HUI3**
Source of weights	Rabin *et.al.*	Brazier *et.al.*	Sintonen *et.al.*	Feeny *et.al.*
Model	Econometric	Econometric	DA additive	DA multiplicative
Source of utility	TTO and Rating scale	Standard gamble	Rating scale	VAS & Standard gamble
Defined states	3125	18,000	3.1×10^{10}	972,000

HUI3 – Health utility index 3; TTO – Time Trade Off; VAS – Visual Analogy Scale;

Source: Chen *et.al.*[22].

3. Methods

3.1 Study Sample

The study is based on data from the multi-instrument comparison (MIC) study that carried out an online survey in 2012. The survey was carried out in six countries (Australia, Canada, Germany, UK, US and Norway) by a global panel company, CINT Pty Ltd [23]. Quotas were applied to obtain two groups of respondents: the healthy group for those who reported no chronic disease and a VAS score greater than 70 and the disease group for those who reported any of the seven chronic diseases (depression, hearing loss, Asthma, Diabetes, Arthritis, heart disease and cancer). For this study, only data pertaining to respondents who completed the EORTC QLQ-C30 questionnaire (i.e. cancer patients) were analysed (N= 772). Further data editing procedure on the MIC study can be found in Richardson *et. al.* [23].

3.2 Instruments

The main purpose is to estimate mapping models that predict utilities for four HSU instruments (EQ-5D, SF-6D, 15D and HUI3) based on EORTC QLQ-C30.

EQ-5D

EQ-5D is a standardized measure of HRQoL developed by the EuroQol Group in order to provide a simple, generic measure of health for clinical and economic appraisal [24]. The EQ-5D descriptive system comprises 5 dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. In the earlier version, (the EQ-5D-3L), each dimension has 3 levels: no problems, some problems, extreme problems. The revised version (EQ-5D-5L) has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problem [25]. The revised England EQ-5D-5L Tariff is considered in this study. The EQ-5D-5L English tariff uses preference weights obtained from TTO and DCE. The respondent is asked to indicate his/her health state by ticking (or placing a cross) in the box against the most appropriate statement in each of the

5 dimensions. A health state is described as comprising one level from each of the five dimensions. Therefore, each health state is referred to in terms of a five-digit code (for example 12413 implying no problems on mobility, slight problems on self-care, severe problems on usual activity, no problems on pain/discomfort and moderate problems on anxiety/depression). A total of 3125 (55) possible health states are defined in this way for the EQ-5D-5L.

SF-6D

The SF-6D is a utility-based instrument that estimates preference-based index scores derived from SF-36 items. Unlike the EQ-5D-5L that uses preference weights obtained from the TTO and DCE valuation techniques, the preference weights used in the SF-6D are obtained from SG. However, like many other HSU instruments, the SF-6D is composed of two parts: the health state classification system that describes health states and a set of values used for scoring the health states. The health state classification system consists of six dimensions comprising of physical functioning, role limitations, social functioning, pain, mental health and vitality, with four to six levels of severity for each, generating a total of 18,000 possible health states [21].

15D

The 15D is a utility-based instrument with 15 dimensions: mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, and sexual activity, each with five possible response levels, structured from the best to the worst possible health condition [10]. The 15D instrument can generate over 30 billion different health states. The valuation system of the 15D is based on the principles of multi-attribute utility theory. A weight for each level of each dimension is obtained by multiplying the level value by the importance weight of the dimension at that level. The preference weights for 15D have been elicited from representative population samples using rating scales (VAS) [10].

HUI3

The Health Utilities Index version-3 (HUI3) consists of eight dimensions (vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain/discomfort), each with five or six levels giving a total of 972,000 possible health states [9, 26]. The HUI3 is predominantly constructed of attributes that relate to physical and emotional health with the exclusion of the social attributes. Like the 15D, the HUI3 applies the multi-attribute utility theory to estimate societal preference scores on an interval scale of 0 for dead and 1 for full health using VAS and SG methods. HUI3 uses multiplicative, multi-attribute utility functions [27].

EORTC QLQ-C30

The EORTC QLQ-C30 is a cancer specific instrument designed for measuring the general cancer quality of life (Table 3). The EORTC QLQ-C30 is composed of 30 questions: 24 of which aggregate into nine multi-item scales representing various dimensions of HRQoL: five functional scales - physical, role, emotional, cognitive and social), three symptom scales - fatigue, pain, nausea and vomiting and one global scale [28, 29]. The other 6 mono-item scales assess other relevant cancer-oriented symptoms - dyspnoea, insomnia, appetite, constipation, diarrhoea and financial difficulties. All EORTC QLQ-C30 items have four response options (i.e. ‘not at all’, ‘a little’, ‘quite a bit’ and ‘very much’) except for the two items (29 and 30) assessing global QLQ that use a seven-point scale [28].

The global, functional and symptom scores of the EORTC QLQ-C30 were constructed following the published EORTC QLQ-C30 rules and standardized to a range of 0 to 100 with higher scores representing higher response levels [28]. Therefore, a high score for the global health status represents a high or better HRQoL, higher functional scale scores indicate better HRQoL, but higher symptom scale/item scores indicate higher level of symptoms or poor HRQoL.

Table 3. Description of the EORTC QLQ-C30 instrument

	Scale	No of items	Item range	Item numbers
Global health status	QLQ	2	6	29 & 3
Functional scales				
Physical functioning	PF	5	3	1 to 5
Role functioning	RF	2	3	6 & 7
Emotional functioning	EF	4	3	21 to 24
Cognitive functioning	CF	2	3	20 & 25
Social functioning	SF	2	3	26 & 27
Symptom scale/items				
Fatigue	FA	3	3	10, 12 & 18
Nausea and Vomiting	NV	2	3	14 & 15
Pain	PA	2	3	9 & 19
Dyspnoea	DY	1	3	8
Insomnia	SL	1	3	11
Loss of appetite	AP	1	3	13
Constipation	CO	1	3	16
Diarrhoea	DI	1	3	17
Financial difficulties	FI	1	3	28

QLQ – Global health scale; PF – Physical functioning; RF – Role functioning; EF – Emotional functioning; CF – cognitive functioning; SF – Social functioning; FA – fatigue; NV – Nausea/Vomiting; PA – Pain; DY – Dyspnoea; SL – Insomnia; AP – Loss of appetite; CO – Constipation; DI – diarrhoea; FI – Financial impact.

Source: Aaronson *et. al.*, [28]

3.3 Validation of EORTC QLQ-C30

Considering that the aim was to predict the unknown HRQoL values for the HSU instruments using the known HRQoL values from the EORTC QLQ-C30, there was need to assess how accurately the EORTC QLQ-C30 captures the multi-dimensionality construct of HRQoL of cancer patients. For that reason, the validity of the EORTC QLQ-C30 was undertaken prior to the analysis. Since the scales should all measure the same construct, they should have internal consistence, and be correlated to each other [30, 31]. Internal consistency reliability for each of the EORTC QLQ-C30 scales was assessed using the Cronbach’s Alpha. This measures the overall reliability and compares the scales hypothesized to measure the same construct of HRQoL. The recommended standard for group comparison (Cronbach’s Alpha coefficient ≥ 0.70) was adopted [28, 30, 31].

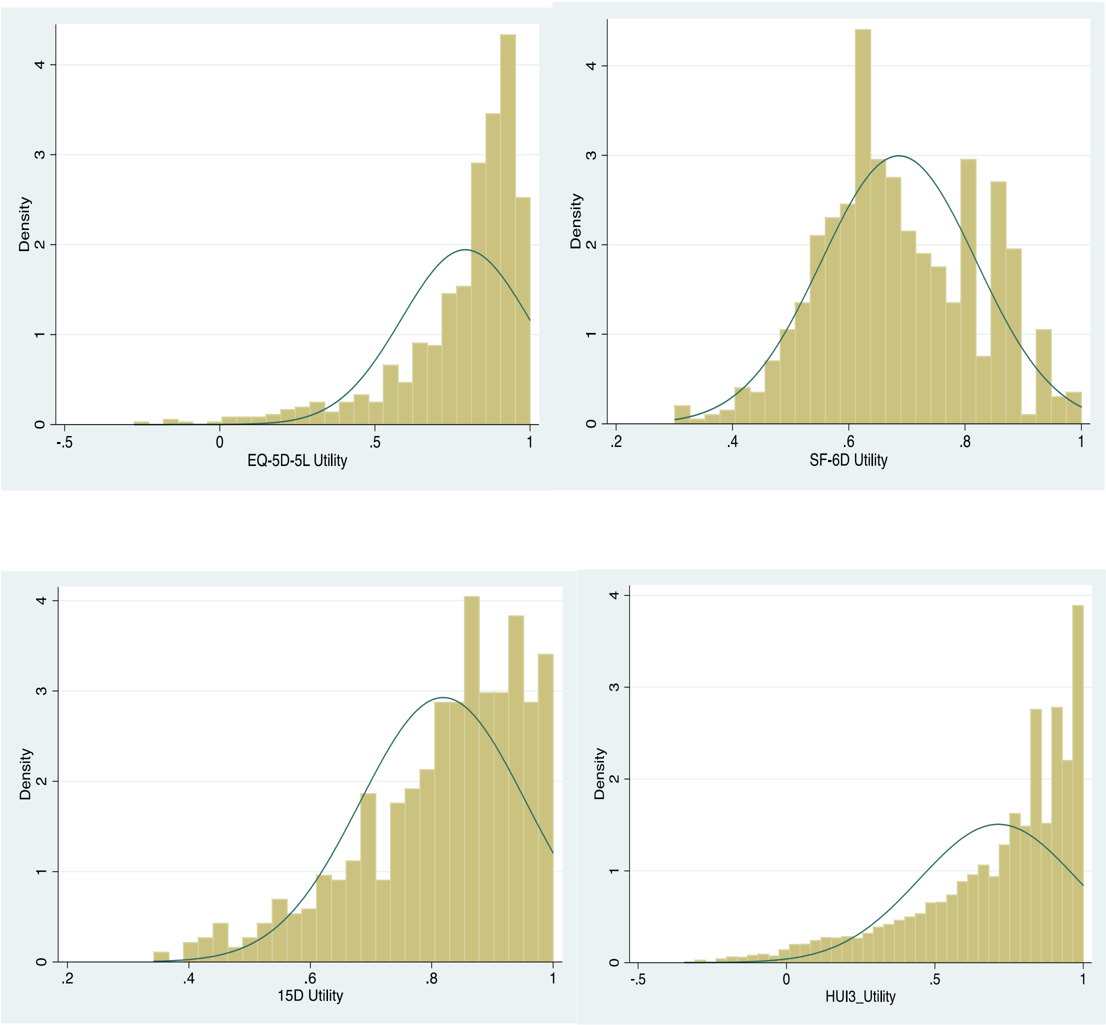
The model was also evaluated for convergent validity, a type of construct validity that looks at the extent of correlation among several measures of the same construct. Convergence validity for the EORTC QLQ-C30 and the four HSU instruments were assessed using Spearman's correlations. However, only results for the EORTC QLQ-C30 and SF-6D are presented and discussed in the paper. The SF-6D has more similar dimensions (e.g. physical, role and social functions) with the EORTC QLQ-C30 compared to the other three HSU instruments. The hypothesis was that scales measuring the same construct of HRQoL, for example physical function scales, should have a high correlation while scales measuring different constructs, for example physical function and cognitive function should have low correlations.

3.4 Statistical Analysis

All the measures included in the study (EQ-5D, SF-6D, 15D, HUI3 and EORTC QLQ-C30) were described using descriptive statistics such as mean, median and range. The distributions of each instrument are depicted in Figure 4. Except for SF-6D, the other three HSU instruments were highly skewed to the left.

The ordinary least square regression method (OLS) is the most widely used method for mapping disease specific instruments onto HSU instruments [7, 32]. Considering that data for HSU instruments have an upper bound of 1, most studies on mapping violates the underlying OLS assumptions of normally distributed errors. However, based on its robustness, the OLS was adopted as the primarily model used to model the EQ-5D, SF-6D, 15D and HUI3 using the scales/items of the EORTC QLQ-C30 as predicting variables. The functional form of the OLS model used was additive, implying that it assumed linear independence between the predictor variables.

Figure 4. Distributions for the EQ-5D, SF-6D, 15D and HUI3 utilities.



Unlike the OLS, the generalized linear regression model (GLM) allow a skewed distribution (i.e., non-normal distribution) of the dependent variable. GLM was applied to compare model performance against the OLS method. The family and link function for GLM was chosen based on the distribution of the data. The selection of the model was also based on which family and link function produced a better prediction [22]. The Gamma family and log link function fitted the data well.

To help select the final set of independent predictors, a stepwise regression technique with forward selection was used for both OLS and GLM. Only statistically significant scales ($p < 0.05$) with economically meaningful coefficients (positive coefficients for global and functional scales and negative coefficients for symptom scales) were included in the final model.

Considering that respondent's demographic characteristics do influence health states measured by HSU instruments, five demographic characteristics (age, gender, education level, marital status and respondents' country) were included as predictor variables. The associations between the demographic characteristics and the HSU instruments were estimated using non-parametric Kruskal-Wallis tests.

3.5 Model Performance

The performance of each model was assessed by the four goodness of fit measures. The first measure is the difference between the predicted mean and observed mean (ΔU) and the second measure is the R^2. The difference between the predicted and observed means is used since most economic studies compare sample means and the R^2 is used to represent the variations in the HSU instruments explained by the models [22]. The higher the R^2, the better the model fit. The third and fourth measures are the mean absolute error (MAE) and the root mean squared error (RMSE). The MAE is the average of the absolute differences between the observed and predicted values while the RMSE is the root of the average of the squared differences [33]. The smaller the values for MAE and RMSE, the better the model performance. Since MAE and RMSE are affected by the scale of the outcome variable, they were normalized by dividing each error by the observed range. The normalized MAE (%MAE) and normalized RMSE (%RMSE) allows for the comparison of models with different scales. Since no consensus exists on the best measure to judge model performance the normalized MAE and RMSE are considered in this study.

Ideally, model performance should be evaluated on data sets that were not used to build the primary model. Doing so provides an unbiased sense of model effectiveness. However, due to the absence of an external validation data set, three cross fold validation data sets were generated and used to estimate and validate the algorithms used for the whole sample. All statistical analyses were conducted using Stata® version 14.2 (StataCorp LP, College Station, Texas, USA).

Ethical consideration

This is a joint study between the University of Tromsø (UiT) and the Multi-Instrument Comparison (MIC) study. This study is part of the master thesis and therefore permission to carry out the study is granted by the UiT thesis committee. All university regulations pertaining to thesis writing were adhered to. The study had no expected budget.

4. Results

4.1. Descriptives

The socio-demographic characteristics of the respondents (age, gender, education level, marital status and country) and variation in HRQoL across these characteristics are shown in Table 4.

Table 4. Socio-demographic characteristics of respondents (N= 772)

Respondent Characteristics	Frequency	Mean (SD)				
Age Group	**N (%)**	**EORTC QLQ C-30**	**EQ-5D**	**SF-6D**	**15D**	**HUI3**
18-44	116 (15)	0.62 (0.22)	0.77 (0.21)	0.65 (0.12)	0.78 (0.17)	0.63 (0.34)
45-54	142 (18)	0.65 (0.21)	0.76 (0.22)	0.65 (0.13)	0.79 (0.14)	0.65 (0.28)
55-64	265 (43)	0.71 (0.20)	0.78 (0.20)	0.69 (0.14)	0.82 (0.12)	0.66 (0.27)
> 65	249(33)	0.75 (0.17)	0.82 (0.18)	0.71 (0.12)	0.84 (0.11)	0.73 (0.23)
P-value*		< 0.001	0.006	< 0.001	0.002	0.016
Gender						
Male	355 (46)	0.71 (0.20)	0.79 (0.20)	0.69 (0.13)	0.82 (0.13)	0.69 (0.27)
Female	417 (54)	0.69 (0.20)	0.78 (0.20)	0.67 (0.13)	0.81 (0.13)	0.67 (0.27)
P-value*		0.281	0.111	0.040	0.580	0.29
Education level						
High school	228 (29)	0.72 (0.19)	0.79 (0.20)	0.69 (0.13)	0.81 (0.13)	0.66 (0.26)
Certificate/Diploma/trade	283 (37)	0.68 (0.21)	0.78 (0.20)	0.67 (0.12)	0.80 (0.14)	0.66 (0.28)
University	261 (34)	0.70 (0.21)	0.79 (0.20)	0.69 (0.13)	0.83 (0.12)	0.71 (0.27)
P-value*		0.021	0.302	0.107	0.063	0.005
Marital status						
Living with spouse/partner	536 (69)	0.71 (0.20)	0.80 (0.19)	0.69 (0.13)	0.82 (0.13)	0.69 (0.27)
Not living with spouse /partner	236 (31)	0.68 (0.21)	0.76 (0.22)	0.66 (0.13)	0.80 (0.13)	0.63 (0.27)
P-value*		0.066	0.009	0.013	0.010	0.002
Country						
Australia	154 (20)	0.71 (0.20)	0.80 (0.17)	0.68 (0.12)	0.81 (0.12)	0.68 (0.20)
USA	148 (19)	0.72 (0.19)	0.78 (0.21)	0.68 (0.13)	0.81 (0.14)	0.68 (0.19)
UK	137 (18)	0.68 (0.21)	0.74 (0.24)	0.66 (0.13)	0.79 (0.13)	0.61 (0.21)
Canada	138 (18)	0.75 (0.18)	0.81 (0.18)	0.70 (0.12)	0.83 (0.13)	0.71 (0.18)
Norway	80 (10)	0.74 (0.18)	0.85 (0.17)	0.72 (0.12)	0.87 (0.10)	0.77 (0.18)
Germany	115 (15)	0.59 (0.23)	0.76 (0.20)	0.66 (0.13)	0.79 (0.16)	0.63 (0.23)
P-value*		< 0.001	< 0.001	< 0.001	< 0.001	0.006

* - Kruskal-Wallis test

All respondents (N=772) who completed questionnaires for the EORTC QLQ-C30 and the four HSU instruments (EQ-5D, SF-6D, 15D, HUI3) were included in the study. The majority of the respondents in the survey are women (54%) and older than 55 years (76%). The HRQoL measured by all the instruments under consideration differs across all age groups with older respondents (>65 years) reporting better HRQoL scores than those <65 years ($p < 0.05$). No gender differences are observed except for SF-6D ($p = 0.04$) where males reported higher HRQoL scores compared to females. The HRQoL reveal significant difference in educational characteristics when measured with the EORTC QLQ-C30 ($p = 0.02$) and HUI3 ($p = 0.05$) whilst the rest of the measures show no difference in educational characteristics among the respondents. Majority of the respondents are married or living with their partners (69%). All HRQoL significantly differs with marital status except for the EORTC QLQ-C30 ($p = 0.06$). Respondents who are married or living with their partners reported better HRQoL scores compared to those who are not married or not living with partners. The proportion of respondents from the six survey countries are as follows: Australia 20%, USA 19%, UK and Canada both had 18%, Germany 15% and Norway 10%. For all the measures included in the study, the HRQoL significantly differs across all the six countries ($p < 0.01$). Respondents from Norway reported higher HRQoL scores for all the four HSU instruments whereas respondents from UK reported lower HRQoL scores across the four HSU instruments. As for the EORTC QLQ-C30 instrument, the highest HRQoL scores are reported by respondents from Canada and the lowest HRQoL scores are reported by respondents from Germany (59).

Table 5 presents the descriptive statistics for the EQ-5D, SF-6D, 15D, HUI3 and the global and functional scales for the EORTC QLQ-C30. Comparing among the four HSU instruments, the lowest observed mean utility index is 0.680 for HUI3 and the highest is 0.818 for 15D. For the EQ-5D and HUI3, the range of utilities is between -0.276 and 1.0 and -0.244 and 1 respectively. With regards to the global and functional scales of the EORTC QLQ-C30, the mean score ranges from 57.3 (global function) to 77.6 (cognitive function), signifying the worst and best HRQoL respectively.

Table 5. Summary statistics for EQ-5D, SF-6D, 15D, HUI3 and EORTC QLQ-C30

STATISTICS	UTILITY SCORES				EORTC QLQ C-30					
					Global*	Functional scores*				
	EQ-5D	SF-6D	15D	HUI3	QLQ	PF	RF	EF	CF	SF
Mean (SD)	0.791 (0.20)	0.686 (0.13)	0.818 (0.14)	0.680 (0.27)	57.3(24)	76.5(23)	71.2(30)	69.6(26)	77.6(25)	66.2(31)
Median	0.851	0.673	0.848	0.777	58.3	86.6	83.3	75.1	83.3	66.6
95% CI	0.776 - 0.805	0.676- 0.695	0.809 – 0.828	0.661 – 0.699	55 - 59	74 -78	69 -73	67 -71	75 -79	64 - 68
Range	-0.276 -1.0	0.301 -1.0	0.342 -1.0	-0.244 -1.0	0 - 100	0 - 100	0 - 100	0 - 100	0 - 100	0 - 100

SD – standard deviation; CI – confidential interval; QLQ – Global health status; PF – physical function; RF – role function; EF – emotional function; CF – Cognitive function; SF – Social function; * higher global health score and functional scale scores indicate better HRQoL.

Table 6 reports the descriptives for the symptom scales. The range for symptom scales is from nausea and vomiting (11.3) with the lowest level of symptoms or better HRQoL to fatigue (39.4) with the highest levels of symptoms or worse HRQoL. The mono-item symptom scores range from Diarrheal (14.1) to Insomnia (37.4), indicating the lowest levels of symptoms or better HRQoL and highest levels of symptoms or worse HRQoL respectively.

Table 6. Summary statistics for the symptom scales of EORTC QLQ-C30

	Symptom scales*			Mono-item symptom scales*					
Statistic	FA	PAIN	N/V	DY	SI	AP	CO	DI	FI
Mean (SD)	39.4 (27.4)	34.4 (31.4)	11.3 (21.4)	23.9 (29.2)	37.4 (32.9)	17.1 (28.3)	15.1 (25.6)	14.1 (25.1)	32.2 (34.8)
Median	33.3	33.3	0	0	33.3	0	0	0	33.3
95% CI	37.5 - 41.4	32.2 -36.7	9.8 - 12.9	21.8 - 25.9	35.1 -39.7	15.1 - 19.1	13.3 - 16.9	12.3 - 15.8	29.7 - 34.6
Range	0-100	0 -100	0 - 100	0 - 100	0 - 100	0 - 100	0 - 100	0 - 100	0 - 100

SD – standard deviation; CI – confidential interval; FA – fatigue; N/V – nausea/vomiting; FI – financial impact; * Higher symptoms scale scores indicate higher levels of symptoms or poor HRQoL.

4.2 Validation of EORTC QLQ-C30

The EORTC QLQ-C30 was validated using the internal-consistency reliability (Cronbach's Alpha) and the construct validity tests (convergence and discriminate validity tests). The internal – consistency reliability for all the EORTC QLQ-C30 scales exceed 0.70 which is the recommended standard. The Cronbach's Alpha for the EORTC QLQ-C30 scales range from 0.729 to 0.920 with the lowest and highest scores coming from the cognitive functional scales and global health status scales respectively.

Construct validity for the EORTC QLQ-C30 is demonstrated by statistically significant ($p < 0.001$) spearman's correlation coefficients for all paired scales of the EORTC QLQ-C30 and the SF-6D (Table 7). Constructs measuring the same dimensions of HRQoL were hypothesized to have higher and statistically significant correlations. As an indication of convergence validity, the spearman's correlation coefficients range from 0.64 (between SF-6D- social scale and EORTC QLQ-C30 - social function) to 0.85 (between the physical scale of SF-6D and the physical scale for EORTC QLQ-C30). On the other hand, constructs measuring different dimensions of HRQoL were hypothesized to have low correlations. For instance, the spearman's correlation of 0.29 (between the physical scale of SF-6D and the nausea and vomiting for EORTC QLQ-C30) indicates discriminate validity.

Table 7. Spearman's correlation coefficients between SF-6D and EORTC -QLQ-C30

	EORTC QLQ C-30								
SF – 6D	**QLQ**	**Physical**	**Role**	**Emotional**	**Cognitive**	**Social**	**Fatigue**	**NV**	**Pain**
Physical scale	0.59	0.85	0.66	0.33	0.36	0.56	0.64	0.29	0.61
Role scale	0.57	0.61	0.64	0.41	0.38	0.56	0.61	0.32	0.52
Bodily pain	0.58	0.62	0.61	0.43	0.36	0.54	0.61	0.33	0.83
Vitality	0.67	0.61	0.54	0.56	0.45	0.56	0.73	0.34	0.54
Social scale	0.61	0.54	0.59	0.51	0.42	0.64	0.61	0.41	0.52
Mental scale	0.58	0.37	0.36	0.73	0.41	0.46	0.49	0.33	0.42

P-value < 0.001 for all paired scales of EORTC QLQ-C30 and SF-6D; QLQ – Global health scale;

N/V – Nausea/vomiting

4.3 Mapping Algorithms

Table 8 presents the results of the regression analysis for the OLS and GLM. The global scale and was identified as a significant predictor of utility for all four HSU instruments except for EQ-5D based on GLM ($p > 0.05$). Two functional scales (physical and emotional) were also significant predictors of utility for all four HSU instruments. The role functional scale is identified as a significant predictor of utility for EQ-5D and SF-6D ($p < 0.05$) but not 15D and HUI3 ($p > 0.05$) for both OLS and GLM models. The opposite is true for cognitive scale which is identified as significant for 15D and HUI3 ($p < 0.001$) but not significant for EQ-5D and SF-6D ($p > 0.05$) for both OLS and GLM models.

As for multi-item symptom scales, pain is the only significant predictor for all four mapping algorithms ($p < 0.05$) whereas fatigue is the only significant predictor of utility for SF-6D ($p < 0.001$) for both OLS and GLM models. Nausea and vomiting is not a significant predictor of utility for any of the HSU instruments using either the OLS or the GLM ($p > 0.05$). With regards to the single-item symptom scales, dyspnea, sleep impairment, constipation and financial impairment are all statistically significant for the 15D ($p < 0.05$) for both OLS and GLM models. In the OLS model, financial impairment is also a significant predictor of utility for SF-6D and HUI3 ($p < 0.05$) whereas in GLM, financial impairment is significant for SF-6D ($p < 0.05$) but not HUI3 ($p > 0.05$).

The country dummy (Germany) is significant for all the four mapping algorithms ($p < 0.001$) based on the OLS and GLM models. The country dummy (UK) is significant for EQ-5D and HUI3 ($p < 0.05$) in the OLS model and EQ-5D, SF-6D and HUI3 in the GLM. ($p < 0.05$). The country dummy (Norway) is a significant predictor of utility for 15D in both the OLS and GLM ($p < 0.001$).

Gender is found to be a significant predictor of utility for 15D ($p < 0.05$) in both the OLS and GLM. In the OLS model, marriage is a significant predictor for two mapping algorithms, from EORTC QLQ-C30 to EQ-5D and HUI3 ($p < 0.05$) whereas in the GLM, marriage is significant only for EQ-5D ($p < 0.05$). For both OLS and GLM, age (25 – 34) was significant for SF-6D whilst age (35 – 44) was significant for EQ-5D ($p < 0.05$).

Table 8. Regression models: OLS and GLM results

OLS Predictors	EQ-5D			SF-6D			15D			HUI3		
	Coeff.	SE	P-value	Coeff.	SE	P-value	Coeff.	SE	P-value	Coeff.	SE	P-value
Constant	0.4704	0.0290	< 0.001	0.4863	0.0239	< 0.001	0.5992	0.0252	< 0.001	0.2105	0.0477	<0.001
Global function	0.0006	0.0003	0.013	0.0012	0.0002	< 0.001	0.0010	0.0002	< 0.001	0.0013	0.0003	< 0.001
Physical function	0.0024	0.0003	< 0.001	0.0011	0.0002	< 0.001	0.0014	0.0002	< 0.001	0.0035	0.0004	< 0.001
Role function	0.0005	0.0002	0.016	0.0004	0.0001	<0.006						
Emotional function	0.0016	0.0002	< 0.001	0.0011	0.0001	< 0.001	0.0005	0.0001	0.001	0.0013	0.0003	< 0.001
Cognitive function							0.0009	0.0001	< 0.001	0.0015	0.0003	< 0.001
Pain	-0.0023	0.0002	< 0.001	-0.0008	0.0001	< 0.001	-0.0003	0.0001	0.005	-0.0023	0.0003	< 0.001
Fatigue				-0.0005	0.0002	< 0.001						
Constipation							-0.0004	0.0001	< 0.001			
Dyspnoea							-0.0007	0.0001	< 0.001			
Sleep Impairment							-0.0005	0.0001	< 0.001	-0.0004	0.0002	0.032
Financial Impairment				-0.0003	0.0001	0.001	-0.0003	0.0001	0.002	-0.0004	0.0002	0.044
Germany	0.0597	0.0127	< 0.001	0.0378	0.0075	< 0.001	0.0411	0.0072	< 0.001	0.0656	0.0168	< 0.001
Norway							0.0312	0.0084	< 0.001			
UK	-0.0279	0.0116	0.016							-0.0441	0.0153	0.004
Age (25 - 34)				-0.0254	0.0128	0.048						
Age (35 - 44)	0.0314	0.0148	0.035									
Gender (female)							0.0130	0.0051	0.011			
Married	0.0193	0.0094	0.036							0.0262	0.0124	0.034

GLM Predictors	EQ-5D			SF-6D			15D			HUI3		
	Coeff.	SE	P-value	Coeff.	SE	P-value	Coeff.	SE	P-value	Coeff.	SE	P-value
Constant	-0.5725	0.0556	< 0.001	-0.7046	0.0353	< 0.001	-0.5294	0.0321	< 0.001	-1.6921	0.1116	< 0.001
Global function				0.0015	0.0002	< 0.001	0.0012	0.0002	< 0.001	0.0022	0.0009	0.026
Physical function	0.0041	0.0005	< 0.001	0.0017	0.0003	< 0.001	0.0021	0.0002	< 0.001	0.0081	0.0010	< 0.001
Role function	0.0010	0.0004	0.013	0.0007	0.0002	< 0.001						
Emotional function	0.0030	0.0004	< 0.001	0.0017	0.0002	< 0.001	0.0007	0.0002	0.001	0.0040	0.0009	< 0.001
Cognitive function							0.0014	0.0002	< 0.001	0.0044	0.0009	< 0.001
Pain	-0.0035	0.0004	< 0.001	-0.0013	0.0002	< 0.001	-0.0004	0.0002	0.010	-0.0047	0.0007	< 0.001
Fatigue				-0.0006	0.0002	0.008						
Constipation							-0.0006	0.0002	0.001			
Dyspnoea							-0.0009	0.0001	< 0.001			
Sleep Impairment							-0.0007	0.0001	< 0.001			
Financial Impairment				-0.0004	0.0001	0.001	-0.0003	0.0001	0.010			
Germany	0.1068	0.0252	< 0.001	0.0579	0.0111	< 0.001	0.0513	0.0103	< 0.001	0.1554	0.0500	0.002
Norway							0.0390	0.0121	0.001			
UK	-0.0517	0.0226	0.022	-0.0218	0.0101	0.032				-0.1003	0.0451	0.026
Age (25-34)				-0.0367	0.0186	0.048						
Age (35-44)	0.0679	0.0291	0.020									
Gender (female)							0.0167	0.0073	0.023			
Married												

4.4 Model performance

The mean and range for the predicted utilities for the four HSU instruments (EQ-5D, SF-6D, 15D and HUI3) were compared with the mean and range for the observed utilities (Table 9). The OLS model generated identical mean utility values to the observed values for EQ-5D, SF-6D, and 15D but over-predicted for HUI3 (0.680 vs 0.681). In the GLM model, only the estimated SF-6D and 15D algorithms generated identical mean utility values to the observed values. GLM over-predicted the mean utility values for EQ-5D (0.791 vs 0.793) and HUI3 (0.680 vs 0.692). The mean difference between the reported and predicted utility values for all four mapping algorithms are small and range from a minimum of 0.001 (mapping from EORTC QLQ-C30 to HUI3 using OLS) to a maximum of 0.012 (mapping from EORTC QLQ -C30 to HUI3 using GLM estimates).

Table 9 shows results for model performance based on the R^2, MAE and RMSE. With regards to the goodness of fit of the regression models measured by the R^2, all the four mapping algorithms generated high explanatory powers with R^2 for OLS ranging from 0.667 (EQ-5D) to 0.762 (15D) and for GLM ranging from 0.601 (HUI3) to 0.753 (15D).

Table 9. Summary results of goodness of fit statistics

HSU instruments	R2		Predicted Utilities (min-max)		OLS		GLM		OLS		GLM	
	OLS	GLM	OLS	GLM	MAE	%MAE	MAE	%MAE	RMSE	%RMSE	RMSE	%RMSE
EQ-5D	**0.665**	0.611	0.791 (0.279-1.082)	0.793 (0.341-1.188)	0.082	**6.426**	0.096	7.523	0.119	**9.326**	0.131	10.266
SF-6D	0.706	**0.727**	0.686 (0.346-0.898)	0.686 (0.396-0.929)	0.058	8.297	0.056	**8.011**	0.072	10.300	0.072	**10.014**
15D	**0.762**	0.753	0.818 (0.441-1.026)	0.818 (0.487-1.058)	0.048	**7.294**	0.049	7.446	0.067	**10.182**	0.068	10.334
HUI3	**0.678**	0.597	0.681 (-0.063-1.073)	0.692 (0.134-1.424)	0.116	**9.324**	0.149	11.977	0.157	**12.540**	0.189	15.112

HSU- Health state utility; OLS - Ordinary least squares; GLM – Generalized linear model; R^2 – coefficient of determination; MAE– Mean absolute error; RMSE – Root mean squared error; %RMSE - Normalized RMSE. %MAE – Normalized mean absolute error; Best results are in bold type.

Therefore, based on R^2, the best and least fitting model is 15D (0.762) using the OLS and the least fitting model is HUI3 (0.601) using GLM. As for the predictive ability of the models reported in terms of %MAE, the least error predictions are generated by the EQ-5D (6.426%) using OLS model and the highest error predictions are generated by the HUI3 (11.977%) using GLM model. In terms of %RMSE, the smallest error values are generated by the EQ-5D algorithm (9.326%) using OLS model and the highest %RMSE values are generated by HUI3 algorithm (15.122%) using GLM model.

Figure 5 illustrates the scatter plots for the predicted vs observed utilities. The OLS and GLM over-predicted at the lower and upper end for EQ-5D utilities whereas both the OLS and GLM over-estimated at the lower end but under-estimated at the upper end for SF-6D utilities. Both models over-predicted the 15D and HUI3 utilities at both lower and upper ends. For all mapping algorithms, the over-prediction by GLM estimates is greater for lower end utilities (representing respondents with poor health) and upper end utilities (representing respondents with better health) compared to OLS estimates. Consistent with the findings in table 9, the over-prediction by OLS model impacts less on the mean utility values compared to the over-prediction by GLM estimates.

4.5 Model Validation

To measure the generalizability of the results, the OLS and GLM models were re-estimated on three within sample validation data sets and the results are presented in Table 10. The reported results are the average for the three cross fold validation data sets. The observed mean utilities for the four HSU instruments fall within the 95% confidence interval of the predicted utilities. For all the four HSU instruments, the mean difference between the reported and predicted utilities did not exceed the minimal clinically important difference (MCID) of 0.03[1].

[1] Minimal clinically important difference (MCID) is defined as the smallest score difference which the respondents perceive as beneficial (Jaeschke *et. al.*1989). The 0.03 is the EQ-5D-MCID but has been reported in literature for SF-6D and 15D (Walters *et. al.* 2005).

Figure 5. Scatter plots for the reported versus predicted utilities

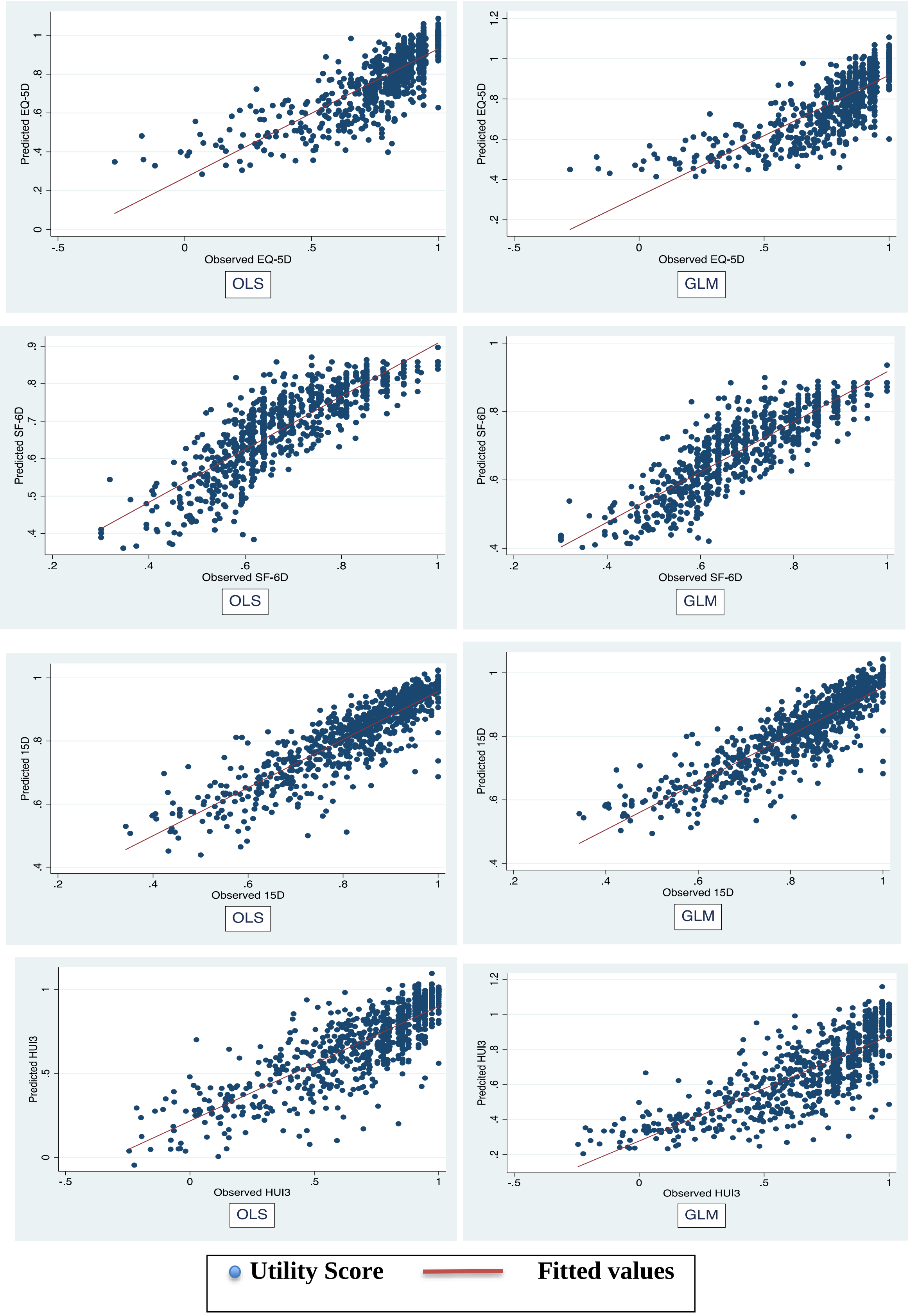


Table 10. Reported versus predicted mean utilities based on cross validation

EQ-5D			**SF-6D**		
	Predicted			**Predicted**	
Observed	**OLS (95% CI)**	**GLM (95% CI)**	**Observed**	**OLS (95% CI)**	**GLM (95% CI)**
0.796	0.791 (0.780, 0.802)	**0.793** (0.781, 0.805)	0.680	**0.686** (0.678, 0.693)	**0.686** (0.678, 0.693)
15D			**HUI3**		
	Predicted			**Predicted**	
Observed	**OLS (95% CI)**	**GLM (95% CI)**	**Observed**	**OLS (95% CI)**	**GLM (95% CI)**
0.818	**0.818** (0.810, 0.826)	**0.818** (0.810, 0.827)	0.680	**0.686** (0.671, 0.709)	0.690 (0.678, 0.693)

OLS - Ordinary least squares; GLM – Generalized linear model; CI – confidence interval; Best results are in bold type.

Table 11 shows the predictive model performance based on the R^2, %MAE and %RMSE. Based on the average for the three samples, the R^2 range from 0.607 (EQ-5D using GLM estimates) to 0.749 (15D using OLS estimates), indicating the model with the lowest and highest explanatory powers respectively. The %MAE range from 6.5% (EQ-5D using OLS estimates) to 10.4% (HUI3 using GLM estimates), representing the least and highest error prediction respectively. The %RMSE range from 10.0% (EQ-5D for both OLS and GLM) to 13.9% (HUI3 using GLM). Consistent with the initial findings, the OLS is the preferred model for mapping the scores of EORTC QLQ-C30 onto HSU instruments except for SF-6D.

Table 11. Predictive model performance

Utility instrument	R^2		MAE (%MAE)		RMSE (%RMSE)	
	OLS	GLM	OLS	GLM	OLS	GLM
EQ-5D	**0.653**	0.607	**0.083 (6.5)**	0.093 (7.2)	**0.128 (10.0)**	0.128 (10.0)
SF-6D	0.704	**0.706**	0.059 (8.4)	**0.056 (8.0)**	0.073 (10.4)	**0.071 (10.1)**
15D	**0.749**	0.735	**0.050 (7.5)**	0.051 (7.7)	**0.069 (10.4)**	0.069 (10.4)
HUI3	**0.664**	0.612	**0.121 (9.7)**	0.130 (10.4)	**0.161 (12.9)**	0.173 (13.9)

OLS - Ordinary least squares; GLM – Generalized linear model; R^2 – coefficient of determination; MAE– Mean absolute error; RMSE – Root mean squared error; %RMSE - Normalized RMSE. Best results in bold type.

Using the OLS and GLM estimates in Table 8, statistically significant scales ($p < 0.05$) with economically meaningful coefficients can be used to generate algorithms mapping EORTC

QLQ-C30 to EQ-5D, SF-6D, 15D and HUI3. OLS is the preferred model to predict all HSU instruments except SF-6D, where GLM provide the best predictive performance (Table 12).

Table 12. Best fitting mapping algorithms

Mapping Algorithms
OLS FOR EQ-5D *= 0.4143 + 0.0008 × QL + 0.0026 × PF+ 0.0006 × RF +0.0018 × EF - 0.0024 × PAIN + 0.0611 × Germany – 0.0277 × UK + 0.0320 × Age (35-44) +0.0198 × Married*
GLM FOR SF-6D *= -0.0007 + 0.0016 × QL + 0.0018 × PF + 0.0007 × RF + 0.0019 × EF - 0.0013 × PAIN – 0.0007 × FATIGUE - 0.0004 × FI + 0.0560 × Germany + -0.0207 × UK - 0.0407 × Age (25-34)*
OLS FOR 15D *= 0.5992 + 0.0010 × QL + 0.0014 × PF + 0.0005 × EF + 0.0009 × CF -0.0003 × PAIN – 0.0007 × DY - 0.0005 × SI- 0.0004 × CO – 0.0003 × FI + 0.0411 × GERMANY + 0.0312 × NORWAY + 0.0130 × FEMALE*
OLS FOR HUI3 *= 0.0101 + 0.0016 × QL + 0.0039 × PF + 0.0015 × EF + 0.0018 × CF - 0.0026 × PAIN – 0.0004 × FI + 0.0670 × Germany – 0.0447 × UK + 0.0268 × MARRIAGE*

QL - Global health; PF - Physical function; EF - Emotion function; RF - Role function; CF - Cognitive function; NV - Nausea and vomiting; SI - Insomnia; DY-Dyspnea; FI - Financial impairment; CO – constipation

5. Discussion

The objective of this study is to develop mapping algorithms that could be used to predict utilities when HSU instruments are not included in a study, such as in clinical trials where the cancer specific EORTC QLQ-C30 instrument is often preferred. To do this, the relationship between the cancer specific EORTC QLQ-C30 instrument and the four widely used HSU instruments (EQ-5D, SF-6D, 15D and HUI3) was estimated using two regression techniques (OLS and GLM). Four goodness of fit measures ($\Delta U,$ R^2, MAE, and RMSE) were used to assess the predictive ability of the algorithms. The preferred algorithm is the one predicting near accurate mean utility values and generating the least predictive errors in terms of %MAE and %RMSE.

Model performance varied between the OLS and GLM across the four goodness of fit measures. The OLS model generated predicted mean utility values which were identical to the observed except for HUI3 while the GLM accurately predicted the 15D and SF-6D utilities but over-predicted the utilities for EQ-5D and HUI3. However, for all four HSU instruments the range of predicted utilities generated by the GLM were narrower compared to the OLS model. Judging by the R^2, %MAE and %RMSE, the OLS estimates produced the highest R^2, lowest %MAE and lowest %RMSE thus making it the preferred model.

Based on the difference between the predicted and observed mean utility values, the mapping algorithms for 15D accurately predicted the mean utility using both OLS and GLM whereas the mapping algorithms for HUI3 over-predicted the mean utilities using both the OLS and GLM. Therefore, judging by the ΔU, the algorithm mapping from EORTC QLQ-C30 to 15D would be the most preferred and the algorithm mapping from EORTC QLQ-C30 to HUI3 would be the least preferred. Overall, the observed values for all the four HSU instruments were within the 95% CI of the predicted values. Moreover, the difference in utility between the predicted and observed for all the four HSU instruments were less than the MCID of 0.03, an indication that in the absence of a

HSU instrument, these mapping algorithms can be used to map the EORTC QLQ-C30 onto any of the four HSU instruments without significantly compromising the results of the intended CUA. Comparable results are reported in other studies that used the OLS method to map scores of EORTC QLQ-C30 onto HSU instruments. In Kontodimopolous *et. al.* [32], the 15D algorithm predicted the closest mean utility values to the observed. Further literature search for comparable studies mapping several HSU instruments from the scores of EORTC QLQ-C30 yielded no results as most of the studies involved either mapping a single HSU instrument to EORTC QLQ-C30 or studies comparing several HSU instruments to disease specific instruments other than the EORTC QLQ-C30.

In terms of the variations explained by the models (R^2), all the HSU instruments had relatively high explanatory powers ($R^2 > 0.50$), with 15D generating the highest R^2 (0.762) using OLS and the HUI3 generating the lowest R^2 (0.601) using GLM. These results are within the range of previously published studies. For instance, Kontodimopolous *et. al*. [32], found 15D with the highest R^2 (0.90) compared to EQ-5D (0.61) and similar to this study, the global score, physical and cognitive functions were significant predictors for the 15D.

However, for the purposes of comparing the predicted abilities of the mapping algorithms, ΔU and R^2 are not the most suitable. The ΔU is unsuitable because the mean utilities from different HSU instruments differ and high R^2 does not imply good predictive ability of the models. For this reason, comparative performance of the mapping algorithms cannot be determined by the ΔU and R^2. The preferred methods are the %MAE and %RMSE, which allows for the comparison of models with different scales.

The study found considerable variations in the error predictions across the mapping algorithms, with the algorithm mapping the EQ-5D from the EORTC QLQ-C30 generating the lowest %MAE (6.4%) using the OLS and the algorithm mapping the HUI3 from the EORTC QLQ-C30 generating the most predictive errors (11.9%) using the GLM. The %RMSE showed similar variations in the error predictions with the least error prediction for EQ-5D (9.3%) and the highest error predictions for HUI3 (15.1%) using the OLS and GLM respectively. Therefore, on a comparative basis and judging by the %MAE and %RMSE, the most preferred-mapping algorithm would be from EORTC QLQ-C30 to EQ-5D and the least preferred would be from EORTC QLQ-C30 to HUI3.

These findings are inconsistent with the results from the only study [32] that map the scores of the EORTC QLQ-C30 onto the same HSU instruments reported here. Using the OLS model, Kontodimopolous *et. al*. [32] found the lowest %RMSE (5.4%) when mapping the EORTC QLQ-C30 onto the SF-6D, whilst the highest %RMSE (12%) was with the prediction of the EQ-5D utility. In this study, the EQ-5D-algorithm produced the lowest %RMSE (6.4%) compared to SF-6D (10.3) and 15D (10.1). This inconsistency can be attributed to the fact that the sample size in the study by Kontodimopolous *et.al*.[32] was small (N=48) and limited to gastric cancer, whereas the sample size in the present study was relatively large (N= 772) and consisted of a heterogenous group of cancer patients.

The mapping algorithms compared in the study yielded different results because of the differences in the scaling properties of HRQoL instruments. These differences arise from differences in their descriptive systems and in the methods used to obtain utility scores. Such differences cause HRQoL instruments to generate significantly different utilities and error predictions for the same population [16, 32, 34-36]. The overlap between the classification systems and the utility weights for HSU instruments and EORTC QLQ-C30 plays a key role in how accurate models predict utilities. For instance, HSU instruments with similar dimensions to EORTC QLQ-C30 tend to produce more accurate predictions whereas those whose important dimensions are not covered by the EORTC QLQ-C30 tend to undermine the model [37].

The EORTC QLQ-C30 was developed specifically for cancer patients and contains dimensions that focus mostly on the symptoms of cancer and its treatment rather than HRQoL. On the other hand, HSU instruments are applicable to all patient groups and contain a combination of dimensions covering both HRQoL and symptoms. On the functional scale level, all the HSU instruments appear to be well covered by the functional dimensions of the EORTC QLQ-C30. For instance, the physical, emotional and role scales were statistically significant predictors when mapping EORTC QLQ-C30 to EQ-5D and SF-6D whereas for the 15D and HUI3, it was the physical, emotional and cognitive scales that were statistically significant ($p < 0.05$).

However, the EQ-5D and HUI3 do not seem to be adequately covered by the symptom scales of the EORTC QLQ-C30 as the SF-6D and 15D. For instance, except for the pain dimension, the EQ-5D - the most widely used HSU instrument and HUI3 have no other symptom dimension to specifically capture the spectrum of symptoms experienced by cancer patients. Consequently, among the symptom dimensions of the EORTC QLQ-C30, only pain scale was a significant predictor ($p < 0.05$) after mapping EORTC QLQ-C30 onto EQ-5D. On the other hand, the mapping of EORTC QLQ-C30 scores on to 15D yielded four significant symptom scales (Insomnia, Dyspnea, Constipation and financial impairment) ($p < 0.05$). This is because the 15D is the most comprehensive in terms of dimensions included and is most sensitive to the symptoms associated with cancer [10].

Since inadequacies in the utility weighting and classification system of the HSU instruments cannot be overcome by mapping, caution needs to be taken when comparing mapping algorithms. In additional, mapping should be undertaken only if the instruments are appropriate for that condition and patient population [37]. Algorithms generating large errors could indicate their inappropriateness for use in heterogenous group of cancer patients.

When it comes to the question of generalizability, it should be noted that the primary focus of this study was to map the cancer EORTC QLQ-C30 onto the four most widely used HSU instruments

(EQ-5D, SF-6D, 15D and HUI3) using data from six countries. Therefore, caution should also be taken in generalizing these results outside of this type of mapping and context.

Like most studies, there are data and methodological limitations with this study. Firstly, data used in this study were obtained from an online survey and like most online data collection methods, respondents who were too ill to complete the survey were excluded. Secondly, there was no external independent dataset to use for validating the mapping algorithms. Fortunately, the psychometric properties of the EORTC QLQ-C30 have been tested and its reliability and validity confirmed in several studies [28, 30, 38]. Therefore, in this study the validation of the EORTC QLQ-C30 served an ad hoc purpose. Whenever possible, the mapping algorithms should be validated against an external sample that is similar. However, such external sample was not available for use.

Methodological limitations include the use of only two regression models (OLS and GLM) for mapping EORTC QLQ-C30 to the four HSU instruments. The OLS method is the most widely used method for mapping DSI onto HSU instruments [34, 39]. However, the OLS models, like most other linear prediction models, suffer from several limitations. Obvious limitations include their poor predictive abilities for low and high utilities values [33, 39, 40]. This is evidenced in Figure 6., where there was clear over prediction for EQ-5D and 15D and underprediction for SF-6D and HUI3 for low utility values using OLS. The GLM, though flexible with regards to non-normally distributed HSU data poorly handles censored dependent variables such as the EQ-5D-5L.

On the other hand, the study has some strength and the obvious being that (i) the sample size was big enough (N= 772), (ii) it consisted of respondents from six countries, (iii) it consisted of a heterogenous group of cancer patients and (iv), it included main respondent characteristics as predictor variables.

Even though the OLS and GLM performed reasonably well in predicting health state values for the EQ-5D, SF-6D, 15D and HUI3 from EORTC QLQ-C30 scores, owing to the above discussed limitations and the mere fact that mapping DSI onto HSU instruments is a complex undertaking, more robust methods such as the two-part model (TPM), Tobit model, latent class mixture model, multinomial logit model and censored least absolute deviations (CLAD) ought to be considered in future studies. Furthermore, even if the within-sample validation is acceptable, future studies incorporating external independent data-sets are recommended. This is because predictions based on the same respondent sample can produce different results when applied to another sample with different respondent characteristics.

6. Conclusion

The study has significant public health implications considering that most of clinical trials on cancer do not use HSU instruments but instead use DSI such as the EORTC QLQ-C30. The main conclusion of the study is that HSU instruments are different regardless of what criterion is used to compare them. They measure different constructs of HRQoL and thus present different definitions of health. We compared the predictive abilities of two regression models in four HSU instruments (EQ-5D, SF-6D, 15D and HUI3) and based on the normalized MAE and RMSE, the study findings indicate that EQ-5D is the preferred instrument for mapping or cross-walking the scores of EORTC QLQ-C30 using either OLS or GLM.

REFERENCES

1. Prieto, L. and J.A. Sacristan, *Problems and solutions in calculating quality-adjusted life years (QALYs).* Health Qual Life Outcomes, 2003. **1**: p. 80.

2. Mariotto, A.B., et al., *Projections of the cost of cancer care in the United States: 2010-2020.* J Natl Cancer Inst, 2011. **103**(2): p. 117-28.

3. Jorid Kalseth, V.H., Thomas Halvorsen, *Cost of Cancer in the Nordic Countries: A comparative study of health care costs and public income loss compesation payments related to cancer in the Nordic countries in 2007.*, in *Health Services Research*. 2011, Nordic Cancer Union: Norway. p. 98.

4. Parker, M., A. Haycox, and J. Graves, *Estimating the relationship between preference-based generic utility instruments and disease-specific quality-of-life measures in severe chronic constipation: challenges in practice.* Pharmacoeconomics, 2011. **29**(8): p. 719-30.

5. Rowen, D., et al., *Comparison of generic, condition-specific, and mapped health state utility values for multiple myeloma cancer.* Value Health, 2012. **15**(8): p. 1059-68.

6. Rowen, D., et al., *Deriving a preference-based measure for cancer using the EORTC QLQ-C30.* Value Health, 2011. **14**(5): p. 721-31.

7. Chen, G., et al., *Diabetes and quality of life: Comparing results from utility instruments and Diabetes-39.* Diabetes Res Clin Pract, 2015. **109**(2): p. 326-33.

8. Richardson, J., et al., *Can multi-attribute utility instruments adequately account for subjective well-being?* Med Decis Making, 2015. **35**(3): p. 292-304.

9. Feeny, D., et al., *Multiattribute and single-attribute utility functions for the health utilities index mark 3 system.* Med Care, 2002. **40**(2): p. 113-28.

10. Sintonen, H., *The 15D instrument of health-related quality of life: properties and applications.* Ann Med, 2001. **33**(5): p. 328-36.

11. Tsuchiya, A., J. Brazier, and J. Roberts, *Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets.* J Health Econ, 2006. **25**(2): p. 334-46.

12. Lin, X.-J., I.M. Lin, and S.-Y. Fan, *Methodological issues in measuring health-related quality of life.* Tzu Chi Medical Journal, 2013. **25**(1): p. 8-12.

13. Patrick, D.L. and R.A. Deyo, *Generic and disease-specific measures in assessing health status and quality of life.* Med Care, 1989. **27**(3 Suppl): p. S217-32.

14. Brazier, J.E., et al., *Generic and condition-specific outcome measures for people with osteoarthritis of the knee.* Rheumatology (Oxford), 1999. **38**(9): p. 870-7.

15. Neumann, P.J., S.J. Goldie, and M.C. Weinstein, *Preference-based measures in economic evaluation in health care.* Annu Rev Public Health, 2000. **21**: p. 587-611.

16. Richardson, J., et al., *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. Med Decis Making, 2015. **35**(3): p. 276-91.

17. Drummond, M.F., *Methods for the Economic Evaluation of Health Care Programmes.* 2015.

18. Blinman, P., King, M., Norman, R., Viney, R., Stockler, M.R., *Preferences for cancer treatments : an overview of methods and applications in oncology.* Annals of oncology, 2012. **23**: p. 1104 - 1110.

19. Dolan, P., *Chapter 32 The measurement of health-related quality of life for use in resource allocation decisions in health care*, in *Handbook of Health Economics*. 2000, Elsevier. p. 1723-1760.

20. Whitehead, S.J. and S. Ali, *Health outcomes in economic evaluation: the QALY and utilities.* Br Med Bull, 2010. **96**: p. 5-21.

21. Brazier, J., J. Roberts, and M. Deverill, *The estimation of a preference-based measure of health from the SF-36.* J Health Econ, 2002. **21**(2): p. 271-92.

22. Chen, G., et al., *Mapping between 6 Multiattribute Utility Instruments.* Med Decis Making, 2016. **36**(2): p. 160-75.

23. Jeff Richardson , L.A., Maxwell, *A Cross national comparison of twelve quality of life instruments: MIC paper1 background , questions, instruments,.* Centre for Health Economics . Research Paper, in Centre for Health Economics. Monash University: Australia., 2012. **76**.

24. *<EQ-5D-3L_UserGuide_2015.pdf>.*

25. Herdman, M., et al., *Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).* Qual Life Res, 2011. **20**(10): p. 1727-36.

26. Torrance, G.W., *Utility approach to measuring health-related quality of life.* J Chronic Dis, 1987. **40**(6): p. 593-603.

27. Horsman, J., et al., *The Health Utilities Index (HUI): concepts, measurement properties and applications.* Health Qual Life Outcomes, 2003. **1**: p. 54.

28. Aaronson, N.K., et al., *The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology.* J Natl Cancer Inst, 1993. **85**(5): p. 365-76.

29. Blazeby, J.M., et al., *Health-related quality of life measurement in randomized clinical trials in surgical oncology.* J Clin Oncol, 2006. **24**(19): p. 3178-86.

30. Apolone, G., et al., *Evaluation of the EORTC QLQ-C30 questionnaire: a comparison with SF-36 Health Survey in a cohort of Italian long-survival cancer patients.* Ann Oncol, 1998. **9**(5): p. 549-57.

31. Bland, J.M. and D.G. Altman, *Cronbach's alpha.* Bmj, 1997. **314**(7080): p. 572.

32. Kontodimopoulos, N., et al., *Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments.* Value Health, 2009. **12**(8): p. 1151-7.

33. Kim, S.H., et al., *Mapping EORTC QLQ-C30 onto EQ-5D for the assessment of cancer patients.* Health Qual Life Outcomes, 2012. **10**: p. 151.

34. Arnold, D.T., et al., *Testing mapping algorithms of the cancer-specific EORTC QLQ-C30 onto EQ-5D in malignant mesothelioma.* Health Qual Life Outcomes, 2015. **13**: p. 6.

35. McKenzie, L. and M. van der Pol, *Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data.* Value Health, 2009. **12**(1): p. 167-71.

36. Richardson, J., et al., *Measuring the Sensitivity and Construct Validity of 6 Utility Instruments in 7 Disease Areas.* Med Decis Making, 2016. **36**(2): p. 147-59.

37. Young, T.A., et al., *Mapping Functions in Health-Related Quality of Life: Mapping from Two Cancer-Specific Health-Related Quality-of-Life Instruments to EQ-5D-3L.* Med Decis Making, 2015. **35**(7): p. 912-26.

38. Fayers, P. and A. Bottomley, *Quality of life research within the EORTC-the EORTC QLQ-C30. European Organisation for Research and Treatment of Cancer.* Eur J Cancer, 2002. **38 Suppl 4**: p. S125-33.

39. Brazier, J.E., et al., *A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures.* Eur J Health Econ, 2010. **11**(2): p. 215-25.

40. Ralph Crott, A.B., *Mapping the QLQ C-30 quality of life cancer questionnaire to EQ-5D patient preferences.* European Journal of Health Economics, 2010. **11**: p. 427 - 434.