Modelling Health-Related Quality of Life Data for Economic Evaluation of Cancer Treatments:
Applications in Lung Cancer

Iftekhar Khan

A thesis submitted for the degree of

Doctor of Philosophy

at the

UCL (University College London)

Department of Applied Health Research

Declaration of Authorship

I, Iftekhar Khan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Signed: ____

Date: ____

Abstract

Introduction: The annual economic burden of treating cancer to the National Health Service (NHS) in the United Kingdom (UK) is over £15 billion; and for non small cell lung cancer (NSCLC), one of the leading causes of cancer deaths in the world, this is £2.4 billion. Economic evaluation plays an essential role in assessing the relative value of lung cancer treatments. Modelling (HRQoL) data is fundamental in determining the cost-effectiveness of cancer treatments. This thesis aims to investigate modelling of HRQoL data collected from lung cancer patients for economic evaluation. In particular, the role of modelling to improve utility prediction is investigated. The sensitivity of disease specific and generic HRQoL measures are also explored. In addition, methods to extrapolate utilities beyond cancer progression and identifying a selection procedure from relevant published algorithms are developed.

Methods: Data from two clinical trials and a prospective observational study in NSCLC patients were designed and executed to develop several mapping models (Linear, Non-Linear, Joint, and Bayesian). The sensitivity of EQ-5D-3L and EQ-5D-5L were compared with a cancer specific measure (QLQ-C30). Simulation methods were used to develop an approach for selecting algorithms.

Results: Two and three-part Beta-Binomial models improve predictions. Joint models also contribute to improved prediction of utilities. Bayesian Networks may help reduce the over-prediction in poor health states. The EQ-5D-5L offers better mapping and is more sensitive for detecting treatment benefit compared to EQ-5D-3L. It is also viable to develop decision criteria for selecting between several published algorithms.

Conclusion: Methodological improvements in modelling HRQoL for the economic evaluation of cancer treatments have been demonstrated. Improvements in model structure, prediction and selection are empirically demonstrated.

Table of Contents

List of Tables and Figures

Acknowledgements

Firstly, I would like to thank God Almighty for His Grace in completing this thesis. I would further extend my gratitude to Professor Steve Morris, who supported my proposal and to Dr Nora Pashayan, who agreed to co-supervise. Their guidance, review, criticisms, and suggestions have been most valuable for improving this work. I would also like to acknowledge Dr Joe Maguire and Dr Bashir Matata at the Liverpool Hospital for working with me on the design of the prospective component of this study and gathering the valuable data. I must also show my appreciation to Professor Allan Hackshaw and Professor Jonathan Lederman of the UCL, Cancer Research UK, Cancer Trials Centre, who despite our differences, permitted and encouraged completion of this Ph.D. Additional recognition is due to Dr Suhailah Khan, Yohanis, and Hanzalah Khan, for seeing yet another thesis through to its completion.

List of Publications Associated with this Thesis

1. 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 (Khan. I and Morris. S, 2014, Health and Quality of Life Outcomes 2014 12:163.). My contribution was the concept, analysis and writing the publication. This paper is based on Chapter 4 of this thesis. Oral presentation at the International Society for Clinical Trials (2014) – Health Economics Theme. This is also a chapter of a book: Chapter 5 of Design & Analysis of Clinical trials for Economic evaluation and Re-imbursement; Chapman & Hall (2016); 313 pages (Iftekhar Khan) [Chapter 4].

2. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients (Khan. I, Morris. S, Pashayan. N, Matata. B, Bashir. Z and Maguirre. J; Health and Quality of Life Outcomes 201614:60). My contribution was to provide the concept including methods for design of the prospective observational study analysis, data management, statistical analyses, writing the publication and responding to reviewer comments. This paper was based on Chapter 5 of this thesis [Chapter 5].

3. *Interpreting small treatment differences from the quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30* (Khan. I**,** Bashir. Z and Forster. M; Health and Quality of Life Outcomes 201513:180.) My contribution was the concept, analysis and write up [Concepts developed in Chapters 4 and 8].

List of Abbreviations and Acronyms

AE	Adverse Events
AIC	Aikakes Information Criterion
ALVDMM	Adjusted Limited Variable Dependent Mixture Model
AP	Appetite Loss
AUC	Area Under the Curve
AX	Anxiety Domain
BB	Beta-Binomial
BN	Bayesian Network
BSC	Best Supportive Care
CDF	Cancer Drugs Fund
CF	Cognitive Functioning
CHART	Continuous Hyper-fractioned Radiotherapy
Chemo-RT	Combination of RT with Chemotherapy
CLAD	Censored Least Absolute Deviations
CO	Constipation
CRUK	Cancer Research UK
CSM	Condition Specific Measure
CUA	Cost Utility Analysis
DAG	Directed Acyclic Graph
DI	Diarrhoea
DY	Dyspnoea
ECOG	Eastern Cooperative Oncology Group
EE	Economic Evaluation
EED	Economic Evaluation Database
EF	Emotional Functioning
EGFR	Epidermal Growth Factor Receptor
EoL	End of Life
QLQ-C30	EORTC-QLQ-C30
EQ-5D	EQ-5D Measure (3L or 5L)
ES	Effect Size
FA	Fatigue
FACT-L	Functional Assessment of Cancer Therapy-Lung
FI	Financial Impact
FMM	Finite Mixture Models

GHS	Global Score
HAQ	Health Assessment Questionnaire
HAS	French National Authority for Health
HR	Hazard Ratio
HRQoL	Health Related Quality of Life
HSD	Half of the standard deviation
HTA	Health Technology Appraisal
HUI	Health Utilities Index
ICER	Incremental Cost-Effectiveness Ratio
IN	Insomnia
INB	Incremental Net Benefit
ISCTM	International Society of Clinical Trials Methodology
KHQ	King's Health Questionnaire
LC	Lung Cancer
LCS	Lung Cancer Subscale
LDVMM	Limited Dependent Variable Mixture Model
MAE	Mean Absolute Error
MESH	Medical Subject Headings
MVN	Multivariate Normal Distribution
NCCN	National Comprehensive Cancer Network (NCCN)
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NR	Not Reported
NSCLC	Non-Small Cell Lung Cancer
NV	Nausea and Vomiting
OLS	Ordinary Least Squares
OR	Odds Ratios
OS	Overall Survival
PA	Pain
PC	Peter Spirtes and Clark Glymour
PD	Progressive Disease
PDF	Probability Density Function
PF	Physical Functioning
PFS	Progression Free Survival
PP	Post-Progression
PPS	Post Progression Survival
PSA	Probabilistic Sensitivity Analysis

PW	Physical Wellbeing
QALY	Quality-Adjusted Life Year
QL	Global Health Status Score
QoL	Quality of Life
QTWiST	Quality of Time without Signs & Symptoms of Toxicity
QWB	Quality of Wellbeing Scales
RCT	Randomized Controlled Trials
RE	Random Effects
RF	Role Functioning
RMSER	Residual/Root Mean Squared Error Reported
RMSEP	Residual/Root Mean Squared Error Predicted
R-Squared	R-Squared = R2 (Coefficient of Determination)
R2	Predictive Power
R2R	R-Squared Reported
R2P	R-Squared Predicted
RT	Radiotherapy
SCLC	Small Cell Lung Cancer
SD	Standard Deviation
SE	Standard Error
SF	Social Functioning
SL	Sleep Disturbance
SMC	Scottish Medicines Consortium
SOCCAR	A Phase II Trial of Sequential Versus Concurrent Chemotherapy and Radiotherapy Using an Accelerated Hypofractionated Radiation Schedule in Stage III NSCLC
SRM	Standardized Response Mean
SW	Social and family Wellbeing
TOI	Treatment Outcome Index
TOPICAL	Randomized phase III Trial of erlotinib compared with placebo in chemotherapy-naive patients with advanced non-small cell lung cancer (NSCLC) and unsuitable for first-line chemotherapy.

Thesis Overview

The thesis is divided into 10 chapters: **Chapter 1** is an introductory chapter on health related quality of life (HRQoL) in the context of cancer and economic evaluation. **Chapters 2 and 3** consist of a literature search, review and statement of objectives of the thesis. **Chapter 4** compares existing models with a new non-linear Beta-Binomial mapping algorithm, using patient-level data from two randomized trials. **Chapter 5** is an extension of chapter 4 as it evaluates other mapping algorithms developed from the more recent EQ-5D-5L. **Chapter 6** seeks to understand the reasons why algorithms may over-predict at poorer health states. **Chapter 7** involves the use of Bayesian networks in order to develop a mapping algorithm. **Chapter 8** compares the sensitivity and responsiveness of generic and condition-specific measures, particularly EQ-5D-5L, EQ-5D-3L, and EORTC-QLQ-C30. **Chapter 9** compares the performance of published mapping algorithms. A selection procedure is proposed, which separates ‘useful’ algorithms from ‘not useful’ ones. Finally, **Chapter 10** provides a summary and conclusion of the above research and discusses the nuances, advantages, limitations and future research ideas.

Chapter 1

Chapter 1: Cancer Epidemiology, Treatment, Quality of Life and Economic Burden

Abstract

Introduction: The annual economic burden of treating cancer to the National Health Service (NHS) in the United Kingdom (UK) is over £15 billion; and for non small cell lung cancer (NSCLC), one of the leading causes of cancer deaths in the world, this is £2.4 billion. Economic evaluation plays an essential role in assessing the relative value of cancer treatments. The aim of this chapter is to introduce important concepts associated with health-related quality of life (HRQoL) and their importance for the methodology that underpins economic evaluation of cancer treatments.

Methods: Medical Subject Headings (MESH) search terms using PUBMED, MEDLINE, and COCHRANE databases of systematic reviews were used to identify articles for this introductory chapter. A narrative review of the epidemiology, treatments, economic costs and HRQoL associated with lung cancer are presented to contextualize the research aims of this dissertation.

Results: Annual worldwide cancer incidence is around 14.1 million with lung, breast, bowel, and prostate being the most common. Treatment options after surgery are often (expensive) chemotherapies. The worldwide economic burden of cancer is at least $895 billion; in the UK alone, the NHS spending is at least £15 billion. Cancer treatments are expensive and their cost-effectiveness may often depend on HRQoL estimates and their uncertainties. Unavailability of suitable HRQoL measures for cost-effectiveness analyses, use of inadequate modelling methods, lack of sensitivity of instruments, unclear definition of relevant effect sizes and rapid disease progression are some of the challenges for the economic evaluation of cancer treatments.

Conclusion: There are challenges identified with HRQoL in the context of the economic evaluation of cancer treatments and these will be addressed in the thesis.

1.1 Introduction and Epidemiology of Cancer

1.1.1 Introduction to Cancer

The term ‘carcinoma,’ is derived from the Greek word ‘karkinos’, meaning crab. Hippocrates associated cancer to the shape of a crab, because of the way it spreads through the body and its persistent nature [1].
The National Cancer Institute Dictionary of Cancer Terms defines cancer as:

“A term for diseases in which abnormal cells divide without control and can invade nearby tissues. Cancer cells can also spread to other parts of the body through the blood and lymph systems. There are several main types of cancer. Carcinoma is cancer that begins in the skin or in tissues that line or cover internal organs. Sarcoma is cancer that begins in bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue. Leukaemia is cancer that starts in blood-forming tissue, such as the bone marrow and causes large numbers of abnormal blood cells to be produced and enter the blood. Lymphoma and multiple myelomas are cancers that begin in the cells of the immune system. Central nervous system cancers are cancers that begin in the tissues of the brain and spinal cord.” [2]

Cells in the body incessantly split and start spreading into various parts of the body [3]. In a normal body, cells in the body grow and divide to make new cells, as and when the body needs. The older or damaged cells die and new cells are generated. However, if the old cells do not die and simultaneously the new cells are generated, the surplus cells divide and result in tumors [4].

1.1.2 Lung Cancer

An estimated 14.1 million new cases of cancer have been identified across the world in 2012. Amongst these, nearly four in ten cases occur in developing countries. The four most common types of cancers recognized worldwide are lung, female breast, bowel and prostate cancer, respectively. These four types of cancer constitute nearly 40% of all cancers diagnosed worldwide.

Importance of Lung Cancer: Mortality

Lung cancer is one of the leading causes of cancer-related deaths and accounts for nearly 1.4 million deaths per year worldwide, with a yearly incidence of over 41,000 in the UK alone [5, 6]. Of these, ≥80% of incidences are non-small cell lung cancers

(NSCLC) [7]. Over 342,000 people in Europe and 162,000 in the US die each year from lung cancer [8]. In addition, approximately 13% of all estimated new cancer cases and 19% of all cancer-related deaths globally are due to lung cancer [8]. Lung cancer incidence and death rates are lower in Europe compared to those in the United States (USA). The survival rates associated with lung cancers do not vary by gender [6, 7, and 8] and the death rate from the disease remains high, at 56 deaths per 100,000 people in the UK population annually.

More than 8 out of 10 lung cancer cases occur in people aged 60 and over. Rates of lung cancer in Scotland are among the highest in the world, owing to high smoking prevalence. In the 1950s, for every 1 lung cancer case diagnosed in women in the UK, there were 6 in men. That ratio is now 3 cases in women for every 4 in men. The lowest lung cancer rates in the world for men and women are in Northern, Western and Middle African countries and South Central Asia; but this will also change if the current trends in the uptake of smoking persist [8]. These facts underline the importance and significance of lung cancer as an important area for research.

Nine out of ten cases of lung cancer are caused by smoking. In 2002, there were 38,410 new cases and 33,602 deaths from lung cancer [3]. Recent statistics in 2009 show lung cancer cases of 42,000 new cases (18,000 of these women, making it the second most common cancer in women after breast and bowel cancer). The majority of cases are inoperable at presentation. This may be due to medical co-morbidity (e.g. at stages I, II and III), or due to tumour extent (e.g. at stage IV): cancer staging is a way of defining the severity of a patient’s cancer, stage I being milder and IV more advanced.

The median survival of lung cancer patients in the UK is about 203 days (about 7 months) [9]. This also means that the time horizons for assessing HRQoL and costs are also relatively short. Moreover, “Lung cancer costs more than any other cancer – mainly because of potential wage losses due to premature deaths from people in employment - about 60% of the total economic costs – and high health care costs. The death rate from the disease remains high at 56 deaths per 100,000 people in the UK population annually, and almost a quarter of these occur before retirement…” [5].

Finding effective treatments has been challenging, with few that extend survival significantly. It is still, therefore, an incurable disorder, and consequently one of the primary aims as part of patient management should be improving HRQoL particularly towards the end of life. Lung cancer is therefore likely to remain a significant burden of illness in the UK as well as worldwide and healthcare resource utilization is likely to remain high in these patients. This is at least one reason why modelling HRQoL in this population remains an important area of research in economic evaluation.

Importance of Lung Cancer: Morbidity

The lung cancer five-year survival rate (17.7 percent) is lower than many other leading cancer types, such as the colon (64.4 percent) breast (89.7 percent) and prostate (98.9 percent) [10]. The five-year survival rate for lung cancer is 55 percent for cases detected when the disease is still localized (within the lungs). However, only 16 percent of lung cancer cases are diagnosed at an early stage. For distant tumours (spread to other organs) the five-year survival rate is only 4 percent.

Smoking, the main cause of lung cancer, contributes to 80 percent and 90 percent of lung cancer deaths in women and men, respectively. Men who smoke are 23 times more likely to develop lung cancer. Women are 13 times more likely, compared to never smokers.10. Exposure to second hand (passive) smoke causes approximately 7,330 lung cancer deaths among non-smokers every year [11]. Lung cancer can also be caused by occupational exposures, including asbestos, uranium, and coke (an important fuel in the manufacture of iron in smelters, blast furnaces and foundries). The combination of asbestos exposure and smoking greatly increases the risk of developing lung cancer. Lung cancer is also associated with poor HRQoL, sometimes exacerbated due to toxicities from treatments (chemotherapy). These include shortness of breath (dyspnoea), cough, anxiety, depression and a marked impact on daily activities.

Risks associated with lung cancer

Risks associated with lung cancer depend on several factors, including age, genetics, and exposure to other risk factors (e.g. smoking). Smoking, insufficient physical activity, alcohol, diet, being overweight and infections account for a high proportion of cancers worldwide. Prevalence of different risk factors varies by region and country; smoking is the single most preventable cause of cancer-related death in the world; around a third of tobacco-caused deaths are due to cancer. Moreover, drinking excessive alcohol causes an estimated 6% of deaths worldwide and about 13% of these deaths (equivalent to 47 million people) are due to alcohol related cancer [5].

*Types of Lung Cancer (*Histological Classifications)

There are two main types of lung cancer: Small Cell Lung Cancer (SCLC) represents about 20 % of lung cancer cases and Non- Small Cell Lung Cancer (NSCLC) about 80% [3]. They are called SCLC and NSCLC because the cancer cells were found to either be small cells, or larger cancer cells, such as adenocarcinomas or squamous cells. Consequently, they have been classified into those separate categories, small cell or non-small cell cancer [12] .

There are two further types of NSCLC, based on histologic subtyping: squamous and non-squamous, where the latter can be divided into two subtypes, adenocarcinoma and large cell carcinoma as presented in Table 1.1 (below).

Lung cancer type	Histology	Subtype	% of total NSCLC
NSCLC	Non-squamous	Adenocarcinoma	40%
		Large cell carcinoma	10 – 15%
	Squamous	Carcinoma	25 – 30%
	Other	–	15 – 25%

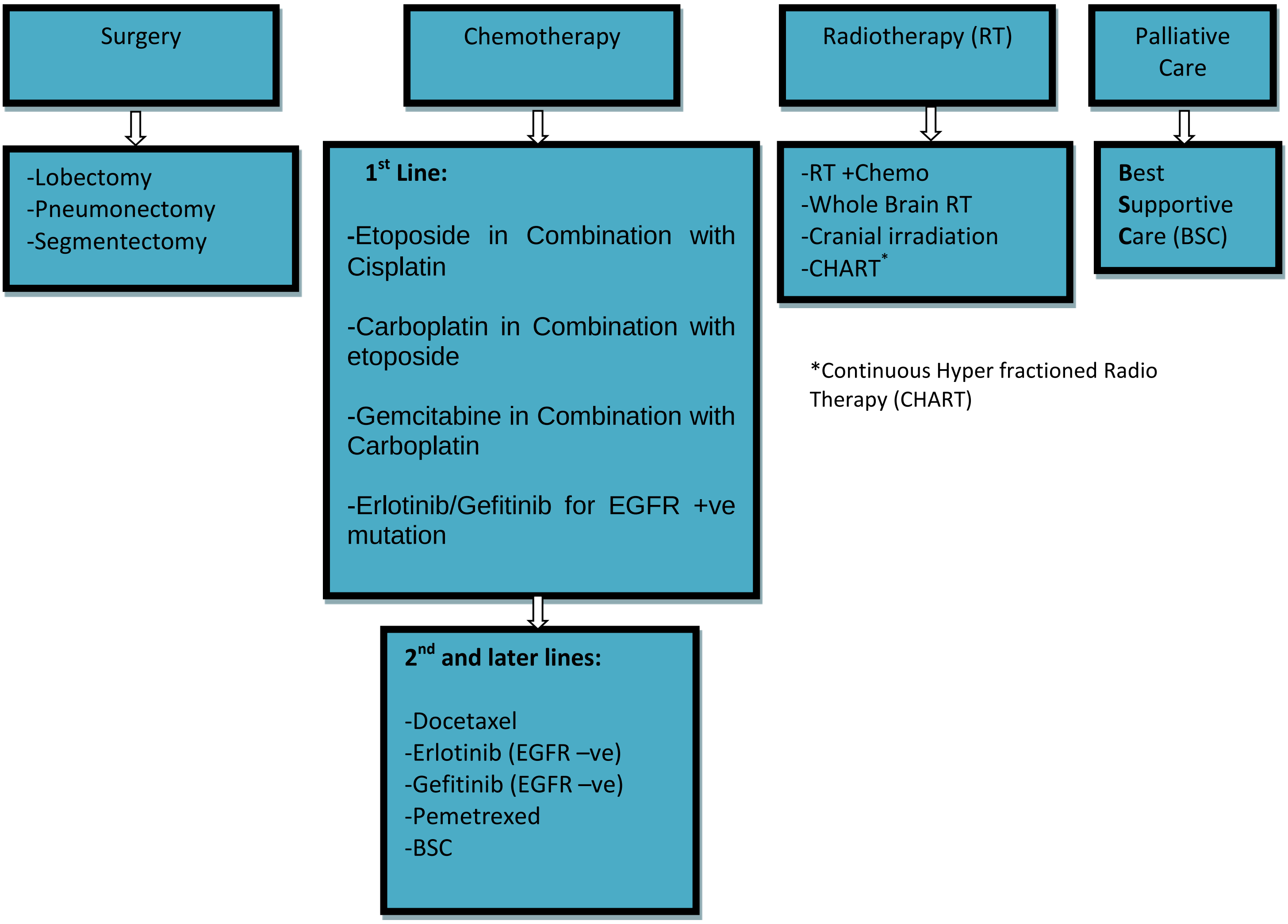
Table 1.1: Summary of NSCLC types (Thunnissen, 2013) [13]

NSCLC presents the largest group of patients for which the economic burden is considerable and for which the need for cost-effectiveness treatments is greater because of the increasing availability of (often expensive) treatments. Hence, I focus on NSCLC primarily in this thesis.

1.2 Treatments for Lung Cancer

Common approaches for treating cancer include (i) surgery, followed by (ii) chemotherapy and (iii) radiotherapy (not necessarily in that order). Despite treatment with chemotherapy, cancer recurrence is not uncommon. Recent novel chemotherapy treatments (e.g. immunotherapy) use the body’s immune system to fight and kill cancer cells. Some of these have proved to be cost-effective, while others have not [14]. ‘Cost-effective’ evaluates the relative costs and benefits of a given healthcare technology to determine its value to the taxpayer or relevant budget holder/payer. Most cancer treatments are associated with side effects that have a marked impact on HRQoL.

As an example, treatment options for a NSCLC patient (the data used in this thesis are from lung cancer patients) are shown in Figure 1.1. A given cancer treatment is often used for different types of tumors and results in common side effects, and therefore, similar patient experiences (e.g. quality of life).



*Continuous Hyper fractioned Radio Therapy (CHART)

Figure 1.1: Common Treatment Options for NSCLC

(Source: NICE guidance [15])

1.3 Economic Burden of Lung Cancer

The worldwide economic burden of treating cancer is high. Cancer has the most devastating economic impact amongst disease related causes of death in the world [16]. The exact worldwide economic costs of cancer are unknown but are estimated to be at least $895 billion ($US) [16].

Drug cost is a major (but not the only) cost component associated with treating cancer. Other costs include costs of treating side effects, surgery or administering chemotherapy, all of which can be significant. As an example, in the UK, the total annual cost of treating lung cancer in 2012 was about £3 billion [17] (20% of cancer costs) - the yearly average cost per patient was £9,071. This is comparable to £2,756 for bowel cancer, £1,584 for prostate cancer and £1,076 for breast cancer. Therefore, the costs associated with treating and managing lung cancer can be three times higher compared to the other types of cancer. In the USA, the mean monthly cost of treating lung cancer patients was estimated at £1,669 (no active treatment) £1=$1.61 and £5,814 (chemo-radiotherapy) [16]. Lung cancer, therefore, is a significant health and economic burden on the UK and worldwide health systems.

1.3.1 Policy Implications

With a growing and aging population, prevention efforts are critical for reducing new cancer cases, human suffering and economic costs [18]. In the United Kingdom (UK), the annual economic burden of cancer is estimated to be £15 billion. The National Health Service (NHS) increased its budget for cancer drugs from £200 million in 2013 to an expected £340 million in 2015 [19], a 70% increase in the Cancer Drug Fund (CDF) [20].

The CDF was set up in 2011 by the UK government to make funds available for paying for cancer drugs. It was changed in 2016 as its current form had become economically unsustainable. One critical change was an explicit reference to the cost-effectiveness of cancer drugs and resolving uncertainty associated with respect to their costs and effects [20]:

> *"Managed access agreements between NHS England and pharmaceutical companies, setting out the terms of a drug's entry into the CDF and the means by which data will be collected to resolve any uncertainty relating to a drug's clinical and cost-effectiveness."*

One objective of government health departments is the desire to optimize the use of cancer drugs by a combination of negotiated price reductions (of drugs) and improved clinical effectiveness. A number of cancer drugs have been removed from the CDF list [20], due to their lack of cost-effectiveness. This is despite NICE defined cost-effectiveness thresholds between £20,000 to £30,000 per quality adjusted life year (QALY)). In some special cases, this is set as a high as £50,000 per QALY where an end of life (EoL) criteria is satisfied. The EoL criteria apply when the life expectancy is short (<24 months), the new treatment improves survival by at least 3 months and the treatment is licensed for a population not exceeding 7,000 (in

England). At the time of writing this thesis, the CDF was under review [21] for which one key initiative was to ensure the drive for stronger value for money.

A more recent example of a drug not recommended by NICE is Nivolumab [14, 22]). Several reasons were cited for the rejection of this drug as not cost-effective. One of these reasons was the absence of adequate HRQoL (utility) data beyond disease progression which influenced the decision to recommend. Hence, HRQoL is central to the decision process for policy design and implementation for treating patients with NSCLC. Table 1.2 shows all the treatments rejected/approved by not only NICE but international re-imbursement agencies. It is interesting to note that the more recent drugs (at the time of writing) were not considered cost-effective at NICE accepted thresholds. A common thread for some of these rejections was related to an absence of or inadequate HRQoL data. It is in this context the reference to resolving “uncertainty” for determining “cost effectiveness” in the above quoted text [20, 21] is meant. Modelling HRQoL (this thesis) will offer ways to estimate otherwise unknown (but not unknowable) HRQoL data for economic evaluation. In this thesis, HRQoL data collected from lung cancer (specifically NSCLC) patients will be used.

	CADTH/ pCODR		**TLV**		**HAS**		**NICE**			**NoMA**		**PBAC**		**SMC**	
	1L	**2L**	**1L**	**2L**	**1L**	**2L**	**1L**	**2L**	**MTx**	**1L**	**2L**	**1L**	**2L**	**1L**	**2L**
Afatinib	pCODR 2013		4116		13272		310					07-2013	07-2013	920/13	920/13
Bevacizumab					5390		148							853/13	
Carboplatin															
Ceritinib								ID729							
Cisplatin															
Combination*							181							N/R	
Crizotinib	pCODR 2012	pCODR 2013						296							865/13
Docetaxel			3094				26	N/R						42/03	
Erlotinib		S0037	1066		12040		258	162	227	12937-7		07-2013		749/11	220/05
Gefitinib		S0003	2115		6839		192	175		18176-34		07-2013		615/10	
Gemcitabine							26								
Methotrexate															
Nintedanib							ID438							SMC 2015	
Paclitaxel							26								
Pemetrexed	pCODR 2013				5800	7892	181	124	190 309			03-2009	03-2010	531/09	342/07
Vinorelbine			377		6288		26					03-2006		179/05	

Table 1.2: Health Technology Assessment Results

Note: CADTH: Canadian Agency for Drugs and Technologies in Health, HAS: French National Authority for Health, HTA: Health Technology Assessment, MTx: Maintenance treatment; NICE: National Institute for Health and Care Excellence, NoMA: Norwegian Medicines Agency, N/R: Not reported; pCODR: pan-Canadian Oncology Review, SMC: Scottish Medicines Consortium, TLV: The Dental and Pharmaceutical Benefits Agency

*Combination therapy: Carboplatin-Taxol, Gemcitabine-cisplatin; 1L: First Line, 2L: Second Line

Green = Approved; Orange = Approved for some indications; Red = Rejected; Blue = currently reviewed

NOTE: Numbers in cells represent the technology appraisal id, if available [see 23 - 41]

1.4 Health-Related Quality of Life in Cancer

The World Health Organization (WHO) defines Quality of Life (QoL) as an individual’s perception of their position in life, in the context of the cultural and value systems in which they reside, and in relation to their goals, expectations, standards, and concerns [42]. HRQoL however, is a broader concept, affected in a complex way by the person's physical health, psychological state, the level of independence, social relationships, personal beliefs and their relationship to salient features of their environment.

Health-Related Quality of Life (HRQoL) is an important endpoint in cancer trials for several reasons. Firstly, when primary endpoint treatment effect sizes (based on mortality) are small, HRQoL offers the potential to ‘add value’ to expensive cancer treatments. Secondly, HRQoL outcomes are essential for cost-effectiveness analysis and drug reimbursement [81,206]. This is particularly true, where some generic HRQoL measures are used to adjust important efficacy outcomes for the purpose of demonstrating the cost-effectiveness of a new treatment. Thirdly, some anti-cancer treatments exhibit serious side-effects, despite improvements in overall survival (OS). The quality of the survival experience in the presence of such side-effects is essential for understanding the value of new cancer treatments from the perspective of the payer. Finally, different HRQoL instruments can result in varied (and sometimes opposing) interpretations of effect sizes. Consequently, the need to compare the effect sizes from different instruments on a common scale is vital, so that the clinicians, decision makers, and patients can align their understanding of HRQoL improvement or deterioration.

HRQoL is measured through various methods – often questionnaires, with specific questions about feelings, symptoms, physical ability and preferences (amongst other questions), in relation to their health. HRQoL data are often collected at several time points during a study (including clinical trials). In clinical trials, an experimental intervention is expected to yield at least equivalent or better clinical benefit (efficacy), compared to usual treatment. However, the new treatment may offer improved HRQoL benefit in addition to or despite lack of improved clinical benefit (for instance, the new treatment may be less toxic, with fewer side effects, leading to improved symptom control).

Limitations of Anti-Cancer Treatments

Cancer patients are concerned about their HRQoL during and after treatment [43]. Anti-cancer treatments have often resulted in some harm (sometimes without benefit) and in some cases clinical benefit with harm [44]. Consequently, HRQoL should be a key outcome measure when assessing the cost-effectiveness of a new intervention [45].

The implications for HRQoL during palliative therapy can be particularly acute because symptom palliation may contribute towards improved quantity and quality of life [46]. Since no further treatments (i.e. chemotherapy) are likely to be used during the end of life (EoL), the HRQoL benefits for patients and their carers from other forms of intervention (e.g. carer support, career education programs) may yield important HRQoL benefits.

Some researchers have suggested that a '*treatment can be recommended ….even without an improvement in survival if HRQoL is shown to improve*…' [47]. For instance, in nearly 8% of the RCTs in breast cancer, HRQoL influenced a treatment decision. In prostate cancer studies, involving chemotherapy and surgery, 25% and 60% of treatment decisions were influenced by HRQoL, respectively [48].

Due to the increasing number of therapy lines, smaller treatment effect sizes and increasing costs of drugs, HRQoL plays an important role in treatment, policy, rationing, and decision-making. This is likely to remain an important factor in the short to mid-term [49,50]; 26 out of 43 (60%) NSCLC studies, including randomized controlled trials (RCT), that assessed HRQoL, included a symptom specific measure (in addition to a cancer-specific measure). This suggests that a generic approach for measuring cancer HRQoL is inadequate [50], and only 2 studies (5%) used a generic measure. In this thesis, this conclusion will be investigated further when comparing condition specific and generic measures of HRQoL.

Why collect HRQoL Data?

HRQoL data are collected because researchers need to maximize the information about how anti-cancer treatments are working so that informed decisions for treating patients can be made. It is essential to know (from both patients' and clinicians' perspectives) not only what the side-effects associated with treatments are, but also *how* these side effects impact the patients' HRQoL. It is now universally accepted that HRQoL should be measured in clinical trials, however, the debate still continues as to what is the most reliable and practical way to obtain this data [50] or what constitutes to be a clinically relevant benefit from these measures*.*

The value of a new healthcare intervention may also have to be considered, through its benefit in terms of HRQoL and not survival. Although some cancers are curable (e.g. testicular cancer), a majority of them (including NSCLC), are considered to be incurable. Therefore, one of the objectives of cancer patient management should be improving HRQoL, particularly towards the end of life, when negligible clinical benefits are realized and fewer treatment options are available [44, 51, 52, and 53]. The risk-benefit and cost-benefit relationship between competing treatments, especially when clinical effects are small, can be guided by HRQoL outcomes [51]. For example, baseline HRQoL might predict survival benefits (e.g. patient who has poor HRQoL prior to treatment and improve on treatment, may also be the ones with improved HRQoL post treatment; whereas those patients who enter a study with good HRQoL may not improve their HRQoL further, even if survival is lengthened) [44]. Therefore, baseline HRQoL can be used to determine whether certain patients are more or less likely to benefit from a given treatment and whether a treatment's cost-effectiveness is likely to be greater for some patients.

Evaluating and measuring HRQoL benefit is an important aspect of economic evaluation. Economic evaluation (EE) is the process of systematic identification, measurement, and evaluation of the inputs and outcomes of two (or more) alternative activities (health interventions), and their subsequent comparative analysis [52]. EE in the context of cancer involves assessing the value of various cancer treatments, often through a metric which combines the quantity (length) of life and the quality of life experienced during that time – called a Quality Adjusted Life Year (QALY). This is particularly valuable when some expensive cancer drugs improve survival only to a minimal extent (e.g. for 1 or 2 weeks). In cancer trials

Challenges with HRQoL in Cancer Studies

There are several features of measuring HRQoL in cancer patients that are important. Firstly, HRQoL for the purposes of economic evaluation is often omitted [53,54] because the belief is that a condition specific measure will capture the necessary HRQoL features. This consequently results in methods adopted for estimating relevant HRQoL data, which is a key objective of this thesis. There are several reasons for this omission. Firstly, for some regions, HRQoL for EE is not important due to the specific health care system (e.g. the USA). Hence measures such as quality adjusted life years (QALYs) which combine quality and quantity of life as a single index, are not relevant. The difficulty is compounded when the same study conducted in a region where QALYs are not relevant (e.g. to the FDA in the USA) is also submitted for licensing to European authorities (at that point, the payer perspective becomes important and demands for estimation for QALYs is essential for patients to access cancer medicines (e.g. see [55,107]).

The second reason why HRQoL for EE is omitted is because the emphasis is placed on the clinical or disease specific aspects of HRQoL. HRQoL for EE are often considered to lack sensitivity. The BR21 trial [55] was primarily submitted to the USA for licensing where economic consideration at trial design was largely ignored [55]. HRQoL was considered only important so far as clinical effects are concerned. This is another objective of this thesis – to investigate to what extent, if any, disease specific and HRQoL measures for EE differ in terms of sensitivity. Another reason might be that two treatments are considered equivalent and therefore collecting HRQoL for an EE may not be useful.

One key feature in cancer is that patients can deteriorate rapidly thereby leading to an absence of both short and long-term data to evaluate efficacy and effectiveness. Economic evaluation is often determined over a life-time horizon and without available HRQoL, EE can become challenging. Some studies report at least 50% of the data missing within 3 months of starting treatment [46] due to disease progression, death or loss to follow up, attributed to the short survival time of patients and rapidly deteriorating health (especially after disease progression). For example, survival times for patients with NSCLC can be short (e.g. only 32% and 10% alive 1 and 5 years after diagnosis, respectively) [56].

Estimation of HRQoL within a study/trial and beyond protocol defined follow-up also plays a significant role in the EE of cancer drugs. For several economic evaluations of NSCLC treatments, estimation of HRQoL beyond study follow up have been performed [22] inadequately. Several technical documents describe methods of estimating both survival data and HRQoL in the absence of available patient level data and suggest further research to improve methods are needed in such circumstances [57] for cost-effectiveness analyses, particularly when such data are not collected (or available). The short survival time also limits the opportunity to collect HRQoL data within a narrow time window. A further challenge is ensuring the appropriate HRQoL is used. For example, the FACT-L and QLQ-C30 can both be used to measure HRQoL in lung cancer patients with different measures and conclusions regarding clinical benefit.

1.4.1 Measuring HRQoL: Generic and Condition-Specific Measures of HRQoL in Cancer

Condition-Specific Measures of HRQoL

Measuring HRQoL can be broadly classified into the two categories - condition-specific measures (CSM), which measure specific HRQoL symptoms (e.g. a cough, dyspnoea, etc.) and generic measures, which measure the broader HRQoL areas (e.g. mobility). Figure 1.2 illustrates the relation between some generic and condition-specific measures (CSM) of HRQoL.

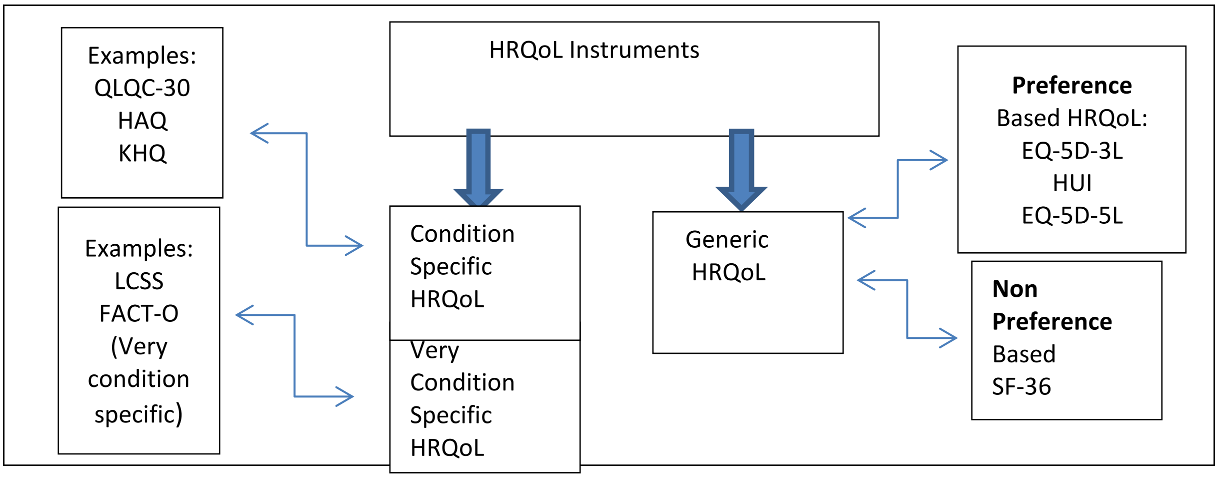


Figure 1.2: Relationship between Generic and Condition-Specific HRQoL Measures

Note: LCSS: Lung Cancer Symptom Specific questionnaire; HAQ: Health Assessment Questionnaire; KHQ: King's Health Questionnaire; HUI: Health Utilities Index; SF-6D: Short –Form 6D

In most cancer studies, HRQoL data wherever collected and reported, have been restricted to condition-specific measures (CSM). A CSM is an instrument that captures the specific quality of life issues in patients linked to a given disease. The wide use of CSMs is due to several reasons. Firstly, CSMs were validated for estimating clinical effects and historically cost-effectiveness was not considered as part of their validation. Secondly, a CSM was considered more sensitive than other generic measures for estimating HRQoL, focusing on specific symptom relief. Thirdly, an economic evaluation was not considered important. As budgets for health care became constrained while demand for health resource use grew, the impetus for rationing health resources became essential. For cost-effectiveness, HRQoL from CSMs are not used unless responses can be converted into a generic preference based measure (section 1.4.2). The two most common lung-specific measures used in studies of lung cancer are:

- EORTC QLQ-C30 and
- FACT-L: Functional Assessment of Cancer Therapy – Lung [49,54,56,58]

These are worthwhile reviewing as the EORTC-QLQ-C30 forms part of later chapters of this thesis.

(i) EORTC-QLQ-C30 Generic Cancer Instrument

The EORTC-QLQ-C30 (QLQ-C30) is a 'generic' cancer instrument [59, 60] consisting of 30 questions, out of which 28 questions are measured on a 4 point scale ('not at all' (1) to 'very much' (4)) and 2 questions are measured on a 7 point scale. Although it is generic across cancer types, it is not a generic instrument across all disease areas. The 30 questions result in **5** functional domains: Physical Functioning (PF), Role Functioning (RF), Emotional Functioning (EF), Cognitive Functioning (CF) and Social Functioning (SF); **8** symptom domains: Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Sleep Disturbance (SL), Appetite Loss (AP), Constipation (CO) and Diarrhoea (DI); and **2** further domains: Financial Impact (FI) and Global Quality of Life (QL). All raw responses are classified to a scale of 0 to 100, where a higher value represents better physical function for the function domains (including global and financial scales) and the converse for the symptom domains (a high value implies poorer symptoms).

For example, for PF, the 5 items (I, or questions) are summed together (questions 1 to 5):

$\mathbf{RS= (I_1+I_2 +\ldots I_5)/5}$ to generate the raw score (RS)

The final score is generated as: (1 – [(RS-1)/Range])*100. This is the score used for further analyses and interpretation of effects. A similar algorithm is used for the symptom scales and adjustments incorporated for missing data [61].

QLQ-C30 has been well documented, has good psychometric properties, is validated and is translated into more than 48 different languages with a large number of possible 'health states'. A health state, in economic evaluation terms, means combinations of different responses. For instance, one outcome for a particular patient from the QLQ-C30 could be 11111….1 (i.e. 30 responses of a value 1). This combination of 1's represents a 'health state'. In this sense, there are 4^{28} + 7^{2} possible health states. Inferences across all the possible health states are practically impossible. Summary statistics are often computed for each domain, in terms of the average (mean) scores and health states are less relevant for measuring clinical benefit.

(ii) FACT-G & FACT-L

The FACT-G is a 27 items cancer specific instrument. FACT-G consists of 5 subscales - physical wellbeing (PW, 7 items), social and family well-being (SW, 7 items), emotional well-being (EW, 6 items), functional well-being (FW, 7 items). FACT-L is a very CSM that consists of 10 additional lung cancer-specific items that supplement FACT-G, making it a 37-item questionnaire. These scores can be produced through three different calculations - a combined total of all the domains (FACT-L total); the Lung Cancer Score (LCS) and a Treatment Outcome Index (TOI), which can be calculated by summing the FACT-G physical, functional domains and the LCS [59]. This instrument also has a large number of possible health states.

The choice between utilizing FACT-G /FACT-L or QLQ-C30 is based on the subjective clinical bias, rather than empirical evidence of the superiority of one over the other. In fact, the empirical evidence presented for the relative superiority can be considered minimal or non-existent [58]. A related issue is which instrument is more (or less) sensitive to detecting a clinically relevant treatment benefit and what is the clinically or economically relevant effect size from these measures. This aspect also remains unknown and is not well understood. Maringwa et al. (2011) have suggested "important" effect sizes of varying magnitudes [60] (e.g. a difference of 10 points). Comparing effect sizes between the varying HRQoL measures has not been widely reported, particularly in NSCLC. In contrast, generic measures of HRQoL (for economic evaluation, in particular) have been criticized for the lack of sensitivity to detect HRQoL benefits. This has implications for later cost-effectiveness [62] because the QALY can be higher or lower depending on the sensitivity of the measure. A further important issue here is that a QALY, a key outcome for measuring cost-effectiveness, cannot be directly constructed from CSMs. These are estimated using generic measures of HRQoL.

1.4.2 Generic Measures of HRQoL

Generic measures of HRQoL capture responses about health in general and not the symptoms that might be associated with the toxicity of chemotherapy. They are useful for comparing effects across a variety of diseases. Although responses are captured from patients from these measures, it is the public’s *preference* for certain health states, termed ‘preference-based’ measures that are reflected in final outcomes. Preference based measures offer a way in which relative preferences (or value) for specific health states can be expressed by individuals in terms of ‘preference weights’ or ‘utilities’. The utilities can subsequently be for later cost-effectiveness analyses.

A utility value is often measured on a continuous scale and depending on which instrument is used, these values have different ranges (lowest and highest values). The health utility index (HUI), for instance, generates utilities on a scale from 0 to 1, where 0 represents ‘dead’, which is the worth state of health possible and 1 represents ‘Full’ health. However, not all generic measures of HRQoL generate utilities on a scale between 0 and 1. Other preference based generic HRQoL measures include the HUI: Health Utilities Index (versions I, II and III) [63], the EQ-5D-3L [64] and EQ-5D-5L [65].

Although responses from QLQ-C30 reflect how a given patient might feel with respect to their symptoms, disease or treatment received, the responses do not necessarily reflect how payers (people who ultimately pay for treatments through taxes) perceive the *value* of a given patient’s health condition, even if the patient is suffering from a disease as severe as cancer. It is possible that society (and not necessarily doctors) might regard a specific patient's health (state) as far worse and therefore believe any funds available to treat a patient’s illness should be spent elsewhere (e.g. preference for a breast cancer sufferer over a lung cancer patient). CSMs do not incorporate a relative valuation of the extent to which a specific symptom or health state affects the overall perception of health. For instance, severe nausea might be considered worse than severe pain for some patients, but not for others. The expression of such relative preferences (utilities) is determined by preference-based HRQoL measures. The QLQ-C30 is not preference-based measures of HRQoL and therefore, cannot be directly used in an economic evaluation. The most common preference based measures used in the UK (and some European countries) for EE are the EQ-5D-3L (EQ-5D-3L) and the more recent EQ-5D-5L.

EQ-5D-3L and 5L

The EQ-5D is a widely used generic measure, which is the shortest and perhaps the least cognitively demanding instrument that appears to be at least as responsive as the other community (preference) weighted instruments [66]. EQ-5D-3L consists of a descriptive health state classification system with five questions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), measured on three severity levels - 'no problems', 'some problems' and 'extreme problems'. A health state defined by the descriptive system of EQ-5D can be described by a five-digit number. For instance, 12113 refers to a patient, who has no problems with mobility (1), some problems with self-care (2), no problems for usual activities (1) or pain/discomfort (1) and extreme problems with anxiety/depression (3). Combining one level from each question defines 243 different possible health states from 11111 to 33333. The utility value associated with a health state 11111 interpreted as full health is 1. For a health state 33333, this is interpreted to be a state 'worse than death' (valued at -0.594). A utility value of zero would be equivalent to a state of death. The EQ-5D also has a visual analogue scale (VAS) which ranges from 0 to 1. The EQ VAS records the patient's self-rated health on a vertical visual analogue scale, where the endpoints are labelled 'The best health you can imagine' and 'The worst health you can imagine'. The VAS can be used as a quantitative measure of health outcome that reflect the patient's own judgement.

EQ-5D-5L

EQ-5D-5L is a revision of EQ-5D-3L. It consists of five questions, identical to EQ-5D-3L (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), but with an expanded 5 point scale and slightly different descriptors for each of the levels compared to the 3 point scale of the EQ-5D-3L [63]. For Mobility, Self-Care and Usual Activities, these are: 1: “No Problems”, 2: “Slight Problems”, 3: “Moderate Problems”, 4: “Severe Problems” and 5: “Unable to”; for the Pain/Discomfort and Anxiety/Depression scale, these were: 1: 'No', 2: 'Slight, 3: 'Moderate', 4: 'Severe' an 5: 'Extreme'. The scores are on a 5 point scale 1 to 5 (for each of the 5 domains). A perfect health state is '11111' and the worst possible health state would be '55555'. There are 3125 health states that can be identified using EQ-5D-5L (5^5).The corresponding minimum and maximal values are -0.281 for a health state of 55555 and a value of 1 for the 11111 health state.

Pre-determined scoring algorithms for EQ-5D have been developed in order to yield community-based health utility estimates (i.e. relative preferences for health states are not based on what the patients think, but what the general population believes) – specific to a given country. The derived utilities are determined from a pre-determined algorithm called a utility function. The utilities from these instruments (such as the EQ-5D, HUI, and other preference based measures) may subsequently be applied to clinical measures, such as survival or Progression Free Survival (PFS) time in order to derive a Quality Adjusted Life Year (QALY). A QALY is used as a generic measure of the effectiveness of a (new) intervention which combines both the quality and the quantity of life (length of life) experienced by each patient. This is a key measure for assessing the cost-effectiveness of health care interventions. One QALY is interpreted as equivalent to one year in 'Full' Health. If a patient's HRQoL (using a preference-based measure) during 1 year is less than in 'Full' Health, the QALY will be less than 1. QALYs are often accumulated at a rate of less than (or equal to) 1 per year. Examples of the EQ-5D along with other instruments mentioned in this thesis are presented in the Appendices.

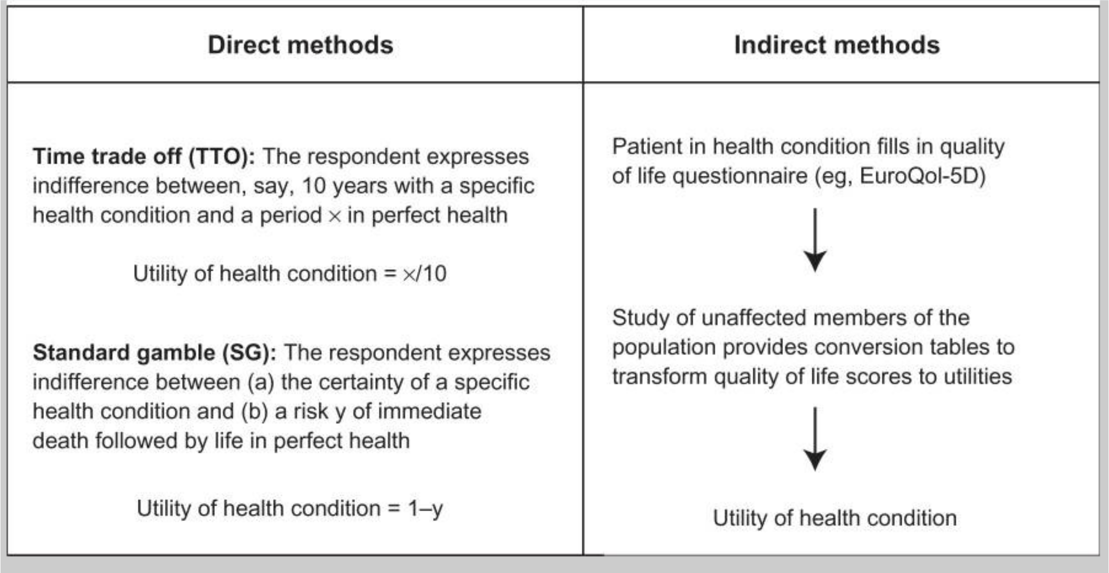
The primary difference between these two instruments (3L and 5L) is that the latter has responses measured on a 5 point scale, with many more health states [187]. EQ-5D-3L is reported to have limited discriminative ability (and also the power to detect differences between groups) compared to EQ-5D-5L [75,187-188]. At the time of writing this thesis, research was ongoing as to the best value sets for use with EQ-5D-5L. An interim scoring was available for EQ-5D-5L, using a crosswalk algorithm from EQ-5D-3L to EQ-5D-5L.

Preferences based condition specific measures

In addition to generic preference based measures, attempts have been made to develop condition specific preference based measures (CSPM) in cancer [67, 68, 69, 70, and 71]. Such measures use condition specific items (vignettes or questions) to derive algorithms to determine utilities. Once these algorithms have been 'validated' the items can then be used alongside or instead of the longer CSMs (such as the QLQ-C30). However, there are some limitations of these algorithms and hence a reason why they were not used in this thesis. Firstly, the estimates of utilities can differ significantly compared to those from the EQ-5D [72]. Secondly, the subset of states is only a fraction of the possible states (choosing 8 items from a possible 30 will be under-representing the true health states). Thirdly, even algorithms can be determined based on a designed valuation study with adequate statistical power for the main effects, the interactions are unlikely to be powered. Finally, in the case of EORTC-8D [69, 70], there are vignettes which are not included that are likely to be valued differently by lung cancer sufferers. As an example, in the 8D, shortness of breath after a long walk (current 8D) is likely to be less important (valued) than shortness of breath after a short walk. More research is needed to compare CSPMs with generic preference measures and other techniques for estimating utilities.

1.4.3 Constructing Utilities from the EQ-5D

Utility valuation methods may be classified as direct or indirect methods of health utility valuation (Figure 1.3) [73]. Most direct methods include trade-off (standard gamble [SG] and time trade-off [TTO]) and visual analogue scale (VAS). The main indirect methods of utility measurement include generic preference instruments, condition-specific preference measures, and mapping from a condition-specific HRQL instrument to a generic instrument. Indirect methods are based on mapping preferences onto the utility scale indirectly via a HRQL questionnaire; these methods are less time consuming, based on simple and versatile questionnaires that stratify data into a number of different dimensions not registered by direct methods. There is, however, no universally accepted theoretical basis for choosing direct or indirect methods.



Direct methods	Indirect methods
Time trade off (TTO): The respondent expresses indifference between, say, 10 years with a specific health condition and a period × in perfect health Utility of health condition = ×/10 **Standard gamble (SG):** The respondent expresses indifference between (a) the certainty of a specific health condition and (b) a risk y of immediate death followed by life in perfect health Utility of health condition = 1–y	Patient in health condition fills in quality of life questionnaire (eg, EuroQol-5D) ↓ Study of unaffected members of the population provides conversion tables to transform quality of life scores to utilities ↓ Utility of health condition

Figure 1.3: The EQ-5D Utility Function in Terms of Health States

Source: reference 73

Direct valuation, such as time trade-off (TTO) involves asking people (members of the public) to trade a given time (e.g. 10 years) in a health state worse than full health (but better than death), for a lower time in full health. The data collected from these options (trade-off) are then analyzed to derive weights for determining a utility index. An example of such elicitation in breast cancer patients has been reported [74].

The benefit of direct elicitation alongside a RCT might be that such valuation of health states within a RCT framework has strong internal validity. In addition, for patients who survive beyond the median survival time, an accurate reflection of the value of health states towards end of life might possible. However, using direct valuation methods from cancer patients alongside a clinical trial is difficult in practice because of the poor prognosis, short median survival times and logistics involved in clinical trial conduct. It may require extending the duration of the trial perhaps by following up patients until death, which would be impractical in most cases. For example, in order to estimate longer-term HRQoL effects (e.g. valuing health states towards the end of life), some patients would need to be followed up in the trial much longer as part of the same protocol. In practice, after cancer progression, follow up for many outcomes (other than necessary longer term safety) is often stopped. Trial governance may also complicate the process of further assessments once a patient's main follow up is completed. Moreover, patients in clinical trials are often excessively ill and there is a serious debate whether the patients are capable of ascertaining their health status through TTO and SG methods. The complex articulations can be daunting and time-consuming for some cancer patients, especially when they are preoccupied with a variety of other tests - scans, blood tests and radiotherapy planning.

An alternative to direct elicitation is to use a pre-scored health status descriptive system (questionnaires such as EQ-5D). Direct responses from patients are converted into utilities by using a standard set of published tariffs (which in turn are based on direct elicitation methods) such as those of Dolan (1997) [75] and Shaw et al. (2005) [76]. The word 'tariff' refers to the EQ-5D value of each of the health. Values such as 11111 or 21333 are not easily analyzed, but converting these responses to a utility value will allow analysis. The term 'health state', introduced earlier is now further elaborated in the context of EQ-5D.

Earlier (Section 1.4.2), it was explained that EQ-5D-3L consist of a descriptive health state classification system with five domains and 243 different health states ranging from full (value of 1) to worst (value of -0.549) health. Each of these health states is converted to a single number, called a utility value; a value from -0.594 to 1 (1 is full health and -0.594 represents the worst state possibly imaginable – even worse than death. If the Shaw et al. (2005) tariff is used the EQ-5D utilities range from -0.109 to 1.0, whereas the range is between -0.594 to 1.0 for the Dolan (1997) tariff [75, 76]. The subsequent utilities are used to compute a QALY by calculating an area under the curve (AUC).

In health economics, the concept of utility and utility functions is central to appreciate why health states from EQ-5D assume values between -0.594 and 1.0. A utility function is a mathematical representation of determining utilities for preferences of given health states. Mathematically, for each subject, i, the utility is [66]:

$$U_i = 1 – (0.081*K – \alpha_1*M_i –\alpha_2*S_i –\alpha_3*US_i –\alpha_4*P_i – \alpha_5*A_i – \alpha_6)$$

where $\alpha_1 \ldots \alpha_6$ are weights based on TTO scores, such that:

α_1= 0.069 if the Mobility score, M_i =2 for patient i and α_1=0.314 if M_i=3.

Similarly,

α_2= 0.104 if the Self-Care score S_i =2, and α_2=0.214 if S_i=3,

α_3= 0.036 if the Usual Activities US_i =2, and α_2=0.094 if US_i=3,

α_4= 0.386 if the Pain score P_i =2, and α_2=0.123 if P_i=3,

α_5= 0.071 if the Anxiety score A_i =2, and α_2=0.236 if A_i=3,

α_6= 0.269 if any of the M_i, S_i, US_i, P_i or A_i is a score of 3, otherwise α_6= 0,

and finally, K is the indicator variable, which takes the value 1 if any health state is dysfunctional (>1), otherwise, it is 0. For instance, for a health state of 12123, U_i would be 1 – (0.081 + 0 + 0.104 + 0 + 0.123 + 0.236+0.269) = 0.187. The utility values are ordered from lowest to highest and a numerical coding can be given to the ordered health states (11111 = 1, 2=21112=0.878…3= 33333= -0.549).

A few attempts have been made to ‘anchor’ the values from CSMs by dividing the raw response by the maximum scale value. For example, if the scale ranges from 0 to 100 (like QLQ-C30), then this would be reduced to a 0 to 1 scale by dividing it by 100. However, this is still not a measure of relative preference, even if 0 was equivalent to a state of death. For instance, if Pain was the clinical outcome, where, pain is scored from 0 to 100 (100 being the worst imaginable pain), then anchoring is not possible because the worst possible pain is not necessarily a state equivalent to ‘death’ (whether cancer pain or not). A utility value of 0, hence, does not imply that the patient is actually in a state of physical ‘death’. Anchoring is a useful method for revealing scale perception bias and evaluating data that are not directly comparable, thereby acting as a reference point.

In order to make a decision about the value (for money) of health care technologies, a generic measure of HRQoL, specifically developed for use in EE for decision-making is used (the EQ-5D). Other multi-attribute instruments, which are also preference-based include the Health Utility Index (HUI) version I, II and III [77], SF-15D [78], Quality of Well-Being Scales (QWB) [43] and SF-6D [80]. All of these are specifically designed to derive utilities.

There are various reasons why the EQ-5D is emphasized. Firstly, it is recommended for use in economic evaluations by the National Institute for Health and Care Excellence (NICE) [81-83]. NICE is a government decision-making body that assesses whether a new health care technology (e.g. new cancer drug) offers the UK tax payer value for money through the use of cost-effectiveness analysis by accessing the relationship between costs and the quality-adjusted life year (QALY). Several reasons exist why NICE prefer the EQ-5D including the fact that it is one of the shortest amongst similar measures, with fewest health states. It also allows NICE to compare consistently between disease areas. The EQ-5D is also used in several countries as a part of the economic evaluation and health technology appraisal (HTA). Moreover, the details of EQ-5D are well documented [66,75] and have been shown to be a reliable [84, 85] and valid HRQoL measure [86,87]. Finally, for the purpose of this thesis, utilities from both EQ-5D-3L and EQ-5D-5L data are readily available. The limitations of the EQ-5D have been raised on its lack of sensitivity and applicability to children [88, 89]. The recent EQ-5D-5L is meant to address the latter issue. The decision support unit (DSU) note the following in response to criticisms against the EQ-5D-3L:

“The expectation amongst its developers is that the five level version of EQ-5D will enhance responsiveness and sensitivity. This will have the impact of reducing the required sample size to detect small changes in health compared to the three level version. How this compares to alternative approaches for addressing inadequate sample sizes, and whether it will eradicate the need to employ these approaches, remains to be seen” [88]

1.4.4 Economic Evaluation in Absence of Utility Data

In EE, the use of HRQoL is particularly vital for a cost-utility analysis (CUA). CUA is a method of estimating the value of a new health technology by combining HRQoL with clinical outcomes (e.g. survival time) to derive a quality-adjusted life year (QALY).

A typical cost-utility analysis involves reporting the incremental cost-effectiveness ratio (ICER), defined as:

$$\textbf{ICER} = (\mu_{C1} - \mu_{C2})/(\mu_{E1} - \mu_{E2}) \qquad \textbf{[1.1]}$$

where, $\mu_{C1} - \mu_{C2}$ is the mean incremental cost for groups 1 and 2 respectively and $\mu_{E1} - \mu_{E2}$ is the mean incremental effectiveness between groups 1 and 2 respectively (groups 1 and 2 are typically patients who are allocated to two different treatments for comparison, which is common in clinical trials). The mean incremental QALY between treatments 1 and 2 can also be written as:

$$\sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{F_t Q_t} + \boldsymbol{F_{t+1} Q_{t+1}})^{1} * (\boldsymbol{T_{t+1}} - \boldsymbol{T_t})^{1} - \sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{F_t Q_t} + \boldsymbol{F_{t+1} Q_{t+1}})^{2} * (\boldsymbol{T_{t+1}} - \boldsymbol{T_t})^{2} \qquad \textbf{[1.2],}$$

where, $\mathbf{F_t}$ is the proportion (probability) that the patient is still alive at the time T=t and Q_t is the corresponding quality of life (utility) at time T = t. In equation [1.2], the first expression in $\boldsymbol{F_t Q_t}$ superscripted with 1 corresponds to μ_{E1} in equation [1.1] and is the mean QALY (area under the quality adjusted survival curve) for Group1. The second part, $\sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{F_t Q_t} + \boldsymbol{F_{t+1} Q_{t+1}})^{2} * (\boldsymbol{T_{t+1}} - \boldsymbol{T_t})^{2}$ corresponds to $\mu_{E2.}$ The values of $\mathbf{Q_t}$ are typically the (EQ-5D) utilities. If a measure of utility is unobtainable from any source, then the ICER can be expressed as an incremental cost per unit of effect (e.g. cost per cases detected), rather than an incremental cost per QALY gained.

When utility data are not available

Utility data are often not collected or unavailable for later cost-effectiveness analysis. There are several reasons why utility data may not be available, although a cost-utility analysis might still be required. Several examples in the literature report the main results of clinical trials where utility data were not collected in the trial, but later cost-utility analyses were performed using historical data [107]. One reason is that in some countries the health system does not require cost utility analyses. Therefore, cost-effectiveness was not part of the trial design (e.g. submission to the FDA in the USA). However, the same data are used for licensing purposes in Europe, where some countries do in fact require QALYs to be reported. Hence, estimates of patient level utilities are not available but a CUA is required. A second reason might be that EQ-5D are not considered sensitive to detecting treatment benefit. When utility data are not available but a cost-utility is required, utilities can be determined in several ways. One way might be through direct elicitation studies for example, such as TTO or SG methods, which may yield higher utilities and therefore QALYs.

Example of QALYs reported in some published cost effectiveness analyses

As an example, from 47 ICERs identified in the literature review (Chapter 2 describes in detail the approach to literature search), more than 20% could not generate a QALY due to an absence of generic HRQoL. This figure is higher once we take into account the number of studies where utilities were estimated from historical data (about 40%). This underlines the need to generate patient level utilities through alternative methods.

Treatment	**Cost (£)**	**QALY**	**Cost/QALY (£)**	**Year**	**Source/Reference**
Paclitaxel	44290	0.53	53227	2011	[90]
	27902	0.923	30230	2010	[91]
	21967	NR	NR	2000	[92]
	24216	NR	NR	2000	[92]
	26228	NR	NR	2000	[92]
	33685	0.4513	74639	2009	[45]
Gemcitabine	27837	0.934	29804	2010	[91]
	27401	0.966	28365	2010	[91]
	18129	NR	NR	2000	[92]
	47876	1.96	24427	2013	[93]
	38859	0.4676	83102	2009	[45]
Vinorelbine	23516	0.888	26482	2010	[91]
	16678	NR	NR	2000	[92]
	17482	NR	NR	2000	[92]
	6901	NR	NR	2010	[94]
Docetaxel	4129	0.1606	25712	2012	[95]
	13956	0.206	67748	2010	[55]
	27409	0.42	65260	2010	[96]
	24798	0.225	110215	2008	[97]
	24904	0.42	59296	2008	[98]
	11622	0.42	27672	2011	[99]

	20903	NR	NR	2011	[100]
Pemetrexed	5791	0.1715	33767	2012	[95]
	29387	0.52	56514	2010	[96]
	27764	0.241	115205	2008	[97]
	37119	0.41	90533	2008	[98]
	14239	0.41	34729	2011	[99]
	17455	0.97	17995	2010	[101]
	41731	0.5016	83195	2009	[45]
	8905	0.41	21720	2012	[102]
Gefitinib	6237	0.1745	22766	2012	[95]
	NR	1.111	NR	2010	[91]
	19787	0.79	25047	2013	[103]
	7704	0.79	9752	2013	[103]
	28471	0.91	31287	2012	[104]
	8980	0.2881	31170	2010	[105]
	10536	0.3188	33048	2010	[105]
Erlotinib	13730	0.238	57689	2010	[55]
	22439	0.25	89756	2008	[97]
	23567	0.42	56112	2008	[98]
	8229	0.1745	30292	2012	[95]
	25546	1.4	18247	2013	[93]
	23503	0.51	46085	2012	[106]
	12909	0.33	39119	2012	[106]
	8104	0.42	19296	2012	[102]
	22744	NR	NR	2011	[100]
	7488	NR	NR	2010	[107]

NR: Not Reported

Table 1.3: Examples of QALYs reported in published cancer studies

Collecting utilities through direct elicitation or valuation (such as Time Trade-Off or Standard Gamble) [68] alongside clinical trial can be expensive and impractical. Only a limited number of subjects might participate, which results in only a few reported health states. For instance, if only 10 subjects are included, then from a possible 243 health states, these 10 patients might be representative of only a few health states.

In certain cases, valuation methods may focus on a reduced set of questions. For instance, Rowen et al. (2012) determine utilities (directly) using a short form of the QLQ-C30 (EORTC-8D) in a cancer population, where patients were milder in terms of their severity (i.e. substantially longer median survival) [68]. With the EORTC-8D, as an instance, one question that arises is whether the patients would have the ability to carry out a “long walk” for assessing the physical function. However, in lung cancer patients, responses about “short walks” are likely to be equally, if not more, relevant [68]. Hence, the number of health states may be under-reported with the EORTC-8D . Furthermore, the purpose of this thesis is not to compare predicted utilities from valuation-based approaches to other (indirect) approaches, but rather, to investigate indirect approaches to estimation through mapping algorithms.

A second method to determine utilities for EE is to use aggregate utilities published in the literature. However, the patient population may be different and the most appropriate utility estimates may not necessarily be reported. Furthermore, adjustments for demographic factors that influence utility response are not possible with aggregate data; and may not take into account HRQoL changes occurring before and after a disease progresses, or heterogeneity in clinical and demographic characteristics. Significantly, they may not also adequately reflect between (and within) patient variability of utility scores as precisely as one could using patient level data. Examples of utilities often cited for cancer economic evaluation are found in Nafees et al. (2008) [108]. However, these results are not based on cancer patients, but on members of the public and moreover, the sample size was small in this study.

Mapping

An alternative approach to estimating utilities is through mapping and extrapolation using statistical modelling techniques. Mapping or ‘cross-walking’ can be useful when patient-level utilities are not available in a clinical trial. A statistical model sometimes termed as ‘mapping algorithm’, is used to predict (estimate) EQ-5D-3L from a disease-specific measure like QLQ-C30. If patient level EQ-5D-3L cannot be obtained, then it becomes challenging to conduct a CUA with patient-level data and reliance is made on published aggregate utilities. Mapping is, therefore, a critical (and sometimes the only) way to estimate patient-level utilities for a trial.

Mapping may refer to estimating utilities between the health states described by a CSM and those described by a generic measure and applying the utility values to the mapped states. Another type of mapping is from the condition specific descriptors directly to utility scores, and the third type of mapping is from summary measures derived from the condition specific descriptors to utilities. However, the second form of mapping as commonly reported in literature is a method can be described as a method where the interrelationship between a generic HRQoL measure like EQ-5D and a condition-specific HRQoL measure (e.g. QLQ-C30) is modelled, so that patient level utilities can be predicted (estimate) [109] for the purposes of an economic evaluation. It is argued that mapping is preferred (e.g. by NICE in the UK) over other valuation methods (e.g. utility studies) which may not be acceptable for a HTAs [81].

A statistical model is developed to facilitate the prediction of EQ-5D-3L from a disease-specific measure, such as QLQ-C30. If patient level EQ-5D-3L cannot be obtained, then it becomes difficult to conduct a Cost Utility Analysis (CUA) with patient-level data and one may need to rely on published aggregate utilities. Mapping is, therefore, an important (and sometimes the only) way to estimate patient-level utilities, which can avoid some (but not all) of the limitations (e.g. differences in populations, disease severity) and uncertainties associated with using published aggregate utilities.

It is fairly common to find a mapping algorithm, which is used to predict (i.e. estimate) patient level utilities for a clinical trial in a particular disease area (e.g. pain), even though the mapping algorithm may be developed using data from a different patient population [110]. Crott et al. (2012) suggest that different algorithms or functional forms may exist for each cancer type [111] and similar issues have also been raised elsewhere [112]. Therefore, it is not often clear as to whether authors of published mapping algorithms intend users of mapping algorithms to generalize their use to *any* patient population or not. Moreover, how factors such as the timing of measurements or presence of differential treatment effects influence predicted utilities do not always seem to be accounted for in modelling. For instance, where algorithms have been developed using only baseline data, it is often not immediately evident how useful they are for predicting post baseline EQ-5D-3L utilities. Users of algorithms are often interested in predicting differential (post baseline) utilities for cost-effectiveness [113]. The effects of treatments on HRQoL occur after treatment has commenced (post baseline) and therefore predicted utilities are likely to be more important and useful where post baseline data are used.

Although EQ-5D-3L is a relatively short instrument, it is surprising that many studies do not collect EQ-5D data. In fact, only 16% of all studies (i.e. observational, surveys, RCTs) in lung cancer gathered preference-based HRQoL data prospectively [114]. In some clinical trials, patient-level utilities were not collected, although formal economic evaluations were conducted at some later point; 25% of HTA submissions to NICE used mapping in such situations [115]. In Australian HTAs, this rate was slightly lower at 24% [116]. The relatively recent introduction of the EQ-5D-5L (measured on a 5 point scale) suggests that EQ-5D-3L may not be adequate to address concerns about sensitivity (i.e. ability to detect treatment benefit). At the time of writing this thesis, the EQ-5D-3L continues to be widely used due to lack of experience and comparison between the two and the absence of specific recommendations for one over the other.

Reasons for Mapping

Given that mapping has an important role in EE, the reasons why mapping is considered important will be outlined. In some early phase cancer trials, preference-based measures are not usually collected, but CSMs, such as QLQ-C30 are gathered to provide early indications of symptom control for future phase II/III trial planning, particularly for drug reimbursement (i.e. the process of negotiating a price for a new healthcare technology) and HTA. A mapping function can be used to estimate EQ-5D-3L from (combined) early trial data for planning the cost-effectiveness argument for later phase III trials. In situations where two identical trials are required for licensing purposes (as in multiple sclerosis), a useful mapping algorithm from one trial can be used to determine utilities in the other trial [117]. If the two trial designs and patient populations are identical, there might be a possibility for developing a mapping algorithm in the first trial and predicting the utilities for the second. In certain trials intended for drug licensing, EQ-5D are not collected or required (e.g. USA, Germany), yet the same set of data are used to support licensing of the new drug within the European Union, where cost-effectiveness is often of great importance.

The usefulness of a mapping algorithm lies in its ability to predict utility values from independent data. A useful mapping function should have good predictive properties, accuracy and model fit, and wherever possible, the differences between the observed versus expected QALY minimized. The closer the predicted utility is to the observed value, the more useful the algorithm is likely to be. Although the true value of the utility being predicted is unknown, simulation methods can lead to conclusions that the predicted utility (or QALY) lies within a quantifiable range of the observed values with a reasonable degree of certainty [109].

Ades (2013), in support of mapping, argues that it is not always necessarily better to directly estimate HRQoL [62]. One theoretical reason for this assumption is that the estimate of HRQoL effects (e.g. mean differences) may have lower variability (i.e. they are minimum variance unbiased estimates) than those from directly estimating

HRQoL. In particular, this is dependent on the relative response between a CSM and the generic HRQoL measure [62].

Despite this, a sizeable proportion of HTAs [109], there are no clear definitions of ‘good’ or ‘bad’ mapping algorithms. Moreover, several questions still remain in relation to modelling HRQoL for economic evaluation and further investigation is needed. Some of these questions are incorporated in the aims and objectives of this thesis stated later in Chapter 3.

Limitations of Mapping

Although mapping can be useful (and sometimes necessary), it is preferable to collect EQ-5D, wherever possible [112]. Some particular problems identified with mapping include the limited ability of models to predict at poorer health states (Rowen et al., 2012) [67]; the assumption being that there is a conceptual relationship (overlap) between the two measures [118] and that the predicted value is a true measure of HRQoL from the base measure [119] is also a concern. The main alternatives to mapping are either to use historically reported estimates or conduct utility studies. In the first case of using historical estimates, since anti-cancer treatment changes over time, so would the associated side effects. The impact on HRQoL may, therefore, be different to what it might have been on previous anti-cancer therapies (e.g. new immunotherapies have specific side effects different to previous/other historical anti-cancer treatments). Consequently, using historical utility estimates may not reflect current treatment trends. In the second case, designing a separate utility study, whether using discrete choice experiments or otherwise is likely to result in fewer health states (e.g. a subset of health) from which inference can be drawn compared to mapping; and in any case may be more expensive and may add to the burden of patients if conducted alongside a clinical trial. A separate study which compares direct elicitation and mapping is an ongoing area of current research.

In addition to the statistical framework of mapping algorithms, questions have been raised about the usefulness and validity of mapping [118]. It is suggested that it is unclear as to what exactly is being predicted from mapping models as the target is unknown [186]. However, this is precisely what a mapping model is supposed to do - to estimate the unknown utilities, which are assumed to be ‘knowable’ on the basis of reasonable assumptions It does not mean the unknown (target utility) is simply ‘unknowable’.

Although this, among other criticisms of mapping, is significant [109,112,118,185], they are perhaps not strong enough to dismiss mapping altogether. Consequently, about 25% of HTA (to NICE) have used mapping [115] in the UK; while in Australia, this was reported to be about 24% (Schuffham, 2008) [116]. Moreover, the published mapping models (for the QLQ-C30), suggest the unknown utilities are likely to be 'knowable' to some extent as some mapping algorithms have shown to yield close approximates of the target mean utility. Therefore, the mapping can be useful in estimating patient level utilities and continues to be used in HTAs of cancer drugs for estimating utilities (or sensitivity analyses), despite the prevalent criticisms. Moreover, simulations have shown that mapping can indeed estimate the sampled utility and is associated with a strong overlap (measured as correlation) between EQ-5D domains and CSMs [109].

A feature common to all published algorithms, including the ones used in cancer, has been the over-prediction of utilities in patients with poorer health states. This means that models predict higher utilities than expected when, in fact, patients have poorer HRQoL. Moreover, most existing models do not appear to have properly addressed over-dispersion at the extremes of the distribution (i.e. many patients have values of 0 or 1). However, the mapping is likely to be less expensive than a separate utility study with the important caveat that a useful algorithm can be identified. Further research has been recommended to address the uncertainty of algorithms in a robust way by using more complex approaches.

Mapping algorithms based on EQ-5D-3L have been shown to consistently over-predict utilities, particularly at poorer health states [109,112]. In order to address some of the limitations, alternative functional and statistical forms of mapping algorithms have been examined [109,134,139,161]. These functional forms in some cases generated improved predictive capability [109,134]. In certain cases, however, changing the functional form did not offer improved prediction over and above simpler models [109,112]. Moreover, when applied to external data, some of the algorithms performed poorly [161,162].

Mapping and Cost-Utility Analysis

A number of health economic evaluations have included estimates of utilities determined from mapping to generate a cost per QALYs [120]. Other uses of mapping range from Rheumatoid Arthritis [121,122]; Multiple Sclerosis [123]; Sarcoma [124]; Alzheimer's and Diabetes [125,126]; Pain [110] and Breast Cancer

[127]. Schuffham (2008) [116] identified about 24% of health technology submissions between 2002 and 2004 made to the Australian Benefits Advisory Committee (PBAC) had included mapping.

Longworth and Rowen (2013) identified from 71 separate CUAs that about 25% used a method of mapping for NICE HTAs [115]. This has increased between 2009 and 2013 [112]. It is interesting to note that despite persistent criticism of mapping algorithms, they are still increasingly used in publications and NICE submissions. In summary, the fact that NICE recommends mapping over utility studies, the increasing use of mapping and research activity in this area and avoiding the need to what can be (expensive) utility studies supports the need for mapping and subsequent research in this area. This provides a rationale to research the effectiveness and sustainability of mapping. The areas of research were identified from a literature research which subsequently led to specific aims and objectives that will be answered in this thesis, discussed in the next two chapters.

Chapter 2

Chapter 2: Literature Search: Methods, Strategy and Thesis Objectives

Abstract

Introduction: Published articles relating to themes such as ‘Modelling HRQoL’, ‘Economic Evaluation’ and/or ‘Cancer’ were identified for further literature searches. The aim of the literature search was to identify gaps in the knowledge for future hypotheses generation. The specific themes of interest were: Mapping, Bayesian approaches in mapping, sensitivity, and responsiveness of the EQ-5D, utility extrapolation and model selection. Aims and objectives of the thesis were identified.

Methods: Medical Subject Headings (MESH) search terms were used with appropriate wildcards in search engines including PUBMED, MEDLINE, and COCHRANE Database of Systematic reviews. Broad search terms were used to identify articles associated with terms such as ‘EQ-5D’, ‘QLQ-C30’, ‘Cancer’, ‘HRQoL’, ‘Mapping’ and ‘Economic Evaluation’. The criteria for selecting articles are presented. In general, relevant articles were initially identified from titles or abstracts. Each of the articles considered relevant was reviewed in full.

Results: From 443 (Mapping) and 559 (Sensitivity and Responsiveness) articles, between 4 (Bayesian Mapping) and 8 (Responsiveness) articles were considered suitable for further review. A further 14 articles on Mapping from 443 were considered relevant to cancer context. Several research themes were identified for further research: a need for improved mapping functions, especially with the more recent EQ-5D-5L; investigation of other structural forms: Joint and Bayesian Network Models; sensitivity and responsiveness of generic versus cancer-specific measures; estimating post-progression utility and an approach to the selection of an optimal mapping model.

Conclusion: From these reviews, the evidence for gaps in knowledge can be demonstrated and aims and objectives of this thesis were subsequently defined.

2.1 Literature Search Methods: Introduction

In this chapter, the methods used for conducting the literature search are presented. A summary of the searches is categorized into five main areas for this thesis (Table 2.1).

Category	Thesis Objectives / Research Question	Number of relevant articles	Chapter
1	Mapping Algorithms	24	4,5,6
2	Bayesian Mapping Algorithms	1	7
3	Sensitivity and Responsiveness of EQ-5D	7	8
4	Extrapolation of Utility after cancer progression	1	9
5	Criteria for selecting mapping models	14	10

Table 2.1: Summary of Thesis objectives, number of articles and relevant chapter

The justification for this categorization becomes clearer in section 2.4, although the aims of this grouping are briefly outlined. The aim of categories 1 and 2 (mapping algorithms) were to identify literature on mapping algorithms used in cancer, with a view to identifying limitations and potential areas for improvement. Research activity in the area of mapping and modelling HRQoL has increased with more mapping algorithms available, but limitations persist in model structure and performance.

The reason for a search in the area of responsiveness and sensitivity (category 3) is largely due to the absence of limited information on the comparative performance of the recent EQ-5D-5L not only in terms of mapping but also its role in economic evaluation when compared with disease specific measures. The aim here is to identify potential areas of improvement and answer the outstanding question as to whether QALYs are over or under estimated as a result of lack of sensitivity of generic measures, a common criticism of the EQ-5D. The reason for a review in the area of ‘Extrapolation of Utilities’ (category 4) is because HTAs of cancer treatments continue to highlight the problems of missing utility data associated with the economic evaluation of cancer drugs. The aim of this search (and review) was to identify potential to develop methods to estimate utility after cancer progression to inform longer-term cost-effectiveness. Finally, given the fact that research in mapping is increasing and more algorithms are available (category 5 in Table 2.1), the need to develop selection criteria was considered to be important. The aim of this review was therefore to determine the methodology for selecting published algorithms. These are some of the reasons why justification for a literature search in these areas is needed.

2.2 Literature Search: General Search Strategy and Search Terms

The literature search comprised of examining published articles and national (UK) and international health technology appraisals (HTAs) of cancer drugs. HTAs were also selected for review because it is important to understand the practical methodological issues raised by HTA reviewers when appraising the cost-effectiveness of cancer drugs.

Broad searches were undertaken for publications (in the English language) on cost-effectiveness, HRQoL and mapping in cancer trials up to 22nd May 2016. No restrictions were placed on the earliest articles in order to maximize the number of possible articles for review (all studies at any time). A final update of searches was undertaken in January 2017 to reflect any recent developments, which had significant implications for this thesis and none were noted.

Databases that were searched included MEDLINE, COCHRANE DATABASE OF SYSTEMATIC REVIEWS, NHS Economic Evaluation Database (EED), NHS HTA, OHE, Cost-effectiveness Analyses Registry, SCOPUS, WEB OF SCIENCE, CANCER TRIALS REGISTRY and country level reimbursement bodies where HTAs were submitted (in English); wild cards (*) were used to maximize the search potential. The following terms (below) were included when devising a search strategy.

a) 'Cancer' Search terms in a) and (cost effectiveness or cost* or QALY* or NICE* or Cost-effectiveness Acceptability Curves* or Cost per QALY or effectiveness* or …..).

b) Search terms in a) and b) and Quality of Life, or Quality* or Health Related* or QALY.

c) Search terms in the above and Missing data or Missing* or

d) Search terms from the above and Condition-Specific QoL or Generic QoL, Generic* or Condition-Specific* or

e) Search terms above and Mapping*, direct elicitation, indirect elicitation, direct*, indirect*

f) *EQ-5D or *Cross-walking.

The abstracts and titles of all articles were initially examined for relevance. The selected articles were then reviewed in detail in relation to the themes identified in (1) to (5) above (Table 2.1). The criteria for comprehensively reviewing articles were based on:

(a) Relevance: whether the article included the key search terms (e.g. in the title or abstract such as 'Cancer', 'Cost Effectiveness', 'HRQoL', 'Mapping', 'QALY', 'Cross-walking'.
(b) The article was in the English language.
(c) The rationale of the article was related to the theme/question of interest in (1) to (5) above (Table 2.1).

The specific details of the searches are provided below.

2.3 Literature Search: Results for each chapter

2.3.1 Literature Search: Mapping (Chapters 4,5,6,7 and 10)

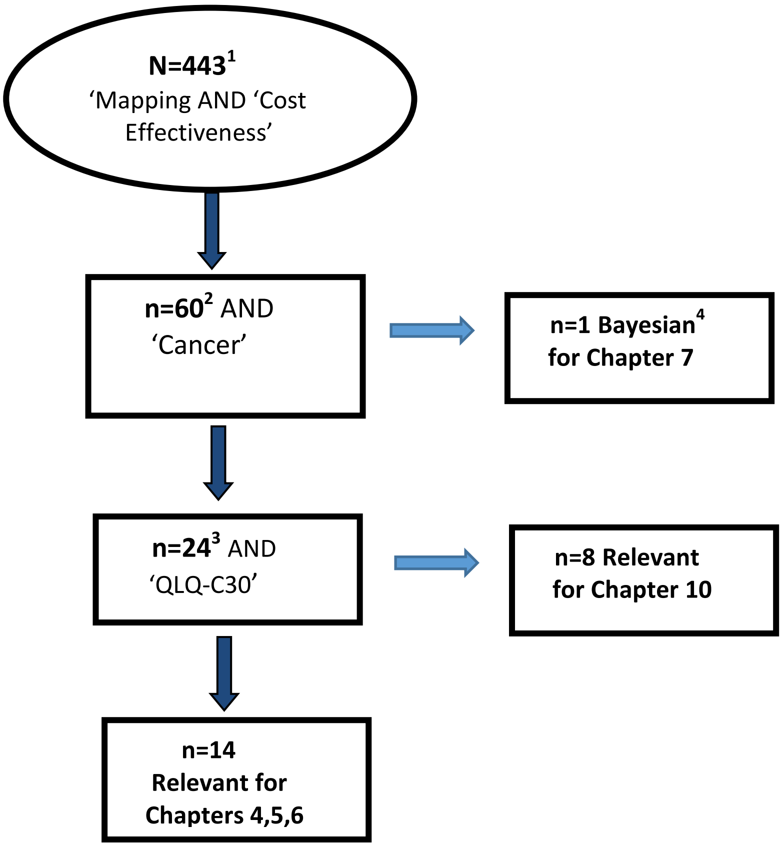

N=443[1]
'Mapping AND 'Cost Effectiveness'
n=60[2] AND 'Cancer'
n=1 Bayesian[4] for Chapter 7
n=24[3] AND 'QLQ-C30'
n=8 Relevant for Chapter 10
n=14 Relevant for Chapters 4,5,6

Figure 2.1: Literature search of number of articles (Mapping)

[1] mapping[All Fields] AND ("cost-benefit analysis"[MeSH Terms] OR ("cost-benefit"[All Fields] AND "analysis"[All Fields]) OR "cost-benefit analysis"[All Fields] OR ("cost"[All Fields] AND "effectiveness"[All Fields]) OR "cost effectiveness"[All Fields])
[2] AND ("neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields])
[3] mapping[All Fields] AND QLQ-C30[All Fields]
[4] (mapping[All Fields] AND ("cost-benefit analysis"[MeSH Terms] OR ("cost-benefit"[All Fields] AND "analysis"[All Fields]) OR "cost-benefit analysis"[All Fields] OR ("cost"[All Fields] AND "effectiveness"[All Fields]) OR "cost effectiveness"[All Fields])) AND Bayesian[All Fields]

The aim of this review was to identify all mapping algorithms related to cancer, to further review the methodological limitations and opportunities for improvements as well as identify areas of research not considered.

From a total of 443 articles on mapping in the context of cost-effectiveness, after reviewing titles, 60 were relevant to cancer. After a further review of abstracts and titles, 14 were relevant to chapters 4, 5 and 6 of this thesis and 8 were relevant to chapter 10. The search terms were used because they broadly define the research themes of this thesis. Only 1 article was a cancer specific Bayesian mapping algorithm.

2.3.2 Literature Search: Sensitivity and Responsiveness (Chapter 8)

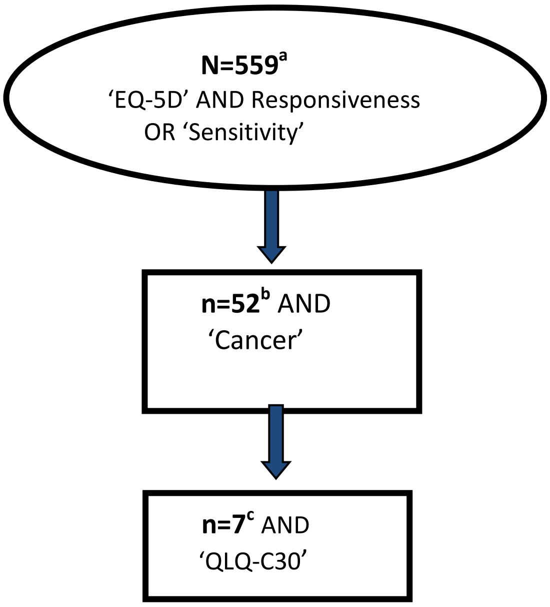


Figure 2.2: Literature search of number of articles (Sensitivity and Responsiveness)

[a] EQ-5D[All Fields] AND (Responsiveness[All Fields] OR ("sensitivity and specificity"[MeSH Terms] OR ("sensitivity"[All Fields] AND "specificity"[All Fields]) OR "sensitivity and specificity"[All Fields] OR "sensitivity"[All Fields]))
[b] AND ("neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields])
[c] AND QLQ-C30[All Fields]

The aim of this review was to identify research articles on the comparative sensitivity and responsiveness of generic (both EQ-5D-3L and EQ-5D-5L) and cancer specific measures. The objective was to identify articles which investigated whether generic measures under-estimated patient benefit and what impact this might have on the QALY and economic evaluation of cancer treatments.

From 559 articles related to EQ-5D responsiveness and sensitivity, of which 52 were cancer studies and amongst these for the purposes of this thesis, 7 articles were considered relevant because the data that will be used in this thesis related to EQ-5D and the cancer specific QLQ-C30.

2.3.3 Literature Search: Utility extrapolation Mapping (Chapters 9)

The aim of this review was to identify articles which proposed methods to estimate utility after cancer progression, for the purposes of economic evaluation. This is an important area of research because many HTAs are critical on how utilities are estimates after cancer progression. The purpose of this review was to investigate gaps in the knowledge for methodological improvement.

Using the search terms: EQ-5D[All Fields] AND ("neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "cancer"[All Fields])", 488 articles related to EQ-5D and cancer, of which 2 were potentially related to utility extrapolation or post-progression utilities. After reviewing the articles in detail, none were related to the theme of post-progression utility extrapolation.

2.3.4 Literature Search: Criteria for selecting Mapping Models (Chapter 10)

Given the publication of several published mapping algorithms, the objective here was to identify articles of published mapping algorithms (used in cancer) and then determine whether a robust selection method is possible to distinguish between 'Useful' and 'Not Useful' algorithms. This is needed because as more algorithms become available, methods that distinguish between poor and good performing algorithms is needed.

The details of the searches are the same as those in section 2.3.1, where n=8 articles are identified which can be used to compare and select published algorithms. These articles were considered relevant on the basis of criteria including:

a) Ensuring that QLQ-C30 was used as a part of the process of developing the algorithm.

b) At least 2 coefficients were reported in the mapping function (the intercept and at least one of the 15 domains (these domains consist of 5 functional domains: Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF); 8 symptom domains: Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI); a domain for Financial Problems (FI) and an overall score: Global Health Status Score (QL)).

c) Algorithms were included irrespective of the tumor type. Most algorithms were reported in such a way that it appeared that the authors intended them to be used across all tumor types.

d) Algorithms were included, whether developed from RCT data or other studies (e.g. surveys, observational studies).

2.4 Summary of Literature Review of Mapping (Chapter 4, 5 and 6)

2.4.1 Previous research on mapping

Model Structure

Most approaches to mapping have generally used OLS models but other models, such as TOBIT and LOGIT, were also utilized at times [112]. Out of 119 models examined, ordinary least squares (OLS) approaches were the most common [112]. Criticism on the use of OLS models is concerned with under-estimation of the uncertainty of mean utilities used for cost-effectiveness. For instance, the variability of mean predictions may be greater than what the OLS models might suggest [128].

Models used for mapping comprising of conditional mean or median regression models had varying degrees of success [111,112,129,130-132] (e.g. TOBIT, Quantile Regression). Response mapping approaches, where ordinal categorical responses are modelled, have also been used with limited success [133]. However, if there are a few responses at extremes, the predictions at these extremes are likely to be imprecise, particularly with smaller sample sizes. Hernandez et al. (2010) compared linear and TOBIT models with adjusted censored models , a class of limited dependent variable mixture models (LDVM) which essentially modify the TOBIT and are applied in a mixture modelling framework (but none were applied to cancer data sets and QLQ-C30 in particular) [134]. These models appear to work well, but seem to need larger sample sizes. For smaller size clinical trials, these models may also have limited ability to predict at the extremes. Longworth and Rowen (2013) suggest several alternative models for investigation, including the Beta-Binomial [115]. Crott et al. (2012) also suggest research of more complex mapping algorithms, as well as a need for greater validation [135]. One aspect of this ‘complexity’ might include adding many interactions and additional demographic variables, which may improve prediction, but on the other hand, may result in an overly complicated algorithm (e.g. rather than 15 QLQ-C30 variables, one may have a model with 30 factors, including interactions to predict EQ-5D-3L). Another way of considering ‘complexity’ could be a more complicated mathematical function, but with fewer independent variables. Table 2.2 shows details of existing mapping models reported in the literature (all used the Dolan [75] tariff for EQ-5D-3L).

Model Comparison

Mapping models have been developed and compared using measures like predictive power (R^2), predictive mean and residual mean squared error (RMSE). Simplistic approaches to reporting model performance were used (e.g. R^2), which may not be sufficient to conclude a successful mapping (e.g. in non-linear models, R^2 may not alone be appropriate). In addition, concerns were raised regarding the over-emphasis on R^2 as a measure of the model fit [112]. It is the predictions of the mean utilities and the uncertainty around them, which are critical in a model choice. It was also unclear as to what constitutes a good or poor R^2 in these models. For instance, on the one hand, a value of R^2 at 50% was considered to be ‘high’, without justification [112]. Testing model performance through the use of simulation may be a way to quantify the uncertainty of algorithms. It is appropriate to include measures of model fit (e.g. Aikakes Information Criteria (AIC)) or visual plots to understand mode fit as well as determining the impact on the QALY, the critical metric of cost-effectiveness.

Model Validation

Few models used independent datasets to validate the model. When the same published models were used in different data sets the results were not as good as those in the original development and reporting of the model. Table 2.2 gives a list of relevant models that used external validation.

Year [Reference]	**Sample Size**	**Population**	**Mapping From/To**	**Time points (month)**	**Model Type**	R^{2a}	Country	validation
1.2016 Mariott et al. [136]	529	Colorectal cancer	QLQ-C30/ EQ-5D-3L	0, 1m, 12m	Mixed TOBIT	65% 51%	Multinational	None
2.2016 Khan et al. [137]	100	NSCLC	QLQ-C30/ EQ-5D-3L EQ-5D-5L	Monthly for 12 months	2 Part Beta OLS	75%	UK	Cross validation & Simulation
3. 2015 Young et al. [138]	530 to 771	Mixed Tumour	FACT-G QLQ-C30/ EQ-5D-3L	Baseline and post baseline	OLS TOBIT 2 part Splines	51%	Multinational	Independent data
4. 2015 Kharroubi et al. [139]	1839	Myeloma	QLQ-C30/ EQ-5D-3L	Cross section	OLS Bayesian Imputation	70%	Multinational	MCMC Simulation
5. 2014 Khan et al. [109]	890	NSCLC	QLQ-C30/ EQ-5D-3L	0, monthly until progression or death	3 part Beta	75%	UK	Simulation & Independent data
*6. 2014 Proskovorsky et al.[140]	154	Myeloma	QLQ-C30 QLQ-MY20 EQ-5D-3L	Baseline and post baseline	OLS	69%	Multinational	None
*7. 2012 Versteegh et al. [141]	137	Mixed	QLQ-C30 HAQ	Baseline and 6 follow-up	OLS	82%	Netherlands	Independent data

			MSIS-29/ EQ-5D-3L	measures				
*8. 2012 Kim SH et al. [142]	893	Mixed	QLQ-C30/ EQ-5D-3L	unknown	OLS	52%	S.Korea	Independent data
*9. 2012 Kim EJ et al. [143]	199	Breast Cancer	QLQ-C30 BR23/ EQ-5D-3L	Cross-sectional survey	OLS	49%	S.Korea	Cross validation
*10. 2010 Crott and Briggs [111]	448	Breast	QLQ-C30/ EQ5D-3L	0,1,2,3,6,12,20 28,36,42,48,54	Quadratic	80%	European	Independent data
*11. 2010 Jang et al. [130]	172	NSCLC	QLQ-C30/ EQ5D-3L	Unspecified	OLS	58%	Canada	Cross validation
*12. 2009 Kontodimopoulo et al. [131]	48	Gastric	QLQ-C30/ EQ5D-3L	Interviewed post-surgery	OLS	91%	Greece	None
*13. 2009 McKenzie et al. [132]	199	Oesophageal	QLQ-C30/ EQ5D-3L	Approx. monthly	OLS PROBIT	61%	UK	Independent data
14. 2007 Wu et al. [144]	280	Prostate	QLQ-C30 FACT-P/ EQ5D-3L	0,3,6, 12	OLS Censored	58%	Multinational	Cross validation

Table 2.2: Summary of Relevant Mapping Algorithms from 443 initial articles on mapping[#]

[a] maximum observed was reported
[#]last search conducted in January 2017
*Relevant to Chapter 10

Limitations of mapping and gaps in research

Authors of articles on mapping noted that the development of mapping algorithms is possibly more complex, citing the need for investigation of complex model structures. The inter-relationship between generic and CSM is likely to be more complex than simple linear models. In addition, no post-progression mapping models were reported (i.e. after the disease has progressed). This is important because pre-progression and post-progression QoL are important in determining the QALY for cost-effectiveness of cancer treatments. Interestingly, at the time of starting this thesis, no mapping algorithm in lung cancer patients was developed [111,112].

Mapping algorithms over-predict at poorer health states, but reasons for this remain unknown. For example, Joint models that incorporate adverse event/toxicity to address the over-prediction of poorer health states have not been considered; no post disease progression mapping algorithms have been developed. There is limited use of Bayesian and Non-Linear mapping Models; an absence of simulation to quantify the uncertainty of mapping algorithms is notable; no mapping algorithms between EQ-5D-5L and QLQ-C30 existed at the time of writing this thesis. In addition, a number of HTAs in cancer (specifically lung cancer as an example) [23-41] report utilities from external studies (direct elicitation) were considered which were criticized for over-estimating utilities in economic evaluations. Direct elicitation studies (utility studies) were criticized for being poorly controlled studies. This is another reason for research in this area. Utility studies were not recommended and mapping appeared to be preferred [23-41].

In summary, further research in mapping functions is proposed for specific research areas (Longworth and Rowen, 2013) [115]:

(i) Mapping general population visual analog scale (VAS) for both measures alongside each other.
(ii) The mapping between Rasch scores and utility scores.
(iii) Using Gaussian Processes.
(iv) Single equation and two-part beta regression models.
(v) Measurement error models.
(vi) Investigation of Bayesian Networks.

To this end, the development of advanced models will be used based on data collected in NSCLC patients to explore alternative mapping algorithms as well as compare them using the more recent EQ-5D-5L.

2.4.2 Summary of Literature Review of Bayesian Algorithms (Chapter 7)

Bayesian methods are being increasingly used in economic evaluation methods. There have been some suggestions Bayesian methods may result in improved mapping [139, 146]. An important feature of Bayesian algorithms is the potential to estimate profiles of health states and not just utility values. Profiles may offer a richer insight and information around the estimated (true) utility values.

Only one Bayesian mapping algorithm was used in cancer mapping the EQ-5D-3L from QLQ-C30 in a myeloma patient population. There were a total of 9 Bayesian mapping algorithms (in varying disease areas and health populations). Basu & Manca (2012) use a Bayesian form of a beta regression (using non-informative priors) in a non-mapping context, compared to an OLS regression model [145]. A 2-part version of the beta model in order to handle the over-dispersion (e.g. over-dispersion of 1's on the EQ-5D-3L scale) of values was used [145].

Kharroubi et al. (2007) suggested a more complex simulation approach in order to better understand the predictive value of the realized mapping function [139] in an attempt to reduce the systematic errors found in mapping functions. However, this may be difficult in practice because one may need to consider setting prior distributions in a multivariate context. This means one would need to set prior distributions for each of the 15 coefficients (e.g. for the QLQ-C30) with potentially large variances (for non-informative priors). In addition, to determine the prior covariance structure of a 15x15 matrix would be a considerable challenge. Generating posterior (mean) coefficients from published algorithms, for use in future predictions may be difficult in practice.

Other Bayesian algorithms [146] were slightly less complicated because the condition specific instrument had fewer coefficients from which to map (using the HAQ); however, in most cases, algorithms were more complex and non-informative priors were used. One such Bayesian approach was to use a Bayesian Networks [147]. The Bayesian Network (BN) approach is based on a probabilistic approach rather than estimating utilities through use of (posterior) mean coefficients. Conditional and joint probabilities of responses using Bayes theorem would allow for computation of posterior probabilities of each EQ-5D-5L (or EQ-5D-3L) response and consequently utilities [148]. However, in this model too, predictions at poor health states in a non cancer context were reported as inadequate, although performed better than other models. In Chapter 7, the use of this approach to mapping with the QLQ-C30 is considered as a 'first' in the application of cancer and with the recent EQ-5D-5L.

2.4.3 Summary of Literature Review: Sensitivity and Responsiveness of EQ-5D (Chapter 8)

Some generic instruments may not be adequate to demonstrate HRQoL benefits, compared to condition-specific measures (CSM) [149-151]. This is because a brief and standardized HRQoL instrument across diseases will lack sensitivity due to its nature. Concerns about the sensitivity and responsiveness of generic measures such as EQ-5D-3L [152,153] have been important enough to lead to the development of the EQ-5D-5L, using a 5 point scale (EQ-5D-3L has a 3 point scale). The issue becomes more acuter and relevant, where a CSM appears to offer an interpretation for a treatment benefit inconsistent with a generic one (or vice versa). Therefore, it is important to explore sensitivity and responsiveness of HRQoL instruments, particularly in a large cancer population, such as lung cancer [154].

Concerns have also been raised about the sensitivity of the EQ-5D-3L and by extension to the derived mapped utilities [151-153]. Most mapping algorithms using the EORTC-QLQ-C30 (QLQ-C30) are based on EQ-5D-3L. Given the reported limitations and criticisms levelled against the EQ-5D-3L and the consequent development of the EQ-5D-5L, a mapping algorithm for the EQ-5D-5L should be an important area of research.

It is also unclear, not just in cancer studies, but also in many trials with HRQoL endpoints as to what a relevant clinical or economic treatment difference size is. Moreover, which among the common HRQoL measures (in cancer) are more sensitive to detecting HRQoL benefits, remains highly uncertain [60,155]. This is relevant whether the HRQoL is a primary or secondary outcome because HRQoL effects are often considered to 'add value' to an intervention, especially when the primary clinical outcome result shows modest or borderline benefits. For example, with QLQ-C30, there are 15 possible effect sizes and only one for EQ-5D. The precise clinical or economic relevance of effect sizes from the QLQ-C30 still remains unknown or not well understood. Maringwa et al. (2011), suggest 'important' effect sizes of varying magnitudes [60]. A suggested effect size of 10 points (it is unclear whether this is a particular domain or any domain of the QLQ-C30) is assumed to be an important difference [156] and it is unclear why this should be the case. This suggested ‘relevant’ effect size (10 point difference) was made many years ago at a time when cancer treatments were comparative to older less effective drugs or even placebo. The standard of care has since improved. For EE the ‘best’ standard of care is required as a comparator which is likely to result in smaller treatment benefits (and QALYs), and therefore, this definition, despite being widely used is likely to be redundant [155].

Wiebe et al. (2003) [207] argued that condition-specific measures (CSM) are likely to be more responsive than generic measures. However, previous research concludes that the responsiveness of the generic EQ-5D-3L was similar to QLQ-C30, using the data from patients with liver metastases [208]. Certain generic instruments may not be adequate to demonstrate HRQoL benefits compared to CSMs [149-152]. A brief and standardized HRQoL instrument across diseases will lack sensitivity because of its inherent nature [151]. Concerns about the sensitivity and responsiveness of generic measures like EQ-5D-3L [153,154] have been important enough to lead to the development of EQ-5D-5L with a 5 point scale (the EQ-5D-3L has a 3 point scale). The issue becomes more acute and relevant, where a CSM appears to provide an interpretation of a treatment benefit, which is inconsistent with a generic one (or vice versa). Therefore, it is vital to explore sensitivity and responsiveness in a cancer population, in particular, lung cancer, which is responsible for most deaths among cancer patients [209]. Establishing the validity and reliability of HRQoL instruments is not sufficient in itself to determine the responsiveness and sensitivity to treatment [210].

Generic measures are considered to lack the sensitivity to detect HRQoL benefits when compared to CSMs. However, if the treatment benefits from a CSM and generic measure are similar, then a simpler and shorter preference-based generic measure (e.g. EQ-5D) could be used alongside a highly condition specific measure (e.g. such as the lung cancer symptom specific LCSS questionnaire) without losing much information. With preference-based measures, there may be a concern that preference weights based on the general population may not reflect the similar relative importance for certain health states that a cancer sufferer might have. A comparison of relative effect sizes between generic measures and CSMs have not been thoroughly evaluated in cancer, particularly with respect to their implications for the QALY. However, suggested ‘important’ effect sizes for the EQ-5D have been suggested to range from 0.03 to 0.074 [157, 158]

In a systematic review of 43 published articles in NSCLC [49], 28 of these studies used QLQ-C30 with the objective of detecting clinical improvements in HRQoL [49]. Among these 28 studies, the vast majority of studies (>80%) did not report improvements with the QLQ-C30, either between treatments or relative to baseline. Moreover, wherever an effect was reported, the sample size was small (e.g. n=19 in the study of Bianco, 2010 cited in the review) [49]. Khan et al. (2015) more recently indicated that treatment effect sizes are rarely large [155] from cancer specific CSMs.

Of the 28 trials (Damm, 2013) which used QLQ-C30, no generic HRQoL was included, yet an economic evaluation of some form was performed on the trial data at a later point (e.g. the BR21 trial [49,107]). Moreover, conclusions regarding HRQoL benefits were provided in terms of “non-worsening HRQoL”- and little or no improvements in HRQoL. The conclusions were often presented such that if patients did not deteriorate in their HRQoL, then this was interpreted worthy of comment or a favorable outcome (see [49] for example). It can be complicated when a CSM suggest a HRQoL improvement and a CSM does not. Given that utilities for the EQ-5D are based on societal preferences and the CSM is based on descriptive assessments, the need for an evaluation of the sensitivity of the recent EQ-5D-5L and the QLQ-C30 is necessary to *inform* an economic evaluation, particularly where the conclusions are borderline (e.g. ICERs close to £20,000 or £30,000/QALY).

Therefore, in summary, the sensitivity of EQ-5D-3L versus EQ-5D-5L versus EORTC-QLQ-C30 has not been addressed extensively; EQ-5D is considered to lack sensitivity for measuring changes in health states in a cancer setting [156]. Small, but important differences in HRQoL should not be ignored [155] and investigated with a view to identifying the implications for an economic evaluation.

2.4.4 Summary of Literature Review: Utility Extrapolation (Chapter 9)

The QALY in cancer trials is often computed as a weighted measure of pre and post-disease progress (PP) survival. Therefore, how post-progression utilities are estimated, either directly or by extrapolation, influences the overall QALY. Although post-progression survival may result in an increase in total QALYs, it is important to understand the impact on the incremental QALY.

For the more recent class of anti-cancer treatments, such as immunotherapies (e.g. Nivolumab), the argument is that ‘pseudo-progression’ (a potentially incorrect conclusion of progression) occurs and that treatment should continue beyond it. Utility data are (often) not collected after the first instance of so-called ‘pseudo’ progression (because some patients may stop treatment) [22]. However, if patients continue to take treatment in anticipation of later benefits, after disease progression, utility estimates become even more important for estimating the post-progression QALY. Such post-progression utilities can be determined from external literature or as proposed in this thesis through extrapolation methods. More importantly, the continued side effects of toxicity from treatments after disease progression (such as hair loss, and other longer term toxicities) can have a continued impact on HRQoL well after the first instance of progression. For these reasons models for predicting utility between disease progression and death are important. Estimating utilities after disease progression through extrapolation has not been considered previously.

When time to progression is short, but the post-progression survival (PPS) is relatively longer, utility data are likely to be missing. Typically, in most cancer studies/trials, utility data are collected until disease progression and not until death. However, in some cases, for some patients, data may be available beyond progression. A recommendation should be to collect utility data beyond progression, particularly if treatment is to be continued beyond progression (e.g. as in the case of more recent immunotherapy advances like Nivolumab [14, 22]). When some patients have utility data after disease progression, this would offer an opportunity to ‘borrow strength’ from between and within patients to model and consequently predict (extrapolate) utility data after disease progression.

Presently, minimal or no data is available for methods in utility extrapolation beyond disease progression. Most methods focus on using estimates from external sources [108, 160]. Stein et al. (2014) [161] report real world utility study in colorectal cancer in a second line setting. However, this is quite different from using a modelling approach to extrapolate utility using patient-level data from a clinical trial. This will be the focus of Chapter 9.

In summary, modelling HRQoL during the end of life (often after disease progression) requires one to explore the relationship between utilities, PD and the probability of death at specified time points. Methods for modelling utility post cancer progression have not been extensively used or are non-existent. Moreover, there is considerable uncertainty in the literature between published values for patients in the pre and post-disease progression stages of their cancer, as well as with utilities for toxicities as a result of their treatment.

2.4.5 Summary of Literature Review: Model Selection (Chapter 10)

While several algorithms to predict utilities are available, classifications of algorithms in terms of performance are still in development. Recently Crott (2014) [161], Arnold et al. [162] and Doble & Lorgelly [163] tested the external validity of some of these algorithms. For instance, an algorithm might report an R^2 of 75% with accurate mean predicted utilities when applied to the same data the model was generated from. However, when this algorithm is applied to an independent dataset, the predicted mean utility might be poor. Only when a mapping algorithm has been tested on a large number of data sets with known observed utility values, judgments about predictive performance can be made. No criteria are available for *deciding* between published algorithms, other than the metrics (e.g. R^2, RMSE) used for testing the algorithm at the time of development. External validity of *published* algorithms was a relatively recent development at the time of writing this thesis.

From 14 published algorithms identified, e*ight* mapping algorithms between EQ-5D-3L and QLQ-C30 were considered. These used the QLQ-C30 and reported at least 2 coefficients. Two algorithms could not be used as they are published as part of this thesis [109, 137]. Only three articles on mapping algorithms were from an NSCLC population, out of which 2 are published as a result of this thesis. A review of these 8 algorithms was provided in Table 2.2:

(i) McKenzie et al. (2009) [132]

This study involved patients with oesophageal cancer (199 UK patients). Data were collected from an RCT with follow up of at least 90 weeks. HRQoL (including QLQ-C30) were collected at 31-time points (baseline, every 3 weeks to at least 90 weeks). Validation of the algorithm was carried out using an independent data set.

(ii) N.Kontodimopoulos et al. (2009) [131]

This was a gastric cancer sample of 48 patients (in Greece). Data were collected from a utility study between November 2007 and March 2008. HRQoL were collected at 2-time points (before and after treatment). Independent data were not used to validate the algorithm.

(iii) Jang et al. (2010) [130]

Jang et al. (2010) mapped EQ-5D-3L in 172 Canadian NSCLC patients using outpatient data (single cohort/cross section) from a single visit. Methodological details were limited in this article.

(iv) Crott and Briggs (2010) [111].

This was a breast cancer sample of 448 patients from several European countries. Data were collected from a randomized controlled trial (RCT) over 54 months. HRQoL were collected at 11-time points (baseline, once per month for the first 3 months (and 6, 12, 20, 28, 36, 42, and 48 up to 54 months thereafter). Independent data were not used to validate this model.

(v) Eun-ju Kim et al. (2012) [142]

This algorithm was developed using data from 199 Korean metastatic breast cancer patients. Data were collected prospectively and HRQoL were collected before and after the patients were diagnosed with breast cancer. Independent data were not used to validate the model.

(vi) Versteegh et al. (2012) [141]

HRQoL data were taken from two separate trials in multiple myelomas (HOVON 24 was a trial with, a sample size of 137 German patients) and non-Hodgkins lymphoma (HOVON 25). The algorithm was developed from HOVON 24 using data from HOVON 25 for validation.

(vii) Kim SH et al. (2012) [143]

A model based on 893 Korean patients with various tumor types was tested in 123 patients with colon cancer. The two data sets were independent. All patients were from a single hospital and independent data were used to validate the model.

(viii) Proskorovsky et al. (2014) [140]

This is an algorithm developed in 154 multiple myeloma patients (89 UK and 64 Germany). Data were from a multinational cohort study. Independent data were not used to validate the model*.*

In summary, 8 algorithms were identified from which a selection procedure could be developed. The development of a selection procedure is needed to identify the more useful algorithms.

2.4.6 Overall Conclusion from Literature Review

The main conclusion from this review is that further development of algorithms, including Bayesian algorithms, could be developed. Moreover, comparisons with the EQ-5D-5L along with its sensitivity is important. Methods for extrapolation of utility beyond PD (particularly in a real world setting) and criterion for selection of useful algorithms are needed. Additional gaps in knowledge identified from a review of the articles include the need for developing methods for adjusting utilities for treatment switching. When patients switch from the standard treatment to the new one (or vice versa), the impact of the switching is not reflected in the QALY or economic model. Handling switching for survival methods is well documented [164], however, for utilities and HRQoL methods are not developed. However, the methods for utility extrapolation may be speculative and a narrower focus on mapping and modelling HRQoL will be the object of this thesis. To this end, the aims and objectives will now be stated.

2.5. Aims and Objectives

2.5.1 Overall Aim of thesis

The overall aim of this thesis is to improve the estimation and valuation of HRQoL benefits in lung cancer for the purposes of economic evaluation. The aims and objectives of this thesis are answered in 7 further chapters (including a concluding chapter: Chapter 4 develops and compares existing models with a new non-linear Beta-Binomial mapping algorithm; Chapter 5 evaluates other mapping algorithms developed from the more recent EQ-5D-5L; Chapter 6 explores other reasons why algorithms may over-predict at poorer health states; Chapter 7 involves the use of Bayesian networks ;Chapter 8 compares the sensitivity and responsiveness of generic and condition-specific measures, particularly EQ-5D-5L, EQ-5D-3L, and EORTC-QLQ-C30. Finally, Chapter 10 proposes a selection procedure, which separates ‘useful’ algorithms from ‘not useful’ ones. To this end, several objectives with specified aims on the theme of modelling HRQoL for the economic evaluation of cancer treatments are outlined.

2.5.2. Underlying Thesis Questions

The key conclusions of the literature search suggested that evidence in some important areas of research were needed. The literature on research themes identified areas such Bayesian mapping algorithms, utility extrapolation, mapping algorithms with the recent EQ-5D-5L and comparison of the sensitivity of the EQ-5D-5L, EQ-5D-3L, and condition specific measures in cancer, was very limited or non-existent. Improvements in methodological and empirical evidence were needed to address some important research questions such as:

(i) Can mapping algorithms be improved with more complex model structures?

(ii) Is it better to use a mapping algorithm from the EQ-5D-5L or the EQ-5D-3L, given a choice?

(iii) Is the more recent EQ-5D-5L sensitive to detecting treatment benefit compared to a CSM and the EQ-5D-3L? What implications does this have for an economic evaluation?

(iv) After patients progress in their disease, can utilities be extrapolated to estimate the cost-effectiveness of new treatments over a life time horizon?

(v) Given the existence of several mapping algorithms, can a selection criterion be developed to identify a more or less useful algorithm?

Based on the literature research and summary of Chapter 2, this thesis will, therefore, examine several areas associated with modelling HRQoL for the economic evaluation of treatments for cancer, with a particular application to lung cancer data sets. To achieve this purpose, seven primary aims are considered:

Aim 1: **To undertake a literature search of methods and approaches used in modelling HRQoL for Economic Evaluation in the context of cancer.**

Objectives of Aim 1 (OA1)

- **OA1.1**: To identify the questions and gaps in knowledge relating to modelling HRQoL from cancer patients in the context of mapping.

- **OA1.2:** To identify the questions and gaps in knowledge relating to Bayesian approaches to mapping.
- **OA1.3:** To identify the literature that compares the sensitivity and responsiveness of the EQ-5D-5L, EQ-5D-3L, and a cancer specific measure.
- **OA1.4:** To identify literature and gaps in knowledge relating to estimating utility for economic evaluation after cancer progression.
- **OA1.5:** To identify literature and gaps in knowledge in terms of objective criteria for selecting mapping models when there are several to choose from.

Aim 2: To investigate methods to improve the performance of mapping algorithms for estimating utilities for economic evaluation (Chapters 4, 5 and 6).

Objectives of Aim 2 (OA2)

- **OA2.1:** To develop and test a novel non-linear Beta-Binomial (BB) approach which takes into account the over-dispersion and skewness of utility data.

- **OA2.2:** To develop and test mapping algorithms that maximize the relationship between clinical, demographic and toxicity factors, through joint modelling.
- **OA2.3:** To compare the performance of mapping algorithms between the recent EQ-5D-5L with the previous EQ-5D-3L.

Aim 3: To test and compare the performance of a Bayesian Mapping Algorithm (Chapter 7)

Objectives of Aim 3 (OA3)

- OA3.1: To develop and test a Bayesian Mapping Algorithm between QLQ-C30 and the EQ-5D-3L.
- OA3.2: To develop and test a Bayesian Mapping Algorithm between QLQ-C30 and the EQ-5D-5L.

Aim 4: To explore the sensitivity and responsiveness of a disease-specific and generic measure using the EQ-5D-5L, EQ-5D-3L and the QLQ-C30 (Chapter 8).

Objectives of Aim 4 (OA4)

OA4.1: To determine whether clinical benefits valued from the EQ-5D-5L, EQ-5D-3L, and those obtained from the cancer-specific QLQ-C30 are similar.
OA4.2: To investigate whether HRQoL benefits from a CSM and those valued from the generic measure are adequate to reflect QALYs.
OA4.2: To compare the HRQoL benefits valued from generic and those obtained disease-specific measures, using standard measures of effect size.

Aim 6: To investigate a method of selection between published algorithms (Chapter 9).

Objectives of Aim 6 (OA6)

OA6.1: To identify relevant published mapping algorithms from the QLQ-C30
OA6.2: To test published algorithms on independent data sets
OA6.3: To develop a decision and classification criteria for published mapping algorithms using simulation.
OA6.4: To present a recommendation of potentially usable algorithms.

Aim 7: To design, conduct and analyse a prospective observational study in cancer patients for collecting HRQoL data using three measures: EQ-5D-3L, EQ-5D-5L, and QLQ-C30 for answering some of the questions identified from the literature review.

Objectives of Aim 7 (OA7)

OA7.1: To collect data from cancer patients in a well designed observational study for investigating the previously defined aims and objectives

2.5.3 Data Sources

The data for this thesis consists of three data sets. Two of these data sets were initially provided by the Cancer Research UK (CRUK), Cancer Trials Centre from two randomized controlled trials, namely data from the TOPICAL and SOCCAR trials [165, 166]. Both the trials were sponsored by CRUK. A third data source, which was a prospectively designed study (Study 3) specifically to meet the objectives for chapters 5, 6, 7, 8, 9 and 10.

2.5.4 Conclusion

In conclusion, based on a detailed review of the literature, a number of objectives were identified relating to: improved mapping methods, comparing mapping with the

EQ-5D-5L, developing methods for utility extrapolation after disease progression, sensitivity of generic versus condition specific measures (particularly EQ-5D-5L) and how to go about selecting a suitable mapping algorithm.

In the following chapters, the aims and objectives associated with improved methods to predict EQ-5D utilities for economic evaluation are considered beginning with a novel approach to mapping using a Beta Binomial modelling approach.

Chapter 3

Chapter 3: Methods

Abstract

Introduction: This methods chapter provides an overview of the key methods used in later chapters for deriving the main conclusions. The methods are chosen based on a review presented in chapter 2.

Methods: Several models were identified that are pertinent to modelling HRQoL for mapping. The models range from simpler Linear regression models to the most advanced and less well known such as Joint Models that model two outcomes simultaneously to predict EQ-5D-utilities. The methods for model fit, model testing and performance have been described including how utilities and QALYs are derived. Advantages and disadvantages of each model have also been provided.

Results: The models identified include (i) Linear Regression (simple and mixed modelling framework), (ii) Beta Binomial models, (iii) Non-Linear (Quadratic), (iv) TOBIT, (v) Quantile, (vi) CLAD, (vii) LDVM, (viii) Joint Models and (ix) Bayesian Networks. The main advantage of the more complex models is improved model fit and better prediction with robust estimates of co-efficient, leading to less uncertainty around the true QALY. The key disadvantages are the complexity of the model and in using the mapping algorithm in practice.

Conclusions: Each of the models has advantages and disadvantages. Whether the disadvantage of greater complexity is offset by improved estimates of utility remains the subject of subsequent chapters this thesis will address.

This chapter details the methods used for reporting the results of each chapter. Since the data used in chapters was the same the methods with respect to data collection and study design were the same across chapters. However the analyses methods naturally differ due to the modelling techniques employed.

3.1 Population

For this thesis the population of patients for which data were collected were NSCLC patients followed up in 3 different studies.

(i) *The TOPICAL Study*

This trial was a RCT comparing Elrotinib for the treatment of NSCLC in elderly patients who were unfit for chemotherapy. . Data were provided by the Cancer Research UK (CRUK) Cancer Trials Centre (CTC) and I was the lead statistician and health economist for this study. The results have already been published [165]. In this RCT, patients included in the trial were recruited from 78 UK hospital sites in the UK. The patient population included in this analysis were newly diagnosed stage IIIb–IV (pathologically confirmed) patients with NSCLC who were chemotherapy naïve with no symptomatic brain metastases, and deemed unsuitable for chemotherapy by treating physicians based on the Eastern Cooperative Oncology Group (ECOG) performance status (PS ≥2) and/or multiple medical comorbidities including renal impairment and estimated life expectancy of at least 8 weeks. The objective was to compare the efficacy of erlotinib (a treatment for NSCLC) with standard treatment (placebo plus best supportive care).

(ii) The SOCCAR Study

This trial was conducted by the CRUK CTC and I was the lead statistician and health economist for this study. Data were provided by the CRUK CTC. The results from the main paper and secondary papers have already been published during the course of writing this thesis [166].

Patients included were recruited from UK sites with histologically or cytologically confirmed stage III NSCLC, Performance status - ECOG 0 or 1, Life expectancy greater than 3 months, No prior chemotherapy, radiotherapy or investigational agents, patients willing and able to give informed consent, patients considered able to tolerate platinum based chemotherapy and radical radiotherapy, adequate renal function. Further details are found in [166].

(iii) Study 3

This trial was designed by me along with Dr Joe Maguire (acknowledgment). Local ethics approval was given by the NHS research and ethics committee (REC) reference number: LH/56/2014, for the design of this study. I entered the data in a Microsoft ACCESS database and performed all analyses. Data were provided by the CRUK CTC and I was the lead statistician and health economist for this study. The results have already been published [137].

Patients included were those aged >18 with histologically or cytologically confirmed stage III NSCLC, ECOG 0-4, Stage I-IV, able to give informed consent and attend routine assessments. These were set broad to reflect the real-world NHS setting as closely as possible. This enhanced generalizability allows estimates of utilities to be used as inputs for future economic evaluation. For the purposes of this thesis only EQ-5D-3L, EQ-5D-5L and QLQ-C30 will be used. The data for other HRQoL and health resource use would be evaluated separately from this thesis.

This study is arguably the first study designed to investigate and compare condition-specific and generic measures in a real-world setting, using *both* EQ-5D-3L and EQ-5D-5L data.

3.2 Study Design

TOPICAL: TOPICAL was a double-blind, randomised (1:1), placebo-controlled, phase 3 trial, done at 78 centres in the UK. Patients were randomised to receive best supportive care plus oral placebo or erlotinib (150 mg/day). Patients were stratified by disease stage, performance status, smoking history, and centre. Patients were followed up until death or progression (whichever occurred first). Investigators, clinicians, and patients were masked to assignment.

SOCCAR: Patients were randomly assigned, in a 1:1 ratio, to receive sequential or concurrent chemo-radiotherapy following a dynamic allocation method. The method of minimisation to stratify for: Radiotherapist, Staging (Stage IIIa, Stage IIIb N3, Stage IIIb Not N3), ECOG Performance (0, 1), Histology (Squamous, Adenocarcinoma, Large Cell, Other NSCLC) and Weight Loss (> 5 %, < 5 %, Unknown) – a total of 5 stratification factors.

Study 3: This was a single cohort prospective (non-interventional) observational follow-up study in 100 NSCLC patients. Patients were follow-up during their routine anti-cancer treatment and cancer management for a period of at least 12 months.

The primary objective of the study was to assess the HRQoL and health resource use using several HRQoL instruments, including EQ-5D-5L, EQ-5D-3L, QLQ-C30, LCSS (Lung Symptom Specific Questionnaire), Hospital Anxiety and Depression Scale (HADS). The primary outcomes were QLQ-C30 and secondary outcomes were EQ-5D, HADS, and LCSS.

3.3 Interventions

TOPICAL: Patients were to take oral erlotinib or matching placebo daily, 1 hour or more before food, or 2 hours after food. The dose could be reduced to 100 mg, then 50 mg in cases of substantial toxic effects. Treatment continued until disease progression, adverse side-effects judged by the treating clinician to warrant discontinuation, or patient withdrawal. Patients continued to receive active supportive care, including palliative radiotherapy, at the discretion of their clinician.

SOCCAR: Patients randomized to the sequential group were given cisplatinum intravenous (I)V as 80 mg/m2 on day 1 and vinorelbine IV 25 mg/m2 on days 1 and 8 for 4 cycles, followed by 55Gy in 20 fractions over 4 weeks. In the concurrent group, vinorelbine 15mg/m^2 was given prior to radiotherapy fractions 1, 6, 15 and 20 whereas cisplatinum was given as 20 mg/m2 with radiotherapy fractions 1-4 and 16 – 19. A further 2 cycles of vinorelbine 25 mg/m2 (days 1 and 8) and 80 mg/m2 (day 1) cisplatinum were given after concurrent chemo-radiotherapy. Radiotherapy was given no more than 6 hours after starting the cisplatin infusion. Patients in both arms of the trial were scheduled to receive a total dose of up to 320 mg/m2 of cisplatinum and a radical radiotherapy schedule comprising 55 Gy in 20 fractions over four weeks. Patients were scheduled to start chemotherapy within four weeks of randomisation and within two weeks of a clinical assessment of fitness.

For the sequential and concurrent group, respectively, patients were scheduled to receive a total of four cycles of cisplatinum and vinorelbine. In the concurrent group, a reduction in the total vinorelbine dose was allowed to permit concurrent administration of radiotherapy with chemotherapy.

Study 3: Patients received their standard chemo-therapy. This was either cisplatin IV as 80 mg/m2 on day 1 and vinorelbine IV 25 mg/m2 on days 1 and 8 for a maximum of 6 cycles (these were standard at the sites). The chemotherapy doses were based on the patient's calculated pre-treatment body surface area using actual body weight. Patients were scheduled to start chemotherapy within four weeks of randomisation and within two weeks of a clinical assessment of fitness.

3.4 Sample Size

TOPICAL: The target sample size was 664 patients, on the basis of the primary study objective to detect an increase in 1 year overall survival from 10% with placebo to 17 ・5% with erlotinib (equivalent to a HR of 0.75 and much the same as that achieved with chemotherapy vs supportive care), with 90% power and 5% two-sided test of significance.

SOCCAR: The target sample size was 130 patients. This sample size would allow the mortality rate to be characterised with acceptable accuracy using a 95% confidence interval (see ref [166])

Study 3: Sample size was determined using an estimation approach, since this was not a study powered to detect differences between treatment groups. Hence, assuming a SD in the Global QoL score at baseline from the SOCCAR trial of about 25 points, with a sample size of about 100, the 95% confidence interval (2 sided, 5% alpha) for the observed change from baseline will lie within 5 points of the observed mean change at 6 months. That is we can be about 95% sure that a sample size will be large enough to give us a good precision around the true change from baseline.

3.5 Outcomes

TOPICAL: Primary outcome: Overall survival (OS). Secondary outcomes: Progression-free survival (PFS), tumour response, HRQoL using QLQ -C30, LC-14 and EQ-5D-3L, health resource use. Pre-specified subgroups included: sex, histological examination, activating EGFR or KRAS mutation, stage, smoking status, ECOG score, and development of first-cycle rash.

SOCCAR: The primary endpoint was treatment related mortality (any cause) defined as an SAE that results in death; and is definitely, probably, or possibly related to any of the trial therapies. Toxicity was assessed according to NCI Common Terminology

Criteria for Adverse Events v 2.0. Secondary endpoints include Overall Survival (OS), progression free survival (PFS) calculated from the date of randomisation to the date of first clinical evidence of progressive disease, or death; Local PFS calculated from the date of randomisation to the date of first clinical evidence of progressive disease at the primary site, or death; tumour response according to RECIST (version 1.0) including the best response in the first 6 months; HRQoL were assessed using the EORTC QLQ –C30, LC-14 and EQ-5D-3L

Study 3: The primary outcome was the QLQ-C30 (Global outcome) at 6 months. Secondary outcomes included QLQ-C30, EQ-5D-3L, EQ-5D-5L, Hospital Anxiety Scale (HADS), SF-36, LC-14, overall survival, progression free survival, adverse events (AE) and health resource use. Adverse events (AEs) and health resource use were collected as and when they occurred. The AEs were graded using National Cancer Institute's (NCI) Common Toxicity Criteria (CTC) Version 2.0 NCI CTC criteria from Grade 1 to Grade 5 (Death).

Adverse events (AEs) and health resource use were collected as and when they occurred. The AEs were graded using National Cancer Institute's (NCI) Common Toxicity Criteria (CTC) Version 2.0 from Grade 1 to Grade 5 (Death). All adverse events were categorized as Grade 0: No adverse events or laboratory data (e.g. CD4 count, neurophils etc.) are within normal limits; Grade 1: Mild Adverse events; 2: Moderate Adverse event; 3: Severe and undesirable adverse event; 4: Life threatening or disabling adverse event; 5 : Death related adverse event. Adverse events were collected prospectively and paired observations of grade 3 and above AEs that correspond to EQ-5D-3L, EQ-5D-5L, and QLQ-C30 data were available for each patient. Adverse events are commonly reported in studies and trials by reporting the worst (maximum) grade – as a single patient can have multiple AEs of varying grades and the worst grade (maximum grade) is used as a response for later modelling. Patients were classified as having a grade 3 or higher AE or not (i.e. dichotomized) for the purposes of joint modelling (Chapter 6).

3.6 HRQoL Assessments

TOPICAL: Patients were followed until disease progression or death. Patients completed HRQoL assessments (QLQ C-30 and LC14, and EQ-5D-3L) at baseline, monthly during the first year, then 18 and 24 months after randomisation.

SOCCAR: Follow up for assessments from randomization continued until August 2011 where the database was closed for analysis. Patients completed HRQoL) assessments using QLQC-30 and EQ-5D at baseline, 3 weekly during treatment, then monthly up to 6 months after randomisation, then 3 monthly for 2nd year, 6 monthly in year 3 and annually thereafter.

Study 3: Assessments for HRQoL (including EQ5D-3L, EQ-5D-5L and QLQ-C30) were planned for collection at baseline and every month for at least 12 months. The EQ-5D-3L and QLQ-C30 were given at the same time and the EQ-5D-5L between 1 to 2 weeks later to avoid potential recall bias and avoid the potential for ‘carry over’. Patients were given the HRQoL forms to take home and were returned by post or when they visited the hospital next (where they were completed with the nurses in the clinic). They were instructed to complete the EQ-5D-3L in the first week and the EQ-5D-5L in the following (or third) week of each month.

The delay of 2 weeks was not considered to be a significant difference in assessing HRQoL because these patients were newly diagnosed NSCLC patients and HRQoL deterioration over one to two weeks was not considered to impact the results in any meaningful way. The order of assessing 3L and 5L was not randomized due to practicality and the reasoning that a sequence (order effect) was likely to be minimal over a short period (1 to 2 weeks) of time.

3.6.1 The EORTC-QLQ-C30

(i) EORTC-QLQ-C30 Generic Cancer Instrument

The EORTC-QLQ-C30 (QLQ-C30) is a ‘generic’ cancer instrument [33, 34] consisting of 30 questions, out of which 28 questions are measured on a 4 point scale (‘not at all’ (1) to ‘very much’ (4)) and 2 questions are measured on a 7 point scale. Although it is generic across cancer types, it is not a generic instrument across all disease areas. Details of the QLQC-30 were provided in Section 1.4.1.

3.6.2 The EQ-5D-3L

The EQ-5D is a widely used generic measure, which is the shortest and perhaps the least cognitively demanding instrument that appears to be at least as responsive as the other community (preference) weighted instruments (Brazier et al., 2009) [35b]. EQ-5D-3L consists of a descriptive health state classification system with five questions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), measured on three severity levels - ‘no problems’, ‘some problems’ and ‘extreme problems’. A health state defined by the descriptive system of EQ-5D can be described by a five-digit number. For instance, 12113 refers to a patient, who has no problems with mobility (1), some problems with self-care (2), no problems for usual activities (1) or pain/discomfort (1) and extreme problems with anxiety/depression (3). Combining one level from each question defines 243 different possible health states from 11111 to 33333. Details of EQ-5D-3L were provided in Section 1.4.1.

3.6.3 The EQ-5D-5L

EQ-5D-5L is a revision of EQ-5D-3L. It consists of five questions, identical to EQ-5D-3L (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), but with an expanded 5 point scale and slightly different descriptors for each of the levels compared to the 3 point scale of the EQ-5D-3L [36]. For Mobility, Self-Care and Usual Activities, these are: 1: “No Problems”, 2: “Slight Problems”, 3: “Moderate Problems”, 4: “Severe Problems” and 5: “Unable to”; for the Pain/Discomfort and Anxiety/Depression scale, these were: 1: ‘No’, 2: ‘Slight, 3: ‘Moderate’, 4: ‘Severe’ and 5: ‘Extreme’. The scores are on a 5 point scale 1 to 5 (for each of the 5 domains). Details of EQ-5D-5L were provided in Section 1.4.1.

3.7 Analysis Methods

Analyses methods depended on the modelling approach for each chapter.

3.7.1 Modelling

Table 3.1 below shows for each chapter the range of modelling methods used along with assumptions, advantages and disadvantages of the models used. The fuller details of methodology are provided in a separate methods section in each chapter.

3.7.2 Criteria for Model Fit and selection

For all models, standard model selection criteria were used. These were Aikakes information Criteria (AIC) where the smaller the value, the better the fit. This was used for all models so that a valid comparison can be made. Although R^2 is used as a measure of model fit, for non-linear models it may not be so relevant [ref]]. In addition to these two measures, the %predicted outside the valid EQ-5D range (i.e. below -0.549 and above 1.0) is also important because models should restrict predictions to valid ranges. A further measure includes root mean squared error (RMSE) or residual error. The smaller this is, the better the fit. The mean absolute error (MAE) is similar to the RMSE in that it is the difference between the individual predicted values versus the observed value (with squaring). A further measure was the mean of the predicted utilities including its standard error (SE). In addition, the proportion the observed and predicted differed (expressed as a percent) was also computed. This latter value is considered more useful that the predicted

Models for Mapping	Chapter	Reasons for selection	Advantages	Disadvantages
Linear	4	-Most commonly reported model and a necessary benchmark model to compare against alternatives below -Simple to use	-Simple to develop and apply -provides unbiased estimates -well known statistical properties	-can predict outside the valid range -assumes linearity and too simplistic
Beta	4	-Responses may be non-linear with EQ-5D -Theoretical basis suggests more flexible to model skewness	-Better statistical properties over the linear models if data skewed and is multi-modal -can restrict predictions to lie within the valid range -can model the variance -treatment effects in terms of odds ratios can contextualize mean differences -can model the clustering (.e mixed effects)	-more complex than the linear -treatment effects in terms of odds ratios may have limited application in health economic decision making -requires a transformation because only valid on a 0 to 1 scale
Quadratic	4	-Responses may be non-linear with EQ-5D -Shown to be the better performing at time of writing thesis and a comparator	-can model curvature -can model the clustering (.e mixed effects)	-difficult to interpret -predictions outside the valid range
TOBIT	4	-values above 1 and <-0.549 or 0 restricted or censored -ensure predictions are restricted within the valid EQ-5D range	-predictions within a valid range -unbiased coefficient estimators -can model the clustering (.e mixed effects), but more complex	-Restricts predictions to within a valid range -depends on normality assumption -difficult to interpret
Quantile	4	Theoretical basis suggests more flexible to model skewness -complexity of interactions between variables could suggest stronger relationship between EQ-5D and other factors with a quantile than with mean	-Predicts only quantiles -models skewness well -estimates more robust to outliers -if data normally distributed, the 50th percentile is the same as the mean	-can predict outside the valid range - if data not normally distributed the mean cannot be predicted -complex to model the clustering (i.e. mixed effects)
CLAD	4	-Previously reported in literature on mapping and shown to have reasonable predictive properties	-no assumption of normality is needed -robust to heteroscedasticity (i.e. variance increasing/decreasing with mean)	- better estimates for larger sample size (asymptotically better) -complex to model the clustering
LDVMM	5	-Previously reported in literature and	-allows modelling of two or more distributions in a	-Very complex difficult to apply

		shown to have good prediction properties and hence a useful benchmark	dataset -valid statistical properties -mixing distribution probability estimates can be uncertain	-practically require too many assumptions to be useful -possibly applicable to no more than a mixture of two distributions
Joint Model	6	Modelling the joint relationship between HRQoL and toxicity seems a natural thing to do	-Better estimates expected -models the correlation between toxicity and HRQoL -Modelling two outcomes	-Complicated to use as a mapping algorithm -still at a relatively early stage of development -more assumptions needed due to complexity
Bayesian Network	7	-Models the relationship between all different aspects of HRQoL and not just between	-Estimates the profile and not just utility, hence more informative	-Depends on prior estimates - complex to model -may only be useful for shorter HRQoL questionnaire

Table 3.1: Summary of models used.

mean, because on average, even if differences are variable, there is a tendency of the mean predicted values to cluster around the mean. Hence, the proportion of differences within ±5, ±10%, ±15%, ±25% and ±30% were computed. Moreover, graphical displays for comparing observed versus predicted, including by health state were generated. Model fit was also assessed by normal probability plots. Finally, where data were available, the impact of the predicted utility on the QALY was also determined.

3.7.3 Model Testing

For each model, the approach to testing was primarily using an independent data set where available. For example, for the TOPICAL and SOCCAR data, the models were tested independently by developing the model from one data set and testing on the other and vice versa. For study 3, cross validation methods were used by a random selection of 50% of the data to build the model and test it on the remaining 50%. In addition to this, extensive simulation was used to test the models. This included bootstrap and monte-carlo simulation techniques. For example, if the sample size of Study 3 was 100, then each bootstrap sample would be a size of 100).The monte-carlo simulation method included multivariate simulation using Fleishman methods [167,168].

3.7.4 Statistical Inference and uncertainty

In general, statistical hypotheses were rejected when the two sided p-value was <0.05. The 95% confidence intervals provided a set a plausible range of value for where the true value (whether utility or QALY) lies. Using the simulations, the proportion of (simulated) 95% confidence intervals that contained the true (observed) mean utility were used as a basis for measuring uncertainty.

3.8 Deriving QALY’s

Where QALYs were determined, these were derived as the area under the baseline-adjusted utility curve using linear interpolation between baseline, and each subsequent (nominal) time point. Specifically for the case of SOCCAR and TOPICAL (Chapter 4), for each model, patient level QALYs were generated using model estimates of patient level utilities and multiplying them by the observed survival times. Simulation was used to estimate the mean overall survival (OS), progression-free survival (PFS) and post-progression survival. The exponential model was chosen to fit the empirical Kaplan-Meier curve for OS and PFS. Using the relationship: OS= 1*Log(1-x_i))/λ where x_i are randomly generated from a uniform distribution and λ is the observed hazard rate. For each realization, the mean (area under the survival curve) OS and PFS were determined for each treatment group. For TOPICAL, there were no censored data and for SOCCAR, the censoring distribution was taken into account (because patients were still alive) so that the simulations resulted in PFS < OS. The pre and post progression utilities were determined from each of the previously described simulations. Hence, a total of 10,000 mean OS, PFS, pre-progression and post-progression utilities were generated to determine QALYs. They were estimated as weighted sums of pre-progression and post-progression mean EQ-5D-3L for each treatment group. Hence, the mean QALY was constructed as:

Mean PFS utility Mean PFS + Mean Post-Progression utility*PPS.*

3.9 Handling Missing Data

Multiple imputation (MI) using the method of chained equations (MCMC) was used for handling missing utility data, taking into account covariates including baseline utilities, age, gender and ECOG. Mean matching using predictive methods was used to improve estimates of imputed values since normality could not be assumed. Each imputed data set was analysed independently using model-based approaches; estimates were pooled to generate mean and variance estimates of utilities using Rubin’s rule to capture within and between variances for imputed samples [169,170].

Information loss from finite imputation sampling was minimized using 20 datasets, resulting in minimal loss of efficiency (<0.5%) [170]. Since the fraction of information missing was reasonably low, n=20 imputation sets were considered adequate. Imputed and observed values were compared to establish that imputation did not introduce bias into subsequent estimation [171].

Chapter 4

Chapter 4: A New Non-Linear Two Part Beta-Binomial Mapping Model

Published: A non-linear Beta-Binomial Model (Khan. I and Morris, 2014, Health and Quality of Life Outcomes, 2014, 12:163)

Published: Book Chapter: Design & Analysis of Clinical Trials for Economic Evaluation and Reimbursement; Iftekhar Khan (2016) – Chapter 5; Chapman & Hall (2016, 312 pages)

Abstract

Introduction: The performance of the Beta-Binomial (BB) model is compared to several existing models for mapping EORTC QLQ-C30 (QLQ-C30) on to EQ-5D-3L, using data from lung cancer trials.

Methods: Data from two separate non-small cell lung cancer clinical trials (TOPICAL and SOCCAR) are used to develop and validate the BB model. Comparisons with Linear, TOBIT, Quantile, Quadratic and CLAD models are executed. The mean prediction error, R^2, proportion predicted outside the valid range, clinical interpretation of coefficients, model fit and estimation of Quality Adjusted Life Years (QALY) are reported and compared. In addition, Monte-Carlo simulation is used.

Results: The Beta-Binomial regression model performed 'best' among all the tested models. The AIC for the BB was lowest (AIC=-2215) compared to all other models for both TOPICAL and SOCCAR data sets. For TOPICAL and SOCCAR trials, respectively, residual mean square error (RMSE) was 0.09 and 0.11; R^2 was 0.75 and 0.71; observed vs. predicted means were 0.612 vs. 0.608 and 0.750 vs. 0.749. Models tested on independent data indicate 95% confidence from the BB model contain the observed mean (77% and 59% of the time for TOPICAL and SOCCAR, respectively) compared to the other models. All algorithms over-predict at poorer health states but the BB model was relatively better, specifically for the SOCCAR data.

Conclusion: The BB model may offer superior predictive properties amongst the considered mapping algorithms and may be more useful when predicting EQ-5D-3L at poorer health states. The algorithm derived from the TOPICAL data, due to better predictive properties and smaller uncertainty is, hence, recommended.

4.1 Introduction

Estimating patient-level utilities to determine Quality Adjusted Life Years (QALYs), which otherwise might be unavailable, is a key objective of mapping. EQ-5D-3L is recommended by the National Institute for Clinical Excellence (NICE) in the UK for use in economic evaluations, specifically, Cost Utility Analysis (CUA) [83].

Several models have been developed and published for mapping QLQ-C30 to predict EQ-5D-3 [111,130-132,143] as indicated in Chapters 1 & 2. These models have been compared using measures like predictive power (R^2), predictive mean and residual mean squared error (RMSE). Models used for mapping include conditional mean or median regression models, with varying degrees of success [112]. Of 119 models examined, ordinary least squares (OLS) approaches were the most common [112].

Longworth and Rowen (2013) suggest several alternative models including the Beta-Binomial for an investigation [115]. Crott et al. (2012) also suggest research on more complex mapping algorithms, as well as a need for greater validation [135]. One aspect of this 'complexity' may involve adding several interactions or additional covariates, which improve prediction. However, this can result in an overly complicated algorithm (e.g. having many factors that make interpretation of the model difficult). Another perspective on 'complexity' might be a more complicated mathematical function (e.g. Joint or non-linear models), but with fewer variables.

The advantage of the BB as a useful mapping algorithm is its flexibility and ability to model skewed and multimodal data, measured on a zero to one interval, directly or through a transformation (because EQ-5D is measured on a scale of -0.549 to 1). The modelling context allows the clustering of data (to model correlations within and between subjects) and is shown to have reported more precise and efficient parameter estimates [172]. In situations where responses are overinflated at extremes (ceiling effects), it is particularly useful as one can attempt to model extreme values, rather than omitting them or considering them as outliers. In addition, effect sizes in terms of odds ratios may be more meaningful to decision-makers (particularly clinicians) than absolute mean differences.

4.2 Methods

Several mapping models applied to QLQ-C30 were identified for the purpose of this study and a useful review is published [115]. Out of the five published algorithms, which mapped QLQ-C30, four used linear models (OLS estimates) [139-144] and one used a quadratic model [111]. A NSCLC data set has been used only once previously [130].

4.2.1 Instruments

The EQ-5D-3L and QLQ-C30 were used and described earlier (Chapters 1& 2) in detail.

4.2.2 Data

Data were from two national (UK) NSCLC clinical trials described in Chapter 3. The first trial (TOPICAL) was a randomized phase III trial in 670 lung cancer patients [165]. The second was the SOCCAR trial [166].

4.2.3 Developing and Testing Alternative Models

Separate mapping algorithms were developed using data from each of the TOPICAL and SOCCAR trials using BB regression and five other models (Linear, TOBIT, Quantile, Censored Least Absolute Deviation (CLAD), and Quadratic regression) for comparison. The five models selected are among the common mapping models reported in a mapping literature review (Brazier et al., 2009) [112]. Estimated utilities from each model were compared using several criteria including RMSE, predicted distributions, MAE, confidence intervals, R^2, residual plots, the proportion of predicted EQ-5D-3L outside the range -0.549 to 1.0, estimated QALYs and Monte-Carlo simulation. The performance of each model was compared using independent data from the SOCCAR trial. In addition, each model was fitted using data from the SOCCAR trial and then tested with data from the TOPICAL trial.

4.2.4 Model Specification and Analysis Methods

For each model, data were combined across time points and treatment groups, following methods of previously reported mapping algorithms [130-132,139-144]. One reason advocated for pooling across all time points is because more health states can be modelled. The models compared were:

(I) Linear Mixed Effect Model

(II) TOBIT Mixed Effect Model

(III) Quadratic Mixed Effects Model following Crott and Briggs (2010)

(IV) Quantile Fixed Effects Model

(V) Censored Least Absolute Deviation (CLAD): Fixed Effects Model

(VI) Mixed Effects Beta-Binomial Regression Model

The linear mixed effects model is a regression model with subject as a random term. The linear mixed and quadratic models model the mean of the EQ-5D-3L utility distribution.

The Quadratic Model of Briggs (2010) (Model III)

Briggs & Crott (2010) present a mapping of EQ-5D-3L from the QLQ-C30 using a quadratic model from a breast cancer population (N=448). Data from this study were collected at multiple time points (baseline, 0, 1 2 ,3,6,12,20, 28,36,42,48 and 54 months). The predictive power of this model using R^2 was reported at around 80%. The model was linear in parameters (i.e. a linear model but included a number of squared terms: PF, EF, SF, SL, and DI). The authors had previously suggested that linear models were inadequate and the over-prediction was an important issue. Hence a non-linear approach was used. The other important feature of this study was the validation of the model using independent data.

The TOBIT models the mean 'plus' the remainder of the distribution in a mixed effects context; however, the slope parameter is adjusted by the probability of censoring. The censoring in the TOBIT can be either from below (censoring EQ-5D-3L to 0) or above (censoring EQ-5D-3L to 1). The BB models the distributions of EQ-5D-3L responses. Models (I) - (III) are not described in detail because a review of the common features of these and other models have been discussed elsewhere [112].

Quantile Regression (Model IV)

In linear regression estimation, problems can exist when a response variable like EQ-5D-3L is skewed, truncated or discrete [172]. A general form of a regression model for estimating a quantile can also be used which takes the form:

$$Y_i = \mathbf{X}^{\tau} * \boldsymbol{\beta}_{\tau} + \varepsilon_i \qquad \textbf{[4.1]}$$

Where **X**, are the predictor variables, β the corresponding vector of coefficients representing changes in response to a specified *quantile,* subscripted τ such that that for instance when $\tau=0.5$, the quantile regression model predicts the median. The mean of the predicted medians can be used for CUA. The quantile regression model is defined as:

In order to predict median patient level EQ-5D-3L values from three variables (such as PF, RF, and EF), each patient's predicted median would be determined in a linear regression model of the form:

$$Y_i = \beta_1 * PF + \beta_2 * RF + \beta_3 * EF \qquad \textbf{[4.2]}$$

Where Y_i is the predicted median (i.e. $\tau = 0.5$) for particular values of PF, RF, and EF. In order to predict the upper quartile (75th percentile), τ would be 0.75 and the values of β_1, β_2 and β_3 would represent the coefficients related to the 75th percentile. In the context of mapping, **X** in the above equation is a matrix of the 15 QLQ-C30 domains, β_τ are a vector of coefficients and the ε_i are assumed to have a median of 0 for the τ^{th} quantile (e.g. if $\tau = 0.5$, the median of the residuals would be equal to 0). The estimates of $\boldsymbol{\beta}_\tau$ are obtained by minimizing the absolute deviation for the τ^{th} quantile or least absolute deviation (LAD) .

Censored Least Absolute Deviation (CLAD) (Model V)

CLAD extends quantile regression with censored responses [173,174,175]. Values >1 are censored at 1 and <-0.549 are censored at 0. The model is now described.

If EQ-5D-3L is the dependent variable, for each patient, the predicted median is computed from the explanatory (QLQ-C30) scores. The approach is similar to quantile regression with the exception that predicted EQ-5D-3L values >1 are restricted to 1, and similarly, values predicted to be < 0 are set to zero. The rationale for censoring assumes that the values below zero are unobservable which is likely to be incorrect and perhaps unnecessary. This is likely to lead to biased estimates (higher than expected as a result of setting known utilities where patients experience states worse than death to values of zero. The form is as follows:

$$\text{EQ-5D-3L} = \begin{cases} 0 & -0.549 < X\beta < 0 \\ 1.0 & \text{if } X\beta > 1.0 \\ X\beta, & \text{elsewhere} \end{cases} \qquad \textbf{[4.3]}$$

If the errors are symmetric (i.e. median of residuals = 0), the estimator is unbiased and consistent, though not efficient [173-175]. Also, conditional medians are estimated at the patient level.

The mean of all the individual (conditional) medians can be used as before for deriving QALYs since the mean is the statistic of choice for decisions relating to health technology assessment. The population mean and median are approximately equal for normally distributed data.

(VI) Beta-Binomial Regression

The BB distribution has been used in Probabilistic Sensitivity Analysis (PSA) in health economic modelling for utility measures such as EQ-5D-3L [176]. One reason for the use in PSA appears to be the convenience of assuming a scale from 0 to 1 for utility (although there is no EQ-5D-3L tariff, which is exactly equal to zero). Essentially, the BB regression can model responses that are unimodal or bimodal with varying levels of skewness [177-179]; utilities are often reported as having skewed or truncated distributions [139]. In addition, the BB estimates the mean of the distribution, whereas, some other models estimate the median. Therefore, the BB approach may be a more suitable model to test for developing a mapping algorithm.

An important feature of the BB approach is that mean predicted estimates of EQ-5D-3L can be estimated while restricting the range between 0 and 1. Although responses are required to be in the (0, 1) interval, BB can still be used in any interval (a, b) for a<b, using the transformation Y-a/b-a. For instance, if the observed EQ-5D-3L value is -0.1, then -0.1 – (-0.549) /1- (-0.549) would give a transformed value of 0.29. However, it may be difficult to correct both asymmetry and heteroscedasticity, resulting in difficult interpretations of parameter estimates in terms of the original response [177-179].

Using BB can be more complicated than linear models, depending upon the need to model the variance. Without modelling the variance (over-dispersion), modelling mean response (EQ-5D-3L) as a function of the 15 QLQ-C30 variables requires using a simple logit function (because values are assumed to lie between 0 and 1), as in a logistic regression model. Hence, the BB regression form for modelling EQ-5D-3L is:

$$\mu_i = \exp(\mathbf{X}\beta)/\{1+\exp(\mathbf{X}\beta)\} \quad \textbf{[4.4]}$$

Where μ_i is the expected (mean) utility for the ith patient, **X** is the set of independent QLQ-C30 variables and β are the corresponding coefficients. Equation **[4.4]**, which is non-linear in its parameters is used to estimate the patient-level EQ-5D-3L utilities.

A linearized form of equation **[4.4]** is

$$g(\mu) = \text{Log}(\mu_i / 1 - \mu_i) = \alpha + \mathbf{X}\boldsymbol{\beta}, \qquad \textbf{[4.5]}$$

where $g(\mu)$ is an equation in **[4.5]** that predicts the mean utility (μ_i) through a transformation of the utilities on a log scale. Equation **[4.4]** is used to estimate the patient-level EQ-5D-3L utilities from the logistic function, with **X** as the set of independent QLQ-C30 variables and **β** as the set of coefficients (parameters) vector:

An additional useful property of the BB model is that $g(\mu)$ is interpreted in a similar way to a log odds ratio (the main difference being that there is no dichotomization of outcomes) for any statistical inference. The mean predicted utility is however determined through equation **[4.4]**.

A further elucidation of this point is that the parameters of a Binomial distribution are (n,p), where n are the number of observations and p is the mean (mean proportion). The value of p is on a (0,1) interval and this can be assumed to follow a Beta(a, b) distribution, where a and b are the shape and scale parameters shown earlier in Figure 4.1.

If we re-write replacing p with a mean utility μ_i, then we assume that the utilities are Beta (a,b). The log transformation, through $g(\mu)$ in **[4.4]** and **[4.5]** offers a way to:

(i) estimate the mean utility

(ii) provide valid inferences for each coefficient b for which the interpretation is similar to that of an odds ratio.

The values of μ_i need to lie in the interval (0,1) and do not need to be probabilities (e.g. they could be very small counts or percentages). This is a feature of the BB which has an advantage over the common logistic regression model that assumes dichotomized responses to estimate the mean (proportion or probability), whereas, the values of μ_i are continuous.

A second model (such as a log function) could be used to model the dispersion in terms of a set of QLQ-C30 variables (not necessarily all 15 variables). The additional precision parameter ϕ, assumes a log function $\ln[g(\phi)]$, i.e. $g(\phi) = \exp(\mathbf{W}\delta)$, in terms a set of predictor variables **W,** with the corresponding parameters δ. Note **W** may not be the same as **X,** because although predictions of EQ-5D-3L might be determined from all 15 QLQ-C30 variables (**X**), the variance may depend on a subset of **X**.

Therefore, two sets of equations are associated with the QLQ-C30 variables; one through the mean EQ-5D-3L and one through the variance. These (two) equations provide the basis to determine the estimates of the parameters β.

Notation for Beta-Binomial

A response variable (EQ-5D-3L) is assumed to follow a Beta (α,β) defined by:

$$F(y|\alpha,\beta) = \{\Gamma\ (\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)\ \} * y^{\ \alpha-1} * (1-y)^{\beta-1} \quad \textbf{[4.6]}$$

The mean and variance of a Y of **[4.6]** are:

$$\alpha/(\alpha+\beta) \text{ and } \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1),$$

where α and β are the shape and scale parameters, respectively. The parameters α and β can be estimated from the observed mean and variance using the method of moments. For instance, the values of β from the TOPICAL and SOCCAR data are shown to be < 1, using the relationship:

$$\alpha = \mu*[((\mu *(1 -\mu)) / \sigma^2)- 1] \qquad \textbf{[4.7]}$$

$$\beta = (1\ -\ \mu)\ *\ [((\mu\ *\ (1\ -\ \mu))\ / \sigma^2\text{-}1] \quad \textbf{[4.8]}$$

When μ and σ^2 are unknown, the sample estimates can be used. However, the above description of the BB distribution is not useful for regression modelling and requires re-parameterization (Ferrari and Cribari-Neto, 2004), so that a response can be defined along with a set of predictors to form a regression model [180]. Setting $\mu = \alpha/\alpha+\beta$ and $\phi = \alpha+\beta$, then **[4.6]** becomes:

$$f(y|\mu,\phi) = \{\Gamma\ (\phi)/\Gamma(\mu\phi)\Gamma((1-\mu)\phi)\ \} * y^{\mu\phi-1} * (1-y)^{(1-\mu)\phi-1} \quad \textbf{[4.7]}$$

The expression in **[4.7]** is a beta distribution: y ~ Beta (μ, ϕ) and the mean, μ is expressed as a link function (to model the mean) in terms of some predictor variables. Typically, the link function $g(\mu) = \mathbf{X\beta}$ is such that $g(\mu) = \log(\mu /1- \mu)$. With this logit link function, the mean response is: $\mu = \varepsilon^{\omega}/1+ e^{\omega}$, where $\omega = \alpha + \mathbf{X\beta}$. The value of μ is restricted to a 0 to 1 scale.

The second parameter ϕ, is a precision parameter. The conditional variance, V(y) can be written as V(y) $= \mu(1-\mu)/1+\phi$, a form more flexible than the binomial $(\mu/(1-\mu))$, which allows greater flexibility to model the over-dispersion; the larger the value of ϕ, the smaller the variance (and higher precision) associated with the response. The dispersion parameter ϕ, can also be expressed as a function g $(\phi(\mathbf{X}))$. For instance, the over-dispersion may be related to one or more of the predictor variables (QLQ-C30).

The responses **Y** (i.e. EQ-5D-3L) are hence, Beta $([(h(\mu), h(\phi)]$, with likelihood function:

$$L(\beta,\delta, \mathbf{Y}, \mathbf{X}, \mathbf{W}) = \Gamma(\exp(\mathbf{W}\delta))/\Gamma(\sigma)\Gamma(\tau)\ \mathbf{Y}^{s-1}\ (1-\mathbf{Y})^{t-1} \quad \textbf{[4.8]}$$

where, $s = \exp(\mathbf{X}\beta+\mathbf{W}\delta)/\{1+\exp(\mathbf{X}\beta)\}$

$t = \exp(\mathbf{W}\delta)/\{1+\exp(\mathbf{X}\beta)\}$

On comparing **[4.7]** with **[4.8]**:

$$\Phi(\psi|\alpha,\beta) = \{\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)\} * y^{\alpha-1} * (1-y)^{\beta-1} \qquad \textbf{[4.9]}$$

where $\alpha-1 = \sigma-1$ and $\beta-1=\tau-1$, the complicated expression relates the observed EQ-5D-3L (y values) through the mean and variance.

This form of parameterization allows a powerful way of modelling utilities not only for mapping but in a generalized mixed modelling context for estimates of utilities and HRQoL which can be scaled to a 0,1 interval. The approach in this chapter is based on a nonlinear mixed modelling approach, where the general likelihood has been programmed directly using the SAS software (version 9.3) [181]. One reason for transforming to a (0,1) range is that the flexibility of the BB model can be properly utilized. For outcomes with bimodal or U shaped distributions, it is important that unbiased estimates can be determined. For example, in Figures 4.1, the shape and scale parameters are useful for modelling distributions which are possible (where overs dispersion of values around 1 or zero are observed). It is necessary that the transformation is made since a key assumption of the 'Binomial' part in the 'Beta-Binomial' framework is that the parameter p lies between the values of 0 and 1.

Moreover, in many probabilistic sensitivity analyses, the assumptions for utility have been suggested as a Beta-Binomial [176].

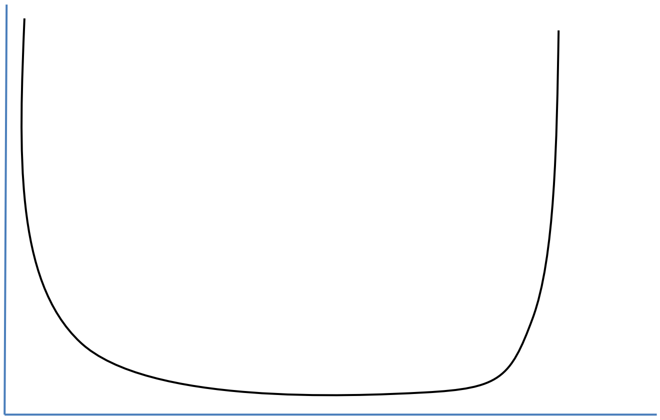


Figure 4.1: Parameters from a BB a < and b <1 resulting in bimodality and U shaped

One reasonable assumption imposed here is that the observed values <0 are set equal to 0. There were < 0.5% of values with EQ-5D-3L responses <0 in each data set; hence, the potential for bias was considered to be minimal. A transformation was, therefore, discarded. However, it is not necessary to set values <0 to zero and this could yield misleading estimates.

Although the BB regression can be very flexible, there exist limitations to the model. The drawback is that the observations existing at either 0 or 1 must be scaled away from these values. That is when there is an inflation of 0 or 1 responses, estimation with a standard BB regression can become problematic. Therefore, a *zero-one inflated* BB model was used to account for potentially over-dispersed 0 and 1 responses [179,181,182].

4.2.5 Testing Algorithms with Lung Cancer Data

For each of the models, using all 15 predictors (one for each of the QLQ-C30, regardless of their statistical significance), the predicted and observed values were compared. Models were developed using the larger TOPICAL data set and validated with SOCCAR data. The proportion of individual predicted EQ-5D-3L responses from each model that lie within ±5% to ±30% of the observed EQ-5D-3L were presented. Estimated utilities from each model were compared using the previously described model statistics.

4.2.6 Model Checking and Adequacy

In all the models, adequacy of fit was considered using residuals, tests for homoscedasticity and Aikakes Information Criterion (AIC). The AIC was used to compare models for the *same* dataset.

4.2.7 Simulations

Monte-Carlo simulations (10,000) from a multivariate distribution for the EQ-5D-3L and QLQ-C30 scores using the method of Fleishman (1979) [167, 168] with the observed correlation structure (Tables 4.7 and 4.8) were carried out to assess the uncertainty of predicted means from each model. The method of Fleishman uses higher order moments (Skewness and kurtosis) as a way of simulating data that approximates the sampling distribution. Each data set simulated contained 670 and 130 patients with 2038 and 1002 observations for TOPICAL and SOCCAR, respectively.

4.2.8 Addressing over Prediction of EQ-5D-3L at the ‘Poorer’ Health States

The over-prediction at ‘poorer’ health states was investigated using a health state of 11321 (EQ-5D-3L utility 0.433) as a cut-off for ‘Poor’ and ‘Good’ health states in TOPICAL and 22222 (EQ-5D-3L utility 0.516) in SOCCAR. Over prediction was defined such that the difference between the estimated patient level utility from the model was greater than the observed utility by any amount (difference >0). The same definition was given for ‘under prediction’ in the opposite direction (difference <0). The selected health states cut-points were chosen as this is where the observed and predicted EQ-5D-3L values start to diverge.

4.2.9 Impact on QALY Estimates

For each model, patient level QALYs were generated using model estimates of patient level utilities and multiplying them by the observed survival times. For PSA, the simulation was used to estimate the mean overall survival (OS), progression-free survival (PFS) and post-progression survival. OS and PFS are important outcomes in cancer trials, often calculated from the time of treatment allocation or randomization until death (OS) or disease progression (PFS). The exponential model was chosen to fit the empirical Kaplan-Meier curve for OS and PFS. Using the relationship: OS= $1*\text{Log}(1-x_i))/\lambda$, where x_i are randomly generated from a uniform distribution and λ is the observed hazard rate. In economic evaluation, in the context of cancer, a common approach is to simulate survival times, which approximates the mean OS and PFS (not the median).

For each realization, the mean (area under the survival curve) OS and PFS were determined for each treatment group. For TOPICAL, there were no censored data and for SOCCAR, the censoring distribution was taken into account (because patients were still alive) so that the simulations resulted in PFS < OS. The pre and post progression utilities were determined from each of the previously described simulations. Hence, a total of 10,000 mean OS, PFS, pre-progression, and post-progression utilities were generated to determine QALYs. They were estimated as weighted sums of pre-progression and post-progression mean EQ-5D-3L for each treatment group.

The mean QALY was constructed as:

Mean PFS utility Mean PFS + Mean Post-Progression utility*PPS.*

4.3 Results

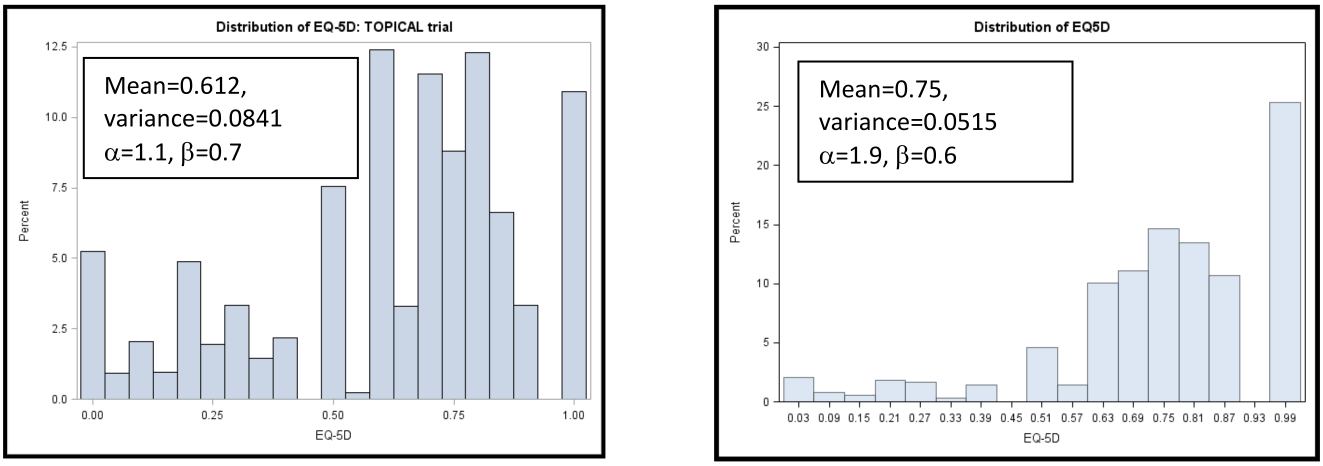
A total of 2038 and 1002 data points with 84 and 54 health states were observed for each of TOPICAL and SOCCAR trials, respectively. The average (median) number of observations per health state were 3 for TOPICAL and 2 for SOCCAR. The most frequent health state in TOPICAL was 21222 (12%) and for SOCCAR was 11111 (25%), followed by 21222 (8%). Patients in the SOCCAR trial had a better performance status (Table 4.1) compared to the TOPICAL patients. Less than 0.5% (3/2038 observations in TOPICAL and 1/1002 in SOCCAR) of EQ-5D-3L observations had values <0 (corresponding to 3 health states in TOPICAL and 2 health states in SOCCAR).

	TOPICAL (N=670)	**SOCCAR (N=130)**
EQ-5D number of observations	2038	1002
Health states (range)	84 (11111, 33312)	54 (11111, 23223)
[EQ-5D Value] {number of HS <0}	[1, -0.043] {3 HS< 0}	[1, -0.028] {2 HS<0}
Median number of observations per HS	3	2
Most Frequent HS	21222 (12%) [value=0.62]	11111 (25%) [value = 1]
ECOG	1-3	0-1
Median Age (years)	77	62
Disease Stage	IIIb-IV	IIIa-IIIb

Table 4.1: Summary of Health states and Baseline Characteristics

HS: Health States

The observed mean (SD) EQ-5D-3L for TOPICAL and SOCCAR were 0.61 (0.29) and 0.75 (0.23), respectively over all post-baseline time points (Table 4.2). Figure 4.2 illustrates the distributions of the EQ-5D-3L, confirming the presence of non-normality, skewness, and multimodality. The parameters α and β are shape parameters which are used to model the distribution; if α and β are the same (i.e. α = β), the distribution tends towards symmetry (if either α, β are < 1, data are considered non-normal, multimodal or skewed). The Kolgomorov-Smirnoff goodness of fit rejects normality (P=0.0093 and P<0.001 for TOPICAL and SOCCAR, respectively).



Mean=0.612,
variance=0.0841
α=1.1, β=0.7

Mean=0.75,
variance=0.0515
α=1.9, β=0.6

Figure 4.2: Distribution of EQ-5D for TOPICAL and SOCCAR

K-Smirnoff Test statistic p-value <0.0093 and <0.001 for TOPICAL and SOCCAR respectively to test normality.

	<u>TOPICAL (N=670)</u>					***<u>SOCCAR (N=130)</u>***				
	Baseline	**3 months**	**6 months**	**12 months**	**Overall**	**Baseline**	**3 months**	**6 months**	**12 months**	**Overall**
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	**Mean (SD)**	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	**Mean (SD)**
EQ-5D	0.56 (0.30)	0.61 (0.30)	0.66 (0.26)	0.58 (0.36)	**0.61 (0.29)**	0.79 (0.17)	0.71 (0.24)	0.71 (0.25)	0.71 (0.33)	**0.75 (0.23)**
PF	51.53(26.44)	51.66 (24.91)	54.00 (25.41)	54.29 (29.14)	**54.15 (26.30)**	86.19 (14.50)	72.06 (23.78)	79.11 (43.57)	98.52 (84.21)	**78.60 (36.53)**
RF	45.18 (36.58)	45.70 (34.26)	51.36 (32.79)	51.41 (33.94)	**49.18 (34.67)**	80.84 (24.40)	63.11 (33.28)	72.67 (49.22)	90.37 (89.43)	**71.28 (43.04)**
EF	70.96 (24.13)	73.17 (24.02)	75.69 (24.49)	74.58 (27.18)	**73.89 (24.41)**	73.41 (23.68)	71.91 (28.30)	80.33 (46.39)	91.30 (91.60)	**77.06 (42.23)**
FI	75.47 (25.75)	77.24 (25.10)	78.15 (24.27)	68.36 (27.28)	**77.17 (24.41)**	85.85 (26.84)	78.28 (24.15)	88.22 (42.34)	103.33 (83.08)	**83.13 (42.32)**
SF	59.47 (36.00)	64.71 (33.76)	68.69 (31.76)	64.94 (31.79)	**66.90 (32.40)**	84.52 (25.58)	64.79 (36.79)	76.67 (47.77)	92.59 (91.05)	**74.55 (45.17)**
QL	47.60 (25.27)	51.64 (23.03)	55.29 (22.84)	52.68 (23.49)	**52.27 (23.25)**	71.36 (18.48)	56.93 (22.93)	59.78 (33.09)	55.56 (59.96)	**63.75 (29.65)**
FA	53.72 (28.59)	51.45 (29.59)	47.41 (28.67)	45.20 (28.03)	**48.52 (28.89)**	26.42 (25.47)	46.44 (30.36)	35.41 (50.09)	13.33 (89.88)	**33.94 (42.20)**
NV	14.14 (22.61)	11.89 (19.55)	8.56 (15.24)	9.89 (16.41)	**10.65 (18.81)**	6.82 (25.92)	17.42 (26.70)	11.78 (43.22)	12.96 (86.84)	**10.58 (37.26)**
PA	32.22 (32.97)	26.11 (29.97)	24.89 (28.54)	26.84 (32.02)	**26.00 (29.70)**	17.45 (22.41)	23.41 (26.32)	23.33 (49.32)	5.93 (91.09)	**20.66 (41.47)**
DY	55.72 (33.49)	50.40 (32.56)	49.89 (34.32)	52.54 (31.07)	**49.76 (32.92)**	29.66 (33.92)	30.71 (30.66)	30.67 (49.85)	14.81 (90.61)	**31.57 (45.98)**
SL	33.71 (33.83)	32.28 (34.47)	24.49 (29.54)	31.61 (33.87)	**29.03 (32.67)**	27.03 (31.63)	32.96 (33.52)	22.22 (48.81)	8.15 (89.66)	**24.00 (42.70)**
AP	44.07 (38.14)	42.15 (36.44)	32.66 (34.22)	32.20 (32.73)	**36.78 (35.57)**	16.54 (25.50)	30.34 (47.31)	24.44 (49.12)	3.70 (89.39)	**20.47 (43.61)**
CO	28.12 (34.81)	21.95 (29.00)	15.86 (24.87)	20.83 (26.64)	**20.33 (28.21)**	13.76 (22.42)	37.45 (47.63)	11.56 (45.02)	10.00 (92.93)	**18.64 (46.78)**
DI	19.59 (30.71)	20.57 (30.91)	16.32 (25.49)	16.67 (25.93)	**17.14 (27.54)**	3.70 (10.52)	7.12 (19.77)	2.22 (38.88)	20.00 (99.75)	**6.06 (40.62)**
CF	14.36 (25.35)	9.57 (19.95)	7.43 (18.98)	12.07 (24.74)	**10.00 (21.02)**	20.11 (31.86)	23.22 (41.26)	19.56 (50.85)	10.74 (86.89	**20.44 (49.53)**

Table 4.2: Summary Statistics of EQ-5D and QLQC30

Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF); Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI), Financial Problems (FI); Global Health Status Score (QL).

4.3.1 Comparison of Models

All terms in the model were retained (regardless of statistical significance); even if some terms were not statistically significant. This was because they can still be relevant later when applying the algorithm [183]. The standard errors of the QLQ-C30 for the BB were smallest (Table 4.3). The AIC values were smallest for the BB model and ranged from -2215 (BB) to -864 (Quantile) for TOPICAL and -1529 (BB) to -587 (Quantile). Smaller AIC values suggest better fit (Table 4.3). For the six models, the estimated R^2 ranged from 0.53 (CLAD) to 0.75 (BB); R^2 was highest (TOPICAL R^2=0.75) for the BB model. Estimated RMSE ranged from 0.09 (BB) to 0.18 (Quadratic, CLAD). The proportion of predicted values of EQ-5D-3L >1 were highest for the quantile model (3.5%), whereas for the Quadratic model, predicted values <0 were more common (5%), despite only 0.3% of observed values <0 were set to zero for modelling purposes (Table 4.3). Only the TOBIT, CLAD, and BB did not predict outside the ‘observed’ range. P-values for coefficients are shown in Table 4.4.

	Linear Mixed		*TOBIT*		*Quadratic*		*Quantile*		*CLAD*		*Beta*	
QLQ-C30	***TOPICAL***	***SOCCAR***	***TOPICAL***	***SOCCAR***	***TOPICAL***	***SOCCAR***	***TOPICAL***	***SOCCAR***	***TOPICAL***	***SOCCAR***	***TOPICAL***	***SOCCAR***
R^2	0.63	0.64	0.65	0.63	0.64	0.62	0.66	0.62	0.55	0.53	0.75	0.71
MAE	0.14	0.14	0.13	0.10	0.16	0.129	0.13	0.09	0.14	0.71	0.10	0.13
RMSE	0.183	0.141	0.17	0.14	0.18	0.14	0.17	0.14	0.18	0.15	0.09	0.11
Predicted Mean (SE)*	0.584 (0.0047)	0.771 (0.0058)	0.631 (0.0057)	0.771 (0.0068)	0.635 (0.0074)	0.774 (0.0057)	0.633 (0.0054)	0.766 (0.0062)	0.593 (0.0059)	0.782 (0.0058)	0.608 (0.0040)	0.749 (0.0049)
Predicted >1 (%)	0.11%	1.04%	0	0	0	1%	2.8%	3.5%	0	0	0	0
Predicted < 0 (%)	0	0	0	0	5%	2%	0.6%	0.4%	0	0	0	0
AIC (lower is better)	-1015	-782	-936.9	-593	-978.6	-782	-864	-587	-926	-601	-2215	-1529
SE of coefficients												
PF	0.000264	0.000335	0.000288	0.000413	0.000686	0.000766	0.002805	0.000513	0.000295	0.000555	0.000190	0.000293
RF	0.000213	0.000231	0.000231	0.000285	0.000710	0.000635	0.000322	0.000267	0.000422	0.000287	0.000106	0.000220
EF	0.000214	0.000250	0.000233	0.000307	0.000532	0.000289	0.000312	0.000333	0.000392	0.000411	0.000289	0.000216
SF	0.000195	0.000203	0.000210	0.000249	n/a	n/a	0.000316	0.000263	0.000419	0.000313	0.000149	0.000204
CF	0.000207	0.000231	0.000225	0.000288	n/a	n/a	0.000244	0.000328	0.000447	0.000348	0.000163	0.000211
FA	0.000257	0.000296	0.000279	0.000373	n/a	n/a	0.000204	0.000373	0.000305	0.000391	0.000224	0.000251
NV	0.000234	0.000217	0.000255	0.000274	n/a	n/a	0.000255	0.000310	0.000255	0.000430	0.000258	0.000204
PA	0.000168	0.000203	0.000182	0.000250	0.000165	0.000200	0.000329	0.000410	0.000329	0.000460	0.000128	0.000211
DY	0.000154	0.000161	0.000168	0.000199	n/a	n/a	0.000201	0.000209	0.000201	0.000277	0.000159	0.000116
SL	0.000144	0.000165	0.000156	0.000203	0.000379	0.000277	0.000241	0.000211	0.000295	0.000291	0.000136	0.000149
AP	0.000138	0.000170	0.000150	0.000208	n/a	n/a	0.000242	0.000221	0.000292	0.000221	0.000129	0.000169
CO	0.000153	0.000158	0.000166	0.000196	0.000151	0.000155	0.000206	0.000207	0.000311	0.000283	0.000149	0.000206
DI	0.000152	0.000218	0.000165	0.000291	0.000405	0.000393	0.000219	0.000426	0.000213	0.000396	0.000133	0.000191
QL	0.000240	0.000294	0.000263	0.000372	n/a	n/a	0.000363	0.000210	0.000299	0.000260	0.000201	0.000241
FI	0.000204	0.000143	0.000224	0.000182	n/a	n/a	0.000101	0.000100	0.000209	0.000210	0.000243	0.000182

Table 4.3: Summary of Model Fit Statistics

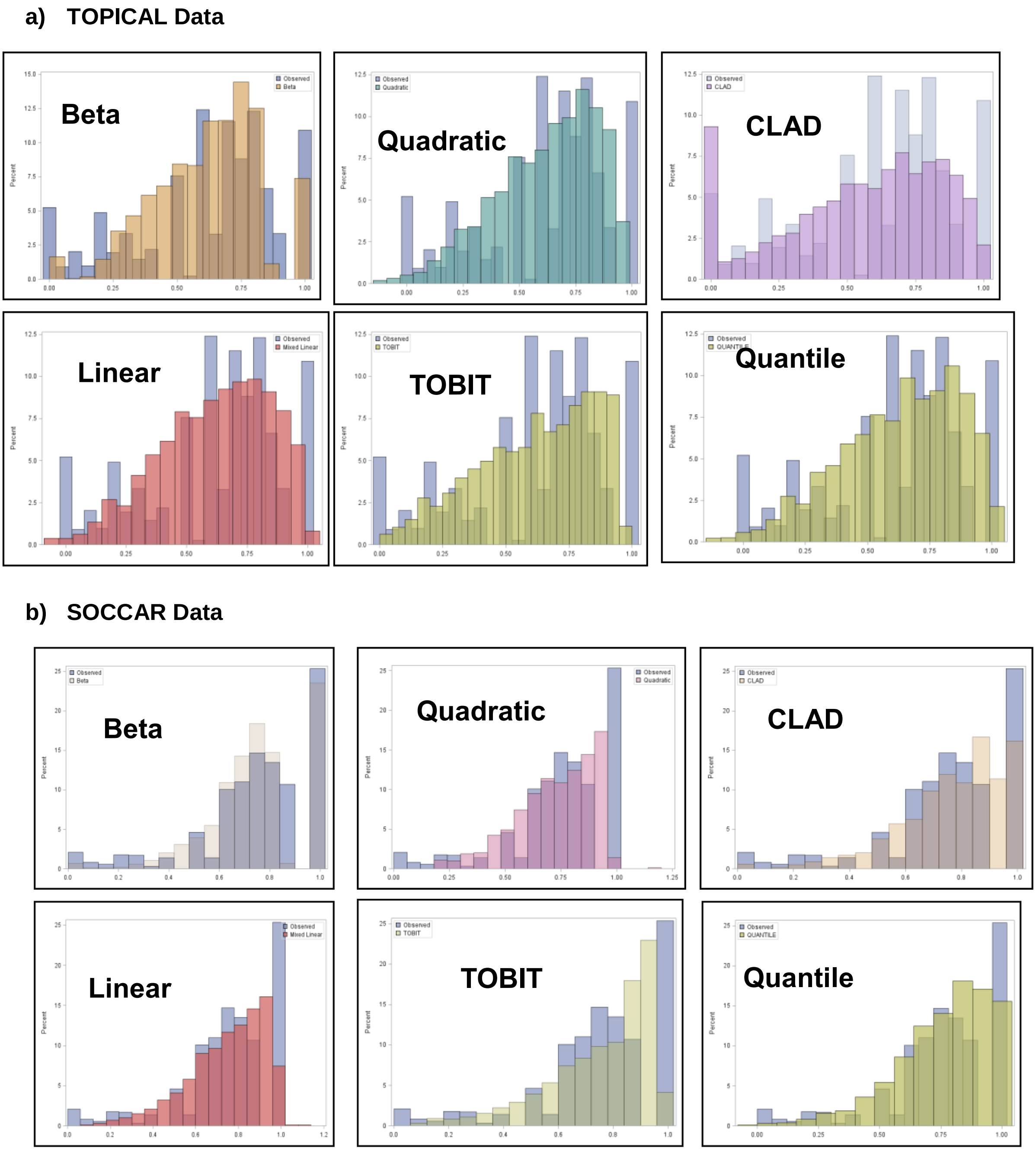
Note: n/a: not applicable because not used in original model of Crott & Briggs (2010); MAE: Mean Absolute Error; RMSE: Residual Mean Squared Error; SE: Standard Error
*Observed post baseline Mean (SD) for TOPICAL was 0.61 (0.29) and for SOCCAR was 0.75 (0.23)

QLQ-C30	*Linear Mixed*	*TOBIT*	*Quadratic#*	*Quantile*	*CLAD*	*Beta*
PF	0.0029, 0.0019 (0.0001, <0.0001)	0.0034, 0.0022 (<0.0001, <0.0001)	0.0061, 0.00016 (<0.0001,0.8247)	0.0028, 0.0022 (<0.0001, <0.0001	0.0028 , 0.0016 (<0.0001, <0.0001	0.518, 0.260 (<0.0001, <0.0001)
RF	0.0010,0.0010 (<0.0001, <0.0001)	0.0009,0.0012 (<0.0001, <0.0001)		0.0006,0.0007 (0.0150, 0.0176)	0.0008, 0.0008 (<0.0001, 0.0012)	0.108,0.2340 (0.0075, <0.0001)
EF	0.0015, 0.0021 (0.0008, <0.0001)	0.0019, 0.0025 (<0.0001, <0.0001)	0.00409, 0.00042 (<0.0001, 0.5028)	0.0019, 0.0026 (<0.0001, <0.0001	0.0017, 0.0028 (<0.0001, <0.0001)	0.067, 0.379 (<0.0001, <0.0001)
SF	0.0009,0.00010 (<0.0001, <0.0001)	0.0009,0.0011 (<0.0001, <0.0001)	0.00172,0.00152 (0.0013, <0.0001)	0.0016,0.0005 (<0.0001, 0.0039)	0.0015, 0.0007 (0.0015, 0.0086)	0.029, 0.257 (0.0452, <0.0001)
CF	-0.0003,-0.0002 (<0.0001, 0.9293)	-0.0003, -0.0007 (0.4422, 0.8041)		0.0003,0.0001 (0.2143, 0.6266)	0.0003, -0.0004 (0.0468, 0.8884)	0.067, 0.061 (<0.0001, <0.0001)
FA	0.0004,0.0001 (0.0163, 0.8389)	0.0005, -0.00021 (0.0513, 0.5763)		0.0008,-0.0004 (0.0244, 0.1797)	0.0007, -0.0005 (0.0114, 0.3422)	0.064, -0.012 (0.0122, 0.8299
NV	-0.0002,0.0004 (0.3131, 0.1281)	-0.0001,0.0005 (0.5699, 0.0715)		-0.0001,0.0003 (0.6821, 0.3951)	0.00001, 0.0004 (0.1922, 0.0998)	0.011, 0.062 (0.6613, 0.0875)
PA	0.0019, 0.0025 (<0.0001, <0.0001)	-0.0029, -0.0017 (<0.0001, <0.0001	-0.0030,-0.0018 (<0.0001, <0.0001	-0.0029, -0.0020 (<0.0001, <0.0001)	-0.0024, -0.0018 (<0.0001, 0.0019)	-0.496, -0.235 (<0.0001, <0.0001)
DY	0.0004,0.00001 (00.330, 0.9469)	0.0004,-0.0002 (0.0361, 0.3748)		0.0003,-0.0001 (0.0670, 0.4068)	0.0002, 0.0001 (0.0466, 0.5325)	0.065, 0.0088 (0.0219, 0.8316)
SL	-0.0004,-0.0004 (0.0161, 0.0044)	-0.0004,-0.0006 (0.0053, 0.0059)	-0.00085, -0.00087 (0.0249, 0.0017)	-0.0004,-0.0005 (0.0410, 0.0032)	-0.0003, -0.0004 (0.0333, 0.0024)	-0.062, -0.102 (0.0195, 0.0036)
AP	0.0001,-0.0002 (0.7427, 0.1192)	0.00007,-0.0002 (0.6337, 0.3325)		0.0001,-0.0001 (0.7616, 0.6294)	0.00001, -0.00007 (0.6257, 0.5998)	0.0265, 0.0017 (0.3421, 0.9660)
CO	-0.0003,-0.0001 (0.0602, 0.0882)	-0.0004,0.0002 (0.0101, 0.2632)	-0.00036, 0.0062 (0.0177, 0.2514)	-0.0002,0.0003 (0.3031, 0.1062)	-0.0002, 0.0004 (0.2145, 0.1145)	-0.082, 0.026 (0.0010, 0.4835)
DI	-0.0002,0.0007 (0.9213, 0.0060)	-0.0001,0.0009 (0.5608, 0.0009)	-0.00062,-0.00029 (0.1257, 0.4490)	-0.0001,0.0001 (0.8258, 0.8379)	-0.0004, 0.0009 (0.7589, 0.4417)	-0.017,0.051 (0.4614, 0.1574)
FI	0.0004,-0.0001 (0.0076, 0.0002)	-0.0007,0.00037 (0.0222, 0.0418)		0.0004,0.0001 (0.0325, 0.4257)	-0.0039, 0.0008 (0.0187, 0.0587)	-0.005, 0.062 (0.0039, 0.0715)
QL	0.0014, 0.0013 (0.0163, 0.0002)	0.0018, 0.00018 (<0.0001, <0.0001		0.0014, 0.0009 (<0.0001, <0.0001	0.0013, 0.0014 (<0.0001, <0.0001	0.237, 0.224 (<0.0001, <0.0001)

Table 4.4: Summary of Models: Coefficients (P-values)

#For Quadratic, P-values for PF^2, PF^2, PF^2, PF^2, PF^2 were (0.0002, <0.0001); (0.003, 0.002); (0.503, 0.261); (0.286, 0.034); (0.251, 0.015) respectively

The models were first tested on the same dataset used to develop the mapping algorithm (Figure 4.3a and 4.3b).



a) TOPICAL Data

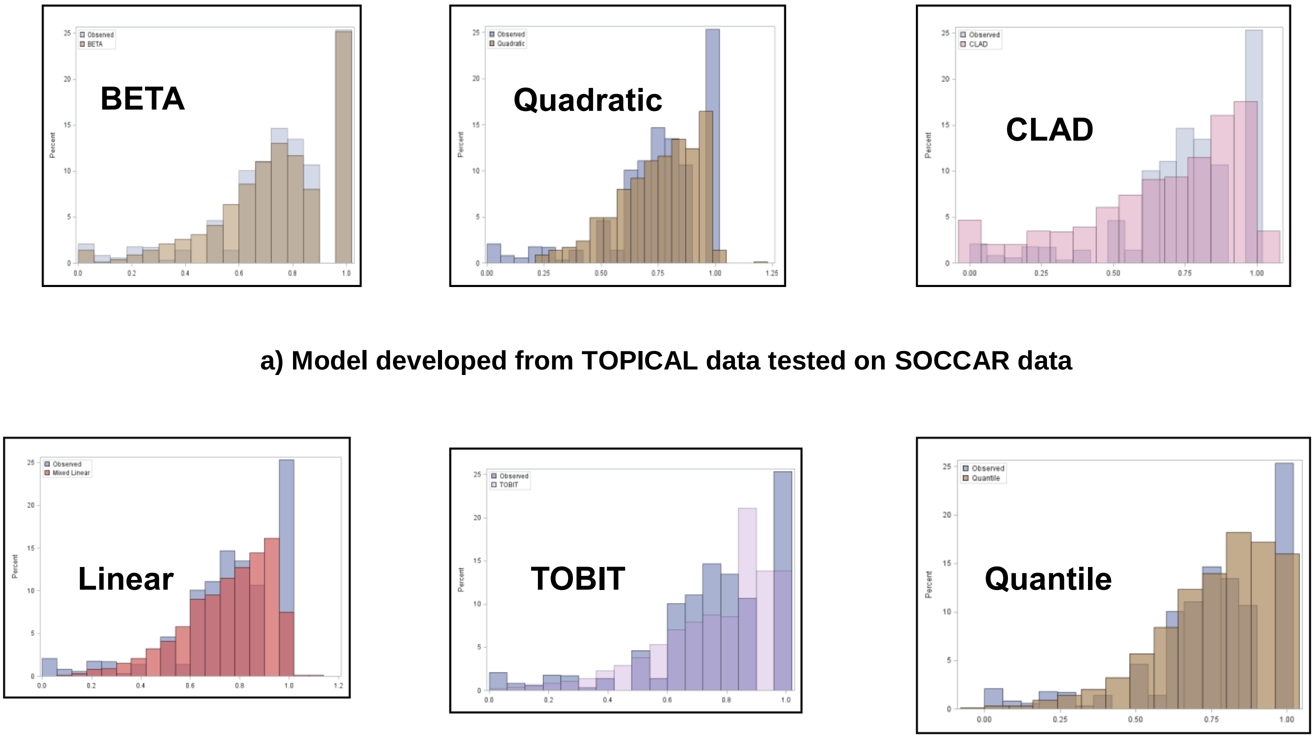
b) SOCCAR Data

Figure 4.3: Observed vs. Mean Predicted EQ-5D value ((a) TOPICAL & (b) SOCCAR)

Mean predicted EQ-5D-3L distributions are also illustrated in Figures 4.2a and 4.2b. The predicted means were 0.608 for BB, 0.584 (Linear), 0.631 (TOBIT), 0.635 (Quadratic), 0.633 (Quantile) and 0.593 (CLAD) in TOPICAL; For SOCCAR these were 0.749 (BB), 0.771 (Linear), 0.771 (TOBIT), 0.774 (Quadratic), 0.766 (Quantile) and 0.782 (CLAD). Predicted mean EQ-5D-3L was closest to the observed with the BB model (Table 4.3).

4.3.2 Testing Models using Independent Data (out of sample predictions)

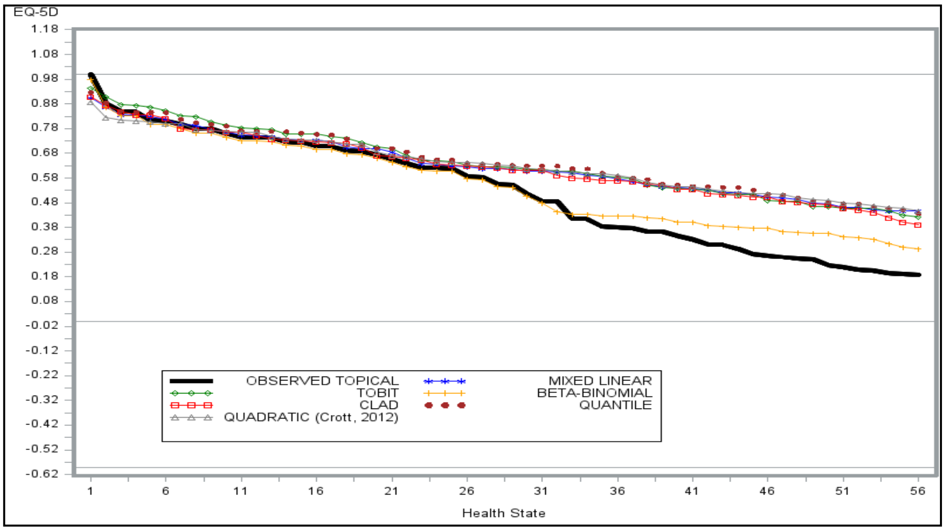
The model developed from TOPICAL was tested on SOCCAR data (Figure 4.4a, 4.4b and Table 4.5).



a) Model developed from TOPICAL data tested on SOCCAR data

b) Model developed from TOPICAL data tested on SOCCAR data: Predictions by health states

Figure 4.4: Out of sample predictions



Model	Predicted Mean (SE) [95% CI]	R^2	RMSE	% of 95% CI containing the observed mean‡
Beta (a)*	**0.747 (0.0069)** [0.733, 0.760]	0.75	0.132	77%
Beta (b)†	**0.622 (0.0057)** [0.608, 0.631]	0.61	0.159	59%
CLAD (a)*	**0.671 (0.0091)** [0.652, 0.689]	0.56	0.027	28%
CLAD (b)†	**0.652 (0.0054)** [0.639, 0.660]	0.47	0.154	19%
Linear Mixed (a)*	**0.738 (0.0051)** [0.728, 0.747]	0.63	0.019	45%
Linear Mixed (b)†	**0.642 (0.0059)** [0.630, 0.653]	0.58	0.095	23%
Quadratic (a)*	**0.768 (0.0056)** [0.757, 0.778]	0.63	0.018	37%
Quadratic (b)†	**0.636 (0.0039)** [0.628, 0.643]	0.55	0.093	14%
TOBIT (a)*	**0.739 (0.014)** [0.702, 0.757]	0.56	0.021	65%
TOBIT (b)†	**0.644 (0.0084)** [0.627, 0.660]	0.59	0.112	24%
Quantile (a)*	**0.772 (0.0060)** [0.754, 0.778]	0.62	0.190	21%
Quantile (b)†	**0.661 (0.0084)** [0.644, 0.677]	0.58	0.148	8%

Table 4.5: Testing of models using independent data

* Model developed from TOPICAL trial and tested using SOCCAR Data

† Model developed from SOCCAR trial and tested using TOPICAL Data

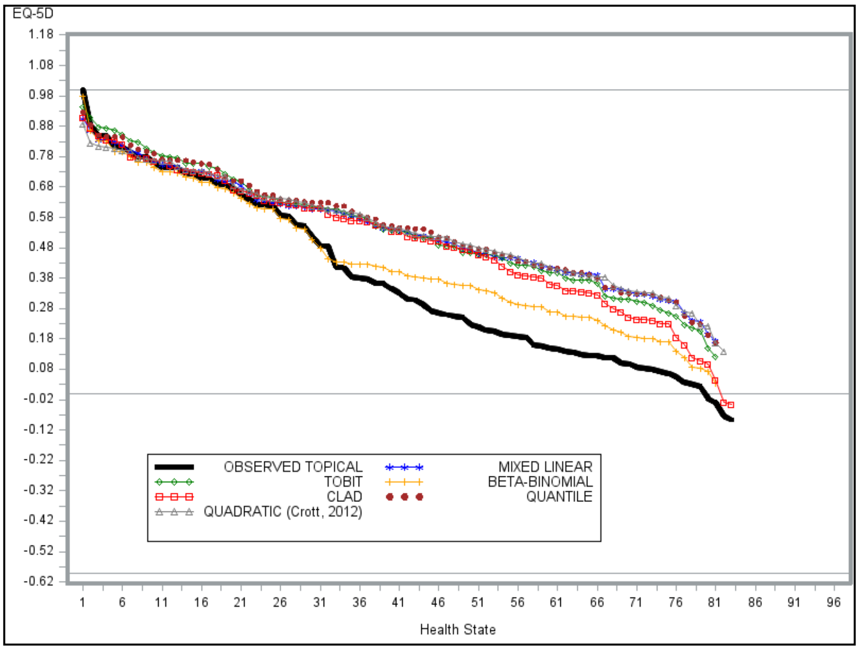
‡ based on 10,000 monte-carlo simulations

Figure 4.4 compares 'out of sample' predicted and observed EQ-5D-3L distributions. In particular, Figure 4.3a shows the predicted vs. observed distributions for the model developed from TOPICAL and tested using the SOCCAR dataset. The BB model

predicts the over-dispersion at values of zero and one better than other models. For SOCCAR, 25% of EQ-5D-3L responses were one, and the BB has predicted these very well. The CLAD and Quantile also predict these with some success, whereas TOBIT and Linear were less accurate. Figure 4.3b shows the predicted mean EQ-5D-3L by health state compared to observed values. The BB also over-predicts EQ-5D-3L at poorer health states, but the extent of the over-prediction is less severe.
The R^2 values were highest with the BB (R^2=0.75) when the model developed from TOPICAL data was tested on SOCCAR data and R^2=0.61 for the model developed from SOCCAR data and tested on TOPICAL; the RMSEs were also higher compared to other models (Table 4.5). Mean predicted EQ-5D-3L for SOCCAR was 0.747 (95% CI: 0.733, 0.760) for the BB.

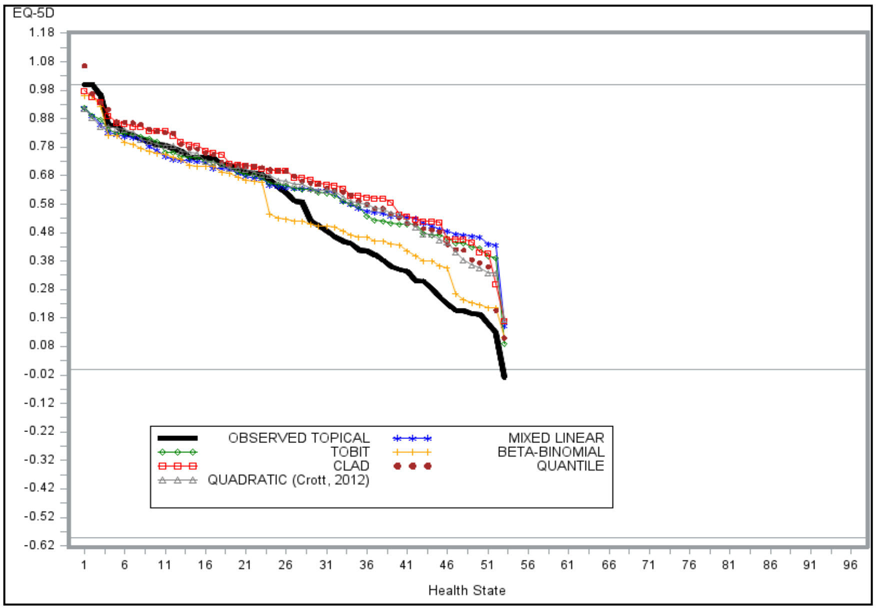
With SOCCAR data, the BB model approximates mean EQ-5D-3L at each health state more closely than all other models (Figure 4.4a). Figure 4.4b shows a similar plot using independent data from the algorithm developed from TOPICAL data.

a) Comparison of models using TOPICAL data



The x-axis indicates ordered health states (1 refers to 11111 and 84 is 33312); these are ordered according to the weighted value of the health state using the UK TTO tariff.

b) Comparison of models using SOCCAR data



The x-axis indicates ordered health states (1 refers to 11111 and 54 is 23223); these are ordered according to the weighted value of the health state using the UK TTO tariff.

Figure 4.5: Predicted EQ-5D versus observed EQ-5D for each Model by health state

developed from TOPICAL data predicted mean EQ-5D-3L to within 0.4%. From 10,000 simulations, 77% of the 95% confidence intervals for the predicted mean EQ-5D-3L contained the observed SOCCAR mean EQ-5D-3L value of 0.75.

When the model was developed using SOCCAR data and then tested on TOPICAL, the mean predicted mean EQ-5D-3L was 0.622 (as compared to the observed 0.61) and 59% of the 95% confidence intervals contained the mean EQ-5D-3L value of 0.61 observed in TOPICAL (Table 4.5). A possible reason for the lower proportion of coverage is that SOCCAR patients had less severe symptoms than NSCLC patients. Consequently, they were in ‘better’ health states compared to the patients in the TOPICAL trial. Normal probability plots of the model tested on SOCCAR data indicate a better fit with the BB model (Figure 4.6).

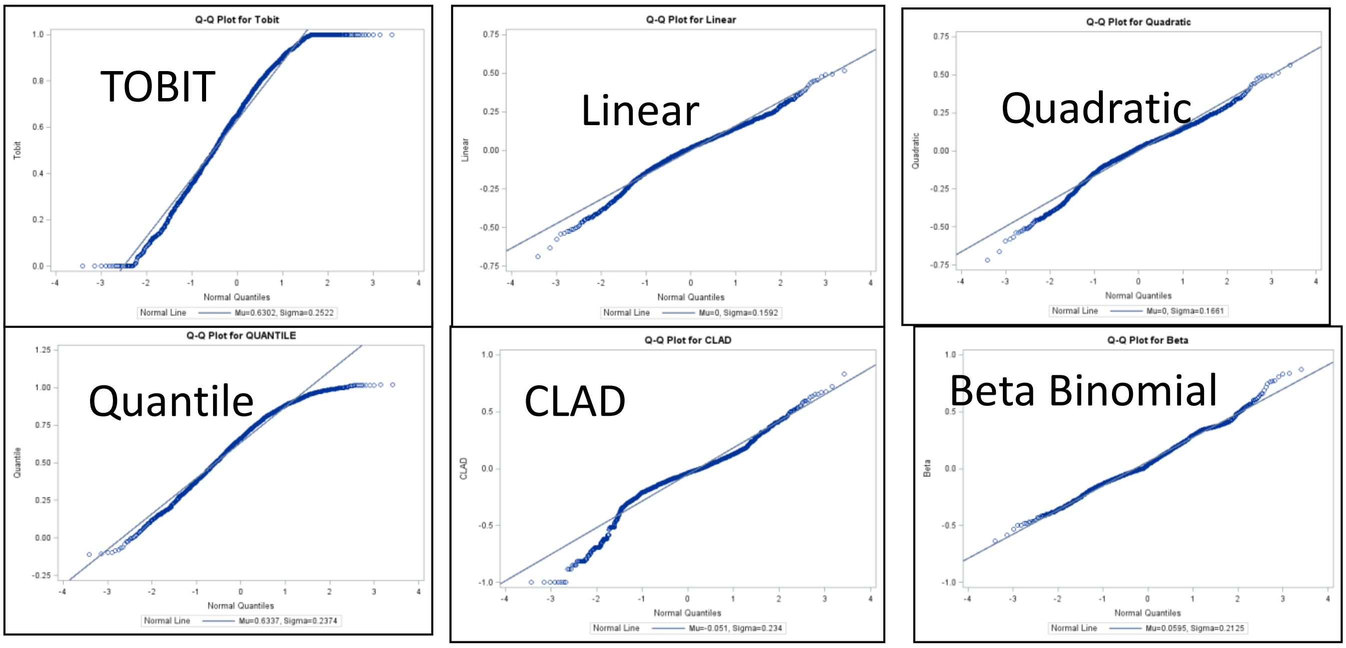
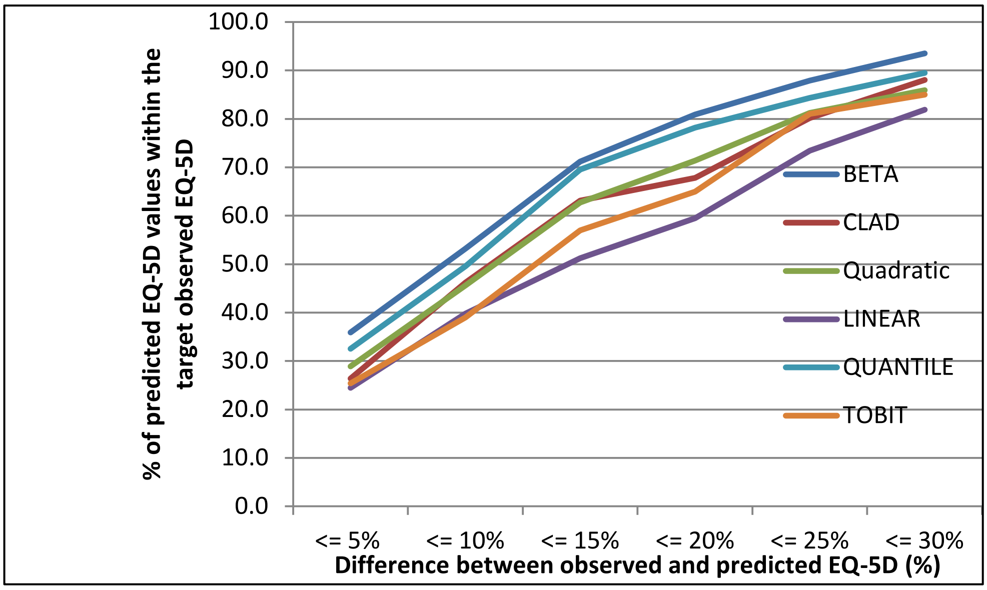


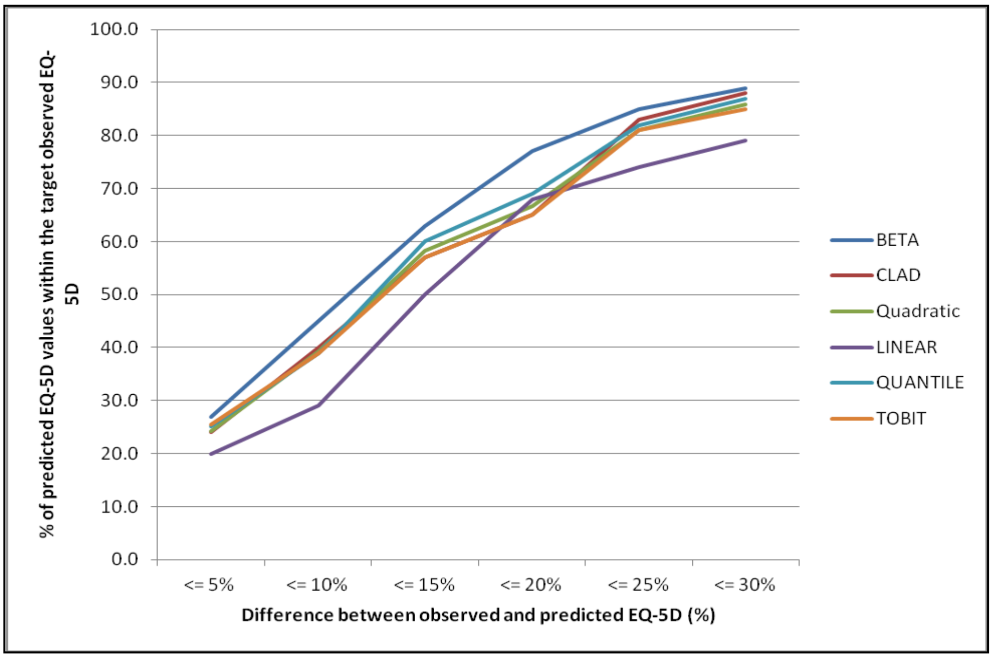
Figure 4.6: Normal Probability plots (TOPICAL)

Patient-level predictions using independent data

Comparing mean predicted EQ-5D-3L with the observed mean may not always be the best way to judge model performance because the distributions of the prediction values tend to cluster around the observed mean. For instance, an observed mean utility of 0.61 when compared with a predicted mean of 0.593 using CLAD (Table 4.3) is a difference of 0.017 (3%). However, about 40% of individual predicted values differed from the observed mean by about 10%. Therefore, for each patient, percentage differences within ±5% to ±30% of observed values were calculated. About 28% (BB) of predicted EQ-5D-3L were within ±5% (Figure 4.6) of the observed EQ-5D-3L, compared to 20% (Linear), 23% (TOBIT), 24% (Quadratic), 22% (Quantile) and 22% (CLAD) with SOCCAR data. Predictions were in general better with the BB model (the curve is above all others). The median prediction error for the BB model is about 10% for both TOPICAL and SOCCAR. Highest prediction errors are observed with the Linear model (median of 15% error for both TOPICAL and SOCCAR). The QLQ-C30 responses ranged from 0 to 100 for 14 out of the 15 domains (scores for the financial domain ranged from 30 to 80).



a) TOPICAL



a) SOCCAR

Figure 4.7: Models compared in terms of patient level predictions (a) TOPICAL, (b) SOCCAR

4.3.3 Over-prediction in Worse Health States

Mean predicted EQ-5D-3L at observed health states for each algorithm were shown earlier in Figures 4.4b, 4.5a, and 4.5b. The BB model had mean predicted EQ-5D-3L estimates closest to the observed values at a given observed health state for TOPICAL and SOCCAR, respectively. Differences between observed and predicted mean EQ-5D-3L for most models occur at about health states of 11321 (value on the x-axis of 32 in Figure 4.4b) for TOPICAL and about 22222 (x-axis value of 26) for SOCCAR. The Quadratic model under-predicted mean EQ-5D-3L at less severe health states compared to the BB model (Figure 4.5a, 4.5b).

Relationship between Health States and Adverse Events

The relationship between adverse event frequency for different definitions of ‘Good’ and ‘Poor’ health states was briefly investigated. The results suggest that patients with ‘Poor’ health (defined roughly here as >11321 for TOPICAL and >22222 for SOCCAR) are also the ones with a higher frequency of adverse events. In the TOPICAL trial, 24% of patients in ‘Poor’ health states (i.e. worse than 11321) experienced more than two grade 3 to 4 adverse events, compared to 15% for patients in ‘Good’ health states (health states 11321 or better); for SOCCAR this was 66% vs. 44%. Hence, there is some evidence that presence of adverse events is likely to influence utility prediction.

One reason why mapping algorithms might over-predict EQ-5D-3L at ‘Poor’ health states might be because treatment-related toxicity is not directly captured into the mapping algorithm, resulting in estimating a higher observed HRQoL. A similar pattern was also seen for other cut-offs that define ‘Poor’ and ‘Good’ health. For instance, when the cut-off for ‘good’ and ‘poor’ was defined as health states 21321 (EQ-5D-3L of 0.364) and > 22111 respectively, fewer patients in ‘good’ health states had AEs compared to patients with ‘poor’ health states: 17% vs. 26% of patients in ‘good’ vs. ‘poor’ health states had at least two adverse events in the TOPICAL trial. A similar pattern was observed for SOCCAR data. This suggests that there may be a more complex underlying mapping algorithm between EQ-5D-3L, QLQ-C30, and toxicity that may better explain the variability and prediction of EQ-5D-3L, chiefly in patients with ‘Poor’ health states.

4.3.4 Impact on QALY Estimates

Table 4.6 compares observed and expected QALYs from each of the models for SOCCAR and TOPICAL. The Observed QALY difference was 0.051 for TOPICAL (Erlotinib vs. Placebo) and 0.164 for SOCCAR (Concurrent vs. Sequential) [184]. Predictions from the BB model generated closest QALY estimates in both trials with a mean QALY difference of 0.053 for TOPICAL and 0.162 for SOCCAR (Table 4.6). QALY predictions from other models ranged from 0.041 (Linear) to 0.072 (Quadratic) for TOPICAL and 0.153 (Linear) to 0.208 (Quantile) for SOCCAR.

	TOPICAL			SOCCAR		
	Erlotinib	**Placebo**	**Difference**	**Concurrent**	**Sequential**	**Difference**
Observed	**0.35**	**0.30**	**0.051**	1.31	1.15	0.164
BB	0.34	0.29	0.053	1.53	1.37	0.162
TOBIT	0.37	0.31	0.064	1.59	1.42	0.174
CLAD	0.33	0.29	0.046	1.62	1.44	0.186
Quadratic	0.42	0.34	0.072	1.92	1.73	0.196
Mixed Linear	0.32	0.28	0.041	1.34	1.19	0.153
Quantile	0.38	0.31	0.070	1.42	1.62	0.208

Table 4.6: Comparison of estimated (mean) QALY's for all Algorithms

4.3.5 Adjustment for Demographic Variables

Several additional factors were added to the model for evaluation. Although R^2 changed slightly from 0.75 to 0.78 in TOPICAL with the inclusion of ECOG (P<0.001) and Gender (P<0.001), the underlying pattern of prediction shown in Table 4.2, Table 4.5 and Figure 4.4a and 4.4b did not vary. Hence, it can be concluded that adding demographic variables does slightly improve the model fit, but it does not have a major impact on predicted means and their standard errors.

The correlation between EQ-5D utilities and each of the 15 domains ranged from -0.60 (FA) to 0.62 (PF) for SOCCAR and -0.62 (PA) to 0.65 (PF) for TOPICAL (Tables 4.7 and 4.8). Higher (positive or negative) correlations suggest better overlap and possibility that mapping will be possible.

	EQ5D	PF	RF	EF	SF	CF	QL	FA	NV	PA	DY	SL	SP	CO	DI	FI
EQ5D	1.00															
PF	0.65	1.00														
RF	0.62	0.76	1.00													
EF	0.53	0.39	0.40	1.00												
SF	0.61	0.63	0.70	0.47	1.00											
CF	0.48	0.43	0.41	0.54	0.44	1.00										
QL	0.57	0.58	0.57	0.43	0.54	0.39	1.00									
FA	-0.62	-0.70	-0.71	-0.54	-0.67	-0.50	-0.63	1.00								
NV	-0.33	-0.29	-0.32	-0.25	-0.33	-0.26	-0.33	0.37	1.00							
PA	-0.62	-0.45	-0.43	-0.40	-0.46	-0.40	-0.40	0.51	0.34	1.00						
DY	-0.42	-0.55	-0.54	-0.37	-0.45	-0.36	-0.44	0.57	0.20	0.31	1.00					
SL	-0.42	-0.34	-0.33	-0.43	-0.32	-0.36	-0.35	0.44	0.20	0.41	0.33	1.00				
AP	-0.39	-0.43	-0.44	-0.34	-0.42	-0.28	-0.43	0.52	0.39	0.31	0.28	0.25	1.00			
CO	-0.30	-0.24	-0.25	-0.26	-0.27	-0.27	-0.21	0.29	0.21	0.32	0.21	0.22	0.23	1.00		
DI	-0.13	-0.11	-0.12	-0.11	-0.13	-0.16	-0.17	0.17	0.21	0.11	0.04	0.10	0.23	-0.03	1.00	
FI	-0.34	-0.28	-0.29	-0.30	-0.37	-0.28	-0.23	0.31	0.18	0.27	0.21	0.25	0.18	0.17	0.04	1.00

Table 4.7: Correlations between EQ-5D-3L and QLQ-C30 TOPICAL Data

	EQ5D	PF	RF	EF	SF	CF	QL	FA	NV	PA	DY	SL	SP	CO	DI	FI
EQ5D	1															
PF	0.62	1.00														
RF	0.60	0.68	1.00													
EF	0.56	0.47	0.44	1.00												
SF	0.57	0.56	0.60	0.47	1.00											
CF	0.43	0.42	0.40	0.56	0.43	1.00										
QL	0.57	0.58	0.56	0.48	0.53	0.37	1.00									
FA	-0.60	-0.66	-0.65	-0.57	-0.57	-0.56	-0.63	1.00								
NV	-0.30	-0.28	-0.29	-0.36	-0.34	-0.28	-0.32	0.39	1.00							
PA	-0.50	-0.40	-0.42	-0.43	-0.39	-0.37	-0.44	0.52	0.30	1.00						
DY	-0.40	-0.52	-0.51	-0.34	-0.41	-0.26	-0.38	0.42	0.30	0.23	1.00					
SL	-0.39	-0.32	-0.33	-0.39	-0.39	-0.34	-0.34	0.42	0.24	0.29	0.30	1.00				
AP	-0.41	-0.40	-0.39	-0.37	-0.36	-0.31	-0.44	0.50	0.41	0.33	0.28	0.27	1.00			
CO	-0.23	-0.20	-0.22	-0.28	-0.26	-0.25	-0.25	0.34	0.19	0.33	0.07	0.15	0.28	1.00		
DI	-0.12	-0.13	-0.15	-0.23	-0.16	-0.28	-0.12	0.26	0.16	0.27	0.17	0.08	0.08	0.25	1.00	
FI	-0.29	-0.37	-0.32	-0.32	-0.37	-0.35	-0.25	0.36	0.25	0.26	0.24	0.26	0.15	0.07	0.12	1.00

Table 4.8: Correlations between EQ-5D-3L and QLQ-C30 SOCCAR Data

Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF), Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI), Financial Problems (FI), Global Health Status Score

4.4 Discussion

Superior predictive properties have been demonstrated with a non-linear BB mapping algorithm, developed and tested using data from two independent lung cancer patient populations. Two mapping algorithms for different types of lung cancer patients (poor and better prognosis) have been shown to perform better than commonly used models. Either algorithm could be used, although there may be a preference for the one derived from the larger TOPICAL trial because of less uncertainty (Table 4.5) and better model fit (Table 4.2). Simulations assessed the uncertainty of mean estimates of EQ-5D-3L utilities. The degree of over-prediction of mean utilities at poorer health states was less with the BB model compared to other models. QALY estimates from models were also closer to the observed values with the BB model. These findings confirm previous untested assertions that the relationship between the EQ-5D-3L and QLQ-C30 may be better understood with a non-linear model structure [111,115,135].

Previous mapping models have mostly used OLS forms and are considered inadequate or over simplistic [112]. Most reported mapping models suffer from over-prediction in poorer health states. Previously reported models have reported R^2 values (using QLQC-30), ranging from 0.23 to 0.83 [112]. These values are similar to those reported in this study. It is very rare that models yield values of R^2 above 70%. Other models (e.g. multinomial, ordinal) have also been used, but have proved to be inadequate [128]. Estimates based on the absolute deviation (CLAD, Quantile, and adjusted censored models) predict patient level medians, whereas the statistic of interest is the mean. The Quadratic model takes into account non-linearity (by having squared terms in the model) but is essentially a linear model (linear in parameters because the coefficients are interpreted in the same way as linear models).

Improving model fit by "discarding" or "weighting" outliers with extreme values [98] is not an optimal solution if the extreme outliers can be modelled (rather than excluding observations). Moreover, there are many possibilities regarding the choice of variables to square and then combine with non-squared variables in quadratic models. For instance, squaring all 15 domain scores of the QLQ-C30 is a possible choice, and so is squaring 14 and having only one non-squared term remaining. Without conducting numerous tests and increasing the type I error, it is challenging to understand the relative merits of one set of variables over another. Therefore, this can lead to some arbitrary selection of combinations of terms in order to improve model fit.

There are several advantages of the BB approach. The BB model applied to this data confirms some superior statistical properties in terms of accuracy and efficiency [177-182]. The reasons for such superior properties was observed in the better fit with several statistics (AIC, R^2, and predictions). Given that these metrics are related to the underlying behaviour of the model in fitting the data, the BB model models the over dispersion and skewness much better than the other models (Figure 4.4). In addition, the parameters (a and b of 1.9 and 0.6 respectively for SOCCAR, Figure 4.1) strongly support a model that is appropriate for modelling skewed, over-dispersed and possible bimodally distributed utilities. The interpretation of the model outputs and parameters are still reasonably clear: mean utilities are estimated and the exponential of coefficients, have the same interpretation as an odds ratio. That is, for every unit change in a given QLQ-C30 domain score, the mean EQ-5D utility will change (increase or decrease) by an amount equivalent to exponentiation of the associated coefficient.

Clinical relevance may be important [128] if a generic measure is sufficient to provide both estimates of utilities as well as a clinical interpretation of HRQoL differences based on an odds ratio (which may be easier to interpret than mean differences). For example, the clinical relevance of a mean treatment difference of 0.012 on the EQ-5D-3L is difficult to judge. However, if this was equivalent to an odds ratio of 1.2 (a 20% improvement in HRQoL), a way of relating both clinical effects and utilities becomes feasible. This can also be extended to the response domains. This makes the BB a powerful and flexible mapping algorithm, relevant to health economic evaluation, policy, and clinical decision making.

The strength of this research lies in the approach to validating the model by using independent data and extensive multivariate simulation from correlated EQ-5D-3L and QLQ-C30 data. Plausible reasons as to why over-prediction at the poorer health states occur were explored by using adverse event data (collected in all trials), often ignored in mapping algorithms. It is possible that joint relationships between adverse events and EQ-5D-3L may offer an explanation for over-prediction because higher toxicity was observed in poorer health states. Some researchers suggest that EQ-5D-3L responses have a bimodal distribution and therefore two separate mapping algorithms might be needed (Veerstegh, 2010) for patients in ‘Poor’ and ‘Good’ health states [129]. The nature of the bimodality could be explored using baseline clinical data (e.g. using baseline ECOG).

This research has several limitations. Firstly, the impact on results for other values of α and β has not been exploited. In this application, α and β were set to model the mean EQ-5D-3L;

other possibilities may include searching for α and β which might optimize R^2, minimize MSE and improve predictions. Secondly, a scale of 0 to 1 was assumed, which may not be suitable for certain diseases, where the states worse than death are significant. Surprisingly, even in this NSCLC population, the proportion of such cases was very low. Statistically, this might seem fine because estimates may not be affected (biased) too much. However, conceptually it is placing someone whose health state is ‘worse than death’ equivalent to death. Thirdly, model validation has been limited to lung cancer data and further testing would be useful in both lung cancer (to see whether algorithms are tumor specific) and non-lung cancer data sets (to check for generalizability). Finally, the BB was not compared with other models such as Bayesian network models that report superior predictive properties when compared to the more common models. However, in the case of such Bayesian models, the choice of the initial (prior) estimates of the probability of EQ-5D-3L responses can influence the predicted utility.

In addition, there are several concerns when using mapping functions, a point that has been repeated in past research [112,112,135,118,185]. One key concern is that it is not possible to know whether the predicted utilities are close to the observed values unless both are known. Secondly, there are questions as to what exactly is being measured or estimated because some key information in one instrument is not included in the other, particularly when predicting EQ-5D-3L from clinical measures alone. One approach might be to evaluate the psychometric properties of the two instruments and also analyse the correlations. A weak correlation (Spearman’s or linear) might explain a poor mapping algorithm.

The first concern regarding mapping can be partially answered with the use of simulation by quantifying uncertainty in how well the predicted values approximate the observed utilities and can be quantified as described in Table 4.5. This does not inform us as to what the predicted EQ-5D-3L is *actually* measuring, but it is assumed that the closer the predicted values are to the observed, the preferences become ‘essentially similar’. If in 90% of simulations, the observed and predicted values are close, it may be reasonable to assume that the mapping algorithm provides estimates that are measuring aspects of “essentially similar” preferences, which for practical purposes might be acceptable. Moreover, the statistical significance of several predictors might also inform us about health state preferences. If a model predicts every EQ-5D-3L perfectly, then one may wish to conclude that the model has correctly predicted the ‘essential nature’ of the preferences (ultimately contained in a single index), or remain skeptical and seek additional evidence to confirm that.

Other concerns with mapping involve time points used when developing and applying an algorithm. For instance, including baseline data in a model, which aims to predict post-baseline treatment differences, may lead to misleading estimates. Assumptions that the rates of change in EQ-5D-3L (the coefficients of the QLQ-C30) are constant from one cancer type to another are also unlikely to hold. If baseline or demographic variables are used, the relevance of these variables for the target population may be important (e.g. if an algorithm is applied only to a male population). Finally, the mapping algorithm should offer a reasonable clinical interpretation. In lung cancer, for example, it might be expected that dyspnoea is an important predictor of HRQoL. In some models, dyspnoea was not shown to be statistically relevant for predicting EQ-5D-3L, although it is an important symptom in lung cancer patients. A model that might have practical relevance is one where for example as dyspnoea symptoms worsen (scores increase) predicted EQ-5D-3L utilities get lower which is not always the case.

4.5 Conclusion

The Beta-Binomial regression approach indicates superior performance compared to published models in terms of predicting the observed EQ-5D-3L from QLQ-C30 in these lung cancer trials. This non-linear approach may offer advantages over existing models for mapping and as a general approach for modelling utilities. These results confirm previous observations that the HRQoL is over-estimated at the poorer heath states. The reasons why current algorithms persistently over-predict at poorer health states requires further interrogation, perhaps incorporating adverse event information into the models. Guidelines on using algorithms may also be beneficial. The mapping may be useful, however, there are still concerns as to whether the predicted utilities are essentially the same as the observed values.

The next chapter (Chapter 5) will consider the impact of mapping using the more recent EQ-5D-5L and investigate whether the change in the scale of measurement improves the prediction and model fit.

Chapter 5

Chapter 5: Comparing Mapping between EQ-5D-5L, EQ-5D-3L, and EORTC-QLQ-C30

Published: Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patient (Iftekhar Khan, Steve Morris, Nora Pashayan, Bashir Matata, Zahid Bashir and Joe Maguirre; Health and Quality of Life Outcomes (2016) 14:60; DOI 10.1186/s12955-016-0455-1)

Abstract

Introduction: Several mapping algorithms have been published with the EORTC-QLQ-C30 for estimating EQ-5D-3L utilities. However, none are available with EQ-5D-5L. Moreover, a comparison between mapping algorithms in the same set of patients has not been simultaneously performed for these two instruments. In this prospective study of 100 non-small cell lung cancer (NSCLC) patients, the performance of three mapping algorithms using the EQ-5D-3L and EQ-5D-5L were compared.

Methods: A prospective non-interventional cohort of 100 NSCLC patients were followed up for a period of at least 12 months. EQ-5D-3L, EQ-5D-5L, and EORTC-QLQ-C30 were assessed on a monthly basis. EQ-5D-5L was completed at least a week after EQ-5D-3L. A random effects linear regression model, a Beta-Binomial (BB) and a Limited Variable Dependent Mixture (LVDM) model were used to determine a mapping algorithm between EQ-5D-3L, EQ-5D-5L, and QLQ-C30. In addition, simulation and other statistical measures were used to compare the performances of the algorithms.

Results: It was identified that mapping from the EQ-5D-5L was better lower AIC, RMSE, MAE, and higher R^2 were reported with the EQ-5D-5L than with EQ-5D-3L, regardless of the functional form of the algorithm. The BB model appeared to be more useful for both instruments; for the EQ-5D-5L, AIC was -485, R^2 of 75%, MAE of 0.075 and RMSE was 0.092. For EQ-5D3L, these values were -385, 69%, 0.099 and 0.113, respectively. The mean observed utilities were 0.572 and 0.515 for EQ-5D-3L and EQ-5D-5L respectively. The mean predicted utilities were 0.577, 0.575 and 0.569 for the random effects, BB and LVDM models for EQ-5D-5L; for EQ-5D-3L, these values were 0.523, 0.518 and 0.532, respectively. Less over-prediction at poorer health states was also observed with EQ-5D-5L.

Conclusion: The BB mapping algorithm is confirmed to offer a better fit for both EQ-5D-3L and EQ-5D-5L. The results are consistent with previous and more recent results on the use of BB type modelling approaches for mapping.

5.1 Introduction

The advantages and limitations of mapping were discussed in the previous chapters. Recently, Crott (2014), Arnold et al. (2015) and Doble and Lorgelly (2015) [161-163] examined the performance of the most common mapping algorithms applied to the QLQ-C30. Several limitations of some of the simpler mapping algorithms from the EQ-5D-3L were also noted. These limitations are related to untenable assumptions around linearity, homoscedasticity, multimodality, skewness, and censoring as the metric of model performance; and in some cases poor over prediction, particularly at poorer health states [109,134,139,161,162].

For this thesis, the performance of three mapping algorithms (from QLQ-C30) - a Random Effects linear model, Beta-Binomial (BB) and Limited Dependent Variable Mixture Model (LDVMM) were compared, for each of two utility measures: EQ-5D-5L and EQ-5D-3L, separately. Currently, no study of mapping compares algorithms from *both* instruments in the same set of patients; and none are available between EQ-5D-5L and QLQ-C30, particularly from a non-small cell lung cancer (NSCLC) patient population. In the previous chapter, using data from a randomized controlled trial (RCT) [109,165,166], a three-part BB model was reported to perform the best amongst other commonly used algorithms. This analysis examines mapping models using data from NSCLC patient in a real world NHS setting. This will offer researchers a way to compute patient-level utilities from the EQ-5D-5L (and EQ-5D-3L) with greater generalizability than algorithms using data from a RCT.

5.2 Methods

Study Design

A single cohort prospective (non-interventional) follow-up study in 100 NSCLC patients was designed and executed. Details are provided in Chapter 3 for Study 3.

Assessments

Described in Chapter 3.

Sample size

Described in Chapter 3.

Statistical Methods

Three models were used for mapping:

(i) Linear Random Effects Model

The linear model with a random effect is an extension of the ordinary least squares (OLS) model. One significant difference is that subject level effects are included in this model. In the context of mapping, as utility scores are observed for each subject on more than one occasion, the responses are not independent. The subject level differences (between subject variability) can be modelled with a random effect. For this reason, the model is often termed as a mixed effects model due to the variability of utilities between and within subjects. This model is relatively easy to use when applied to an external data set for predicting patient level utilities. This is important because, in practice, a mapping algorithm should also have a feature that can be used practically and conveniently. More complicated models require more assumptions and hence introduce greater uncertainty. The model form in a general linear mixed model framework is:

$$\mathbf{Y= X\beta + Z^{*}u + \varepsilon}$$

Where $\boldsymbol{\beta}$ is a matrix with the fixed effects parameters (e.g. the 15 coefficients of the QLQ-C30, as continuous outcomes) and **u** is a matrix (or vector) with the random (subject) terms and $\boldsymbol{\varepsilon}$ is the experimental error term (corresponding to the fixed effects). Mapping models for the QLQ-C30 typically use the 0 to 100 scoring range to estimate EQ-5D values

(ii) Limited Dependent Variable Mixture Model (LDVMM)

A second model [134] belonging to the class of limited dependent variable (LDV) models is the Adjusted Limited Variable Dependent Mixture Model (ALVDMM) [134]. This particular model has several noteworthy features. Firstly, it assumes additivity of effects (as in a linear model). Moreover, it involves a latent variable that is censored. The censoring occurs (similarly to that applied in a TOBIT model) because values are considered to be unobservable. Hernandez et al. (2012) [134] noted that since there is a gap in utilities between the values 0.833 and 1 for the EQ-5D-3L, the preferences for health states are in effect ‘cut-off’ on the higher side of values at (or above) 0.833 to a value of 1 (essentially capturing the ceiling effect). This means that if a patient’s (true) utility is >0.833 and $\leq$1, the instrument (EQ-5D) cannot capture this and a value of 1 is assumed.

The LDV type models generate predicted estimates in a more complex way, which involves finding the probability that the unobserved (latent) value is above or below the censored threshold value (e.g. 0.833) using the ratio of the probability density function (PDF) to the cumulative density functions (CDF). This feature of the LDVs allows the possibility to simultaneously model the presence of several distributions. Previously, mapping was determined using data from a relatively short health assessment questionnaire (HAQ) in an arthritis population [134]. The greater the number of latent classes, the more complex the interpretation. Application of three classes (thresholds) in the context of 15 QLQ-C30 domain parameters (using a scale of 0 to 100 as used in some other published approaches) is likely to lead to a much more complex latent class structure and hence, two classes (two mixed distributions) are used for both the EQ-5D-3L and EQ-5D-5L in this analysis. This is justified by observing the kernel density estimates that suggest a bimodal distribution for EQ-5D-3L (values between about -0.549 to <0.3 and ≥0.3 to 1) in this data set (see Figure 5.1 below). Although Khan et al (2015) showed that the QLQ-C30 appear to fall into discrete categories, it does not follow that the true distribution is discrete [155]. For later Bayesian approaches (Chapter 7) a discrete and a continuous form of the model is implemented. The implication of using a discrete or continuous scale is likely to be reflected in the precision of estimates. A Bayesian approach is probabilistic and with 15 domains and 5 categorical responses is likely to be more uncertain (as will be shown in Chapter 7).

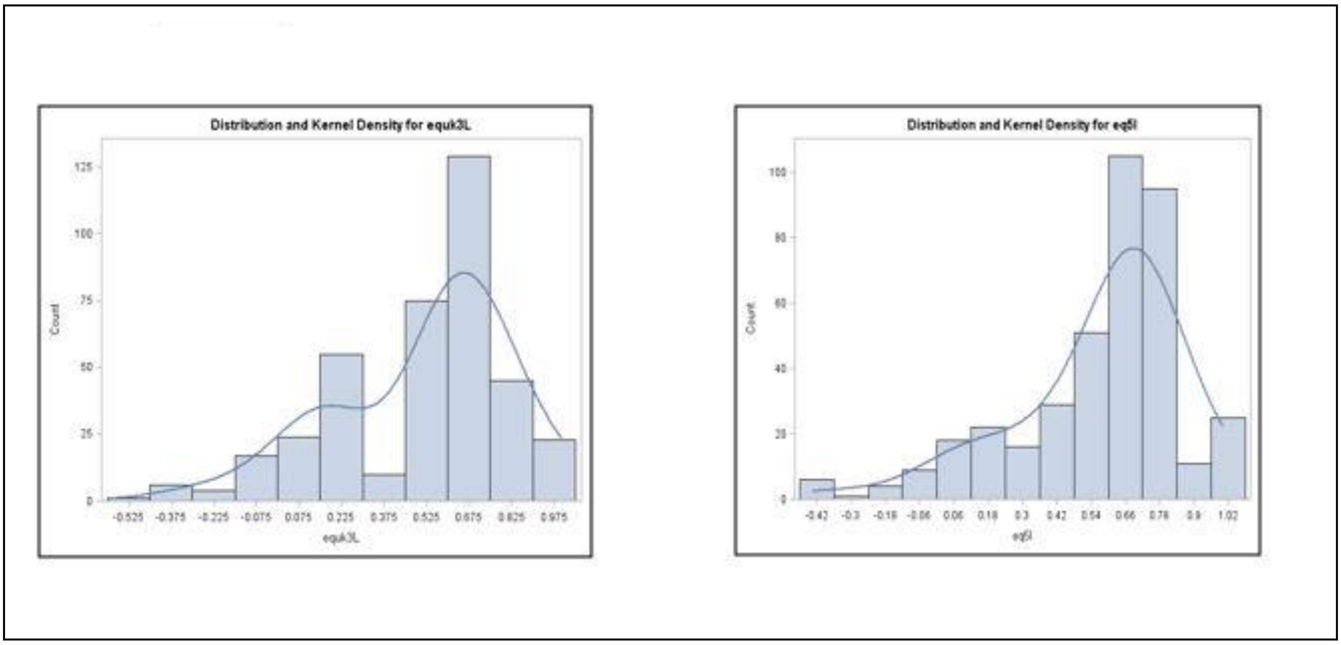


Figure 5.1: Distribution of EQ-5D-3L (left) and EQ-5D-5L (right) utilities

For the EQ-5D-5L, the mixture of distributions is not obvious, although there is marked skewness. The form for the mixture model used in this context is described below.

Assuming responses ***Y*** (i.e. EQ-5D utilities), whose distribution depends on an unobservable random variable ***S***; ***S*** can occupy one of ***k*** states (***k***=2 in this instance), the number of which might be unknown, but is at least known to be finite. Since ***S*** is not observable, it is referred to as a latent variable. Let π_j denote the probability that ***S*** takes on state ***j***. For instance, in the case of the EQ-5D-3L for the ALVDMM, **j=1** might refer to values of EQ-5D-3L < 0.833 and ***j=2*** would refer to states, such that EQ-5D-3L utilities are $\geq$ 0.833 and $\leq$1.0.

Conditional on ***S***, the distribution of the response ***Y,*** is assumed to be $\mathbf{f_j\ (y;\alpha_j,\ \beta\ _j|}$ ***S=j).*** What this expression (i.e. ($\mathbf{f_j(y;\alpha_j,\ \beta\ _j|\ S=j)}$) means is that depending on the number of states (S), a model (with a form $\mathbf{f_j(y;\alpha_j,\ \beta}$) can be used to determine the relationship between Y (the EQ-5D) and a set of predictors, β **(**e.g. the 15 QLQ-C30 coefficients). For instance, for **j=1** (values of EQ-5D-3L between -0.549 and 0.3), the EQ-5D-3L are assumed to follow a Normal distribution. For values between 0.3 and 1 (**j=2**), the data can be considered to follow a Beta-Binomial (BB) distribution. In another scenario, for **j=1**, a Weibull function could be used, and for **j=2** a Normal distribution could be used. There would be six parameters to estimate (two parameters for the Weibull, two parameters for the Normal and consequently two mixing probabilities (π_1 and π_2), the probability of observations belonging to one or another class. The six parameters to be predicted do not include any of the QLQ-C30 predictors (parameters), where a further 16 parameters are estimated.

The following mixture models were simultaneously fitted:

(i) EQ-5D as a function of 15 QLQ-C30 domain scores (for example, Normal Distribution assumed between -0.549 and 0.30)
(ii) EQ-5D as a function of 15 QLQ-C30 domain scores (for example, Beta-Binomial distribution assumed between 0.30 and 1)
(iii) The mixing probabilities as a function of the 15 QLQ-C30 domain scores (two mixing probabilities that classify observations as belonging to distributions in (i) or (ii)).

Evidently, the above modelling approach is complex and perhaps unnecessary; which can lead to model non-convergence. The models practical implementation as an external algorithm is, hence, an important consideration. A transformation may be carried out if specific distributions are assumed (e.g. modelling negative values). For instance, for values between -0.549 and $\geq$0.30, a Gamma (or Beta-Binomial) distribution would not be possible.

Hence, in this analysis two distributions are considered for modelling:

(i) Assume Normality between -0.549 and <0.30 for the 15 predictor variables
(ii) Assume Beta-Binomial between 0.30 and 1.0 for the 15 predictor variables

The predicted estimates are determined in a complicated way from the ratio of the CDF to the PDF of the EQ-5D responses and using the estimated mixing probabilities. The mixing probabilities can be interpreted as the ratio of observations belonging to one of two distributions. If the mixing probabilities were 0.5, then 50% of the EQ-5D-3L might be considered to follow a normal distribution and the remaining 50% a different distribution. A useful exposition of finite mixture models can be found in Schlattman (2009) [189].

Maximum likelihood estimation for continuous and discrete response distributions is used based on a dual quasi-Newton optimization algorithm using SAS® software [190]. A global maximum was sought using initial starting values to search for local maxima, followed by re-running the model using estimates generated from previous model runs.

(iii) Beta-Binomial Model

For the previously used ALVDMM, censoring occurs for values at 0.833 for the EQ-5D-3L. This is not the case for the EQ-5D-5L, where values between 0.833 and 1 do exist. For this reason (Figure 5.1) the distribution of the EQ-5D-5L can be considered appropriate for modelling on a continuous type scale between -0.549 and 1.0 (after a transformation of Y-a/b-a), and hence, the BB model is the third model that is considered for mapping. The details of the BB model were elaborated in the previous chapter and discussed in Khan and Morris (2014) [109]. The BB model demonstrated an improved fit compared to simpler linear and LDV models (e.g. TOBIT and CLAD).

Model Performance Criteria

Several model performance statistics were used, including the root mean square error (RMSE), which is a type of model fit measure (lower values indicate better fit), mean prediction error, R^2, mean absolute error (MAE), and percentage predicted >1 and < -0.594. Chai (2014) argues that the RMSE is more appropriate than the MAE, specifically when the error distribution is Normally Distributed [191]. The Aikakes Information Criteria (AIC) values and percentage predicted within a target range (e.g. ±5%, ±10%) of the observed values were also computed.

Simulation and Cross Validation

Multivariate simulation (1,000 simulations using Fleishman methods) [167,168] were used to test the uncertainty of the models. The method of Fleishman uses higher order moments (e.g. kurtosis and skewness) to generate correlated simulated data, regardless of the distribution of each of the original variables. The steps involved in simulation require computing the mean, SD, skewness and kurtosis for each of the observed 15 QLQ-C30 domain scores. Using the Fleishman (1978) [167,168] power transform:

$$\mathbf{Y=\alpha + \beta*Z +\delta*Z^2+\gamma*Z^3,}$$

The values of α, β, δ and γ are estimated from randomly generated data **Z**, normally distributed with a mean of zero and a variance of 1 and the observed measures of kurtosis and skewness. The values of α, β, δ and γ are estimated through a process of iteration so that **Y** can be determined. The derived **Y** (e.g. 15 QLQ-C30 scores) are simulated (correlated) responses, which are not necessarily normally distributed. Khan & Morris (2015) [109] have shown that the QLQ-C30 scores are unlikely to follow a normal distribution in most cases.

For each simulated data set, cross-validation was used. Half (50%) of the simulated data set (randomly selected) was used to develop the mapping model and the other half was used to test the model (out of sample predictions). For each realization (i.e. data set simulated), the model performance statistics (e.g. RMSE and R^2) were generated and reported. Although there is no theoretical reason for 50% of the data used for developing the model, other cut-offs (e.g. 75% vs. 25%) were also considered.

Handling missing data

Where missing data occurred multiple imputation approaches were used – using chain equations (ref). This involved generating complete data sets (3 were generated) and modelling using each of the data sets was used and the results were summarized for sensitivity analyses.

5.3 Results

Between the period of March 2014 and July 2015, a total of 100 patients consented and were registered for follow-up. Out of these, two patients withdrew before the follow-up started. Consequently, 98 patients (98%) were included in the statistical analysis; 23 patients (23%) died during the follow-up and 2 patients (2%) dropped out due to personal reasons (Figure 5.2 CONSORT). There was a total of 985 observations (responses), across 98 patients for EQ-5D-5L and EQ-5D-3L HRQoL forms, respectively. HRQoL forms were completed by 97/98 (99%) patients at baseline; completion rates at 3 and 6 months were 78/98 (79%) and 41/98 (55%) respectively. Also, completion rates were similar for all the three (EQ-5D-5L, EQ-5D-3L, and QLQ-C30) instruments.

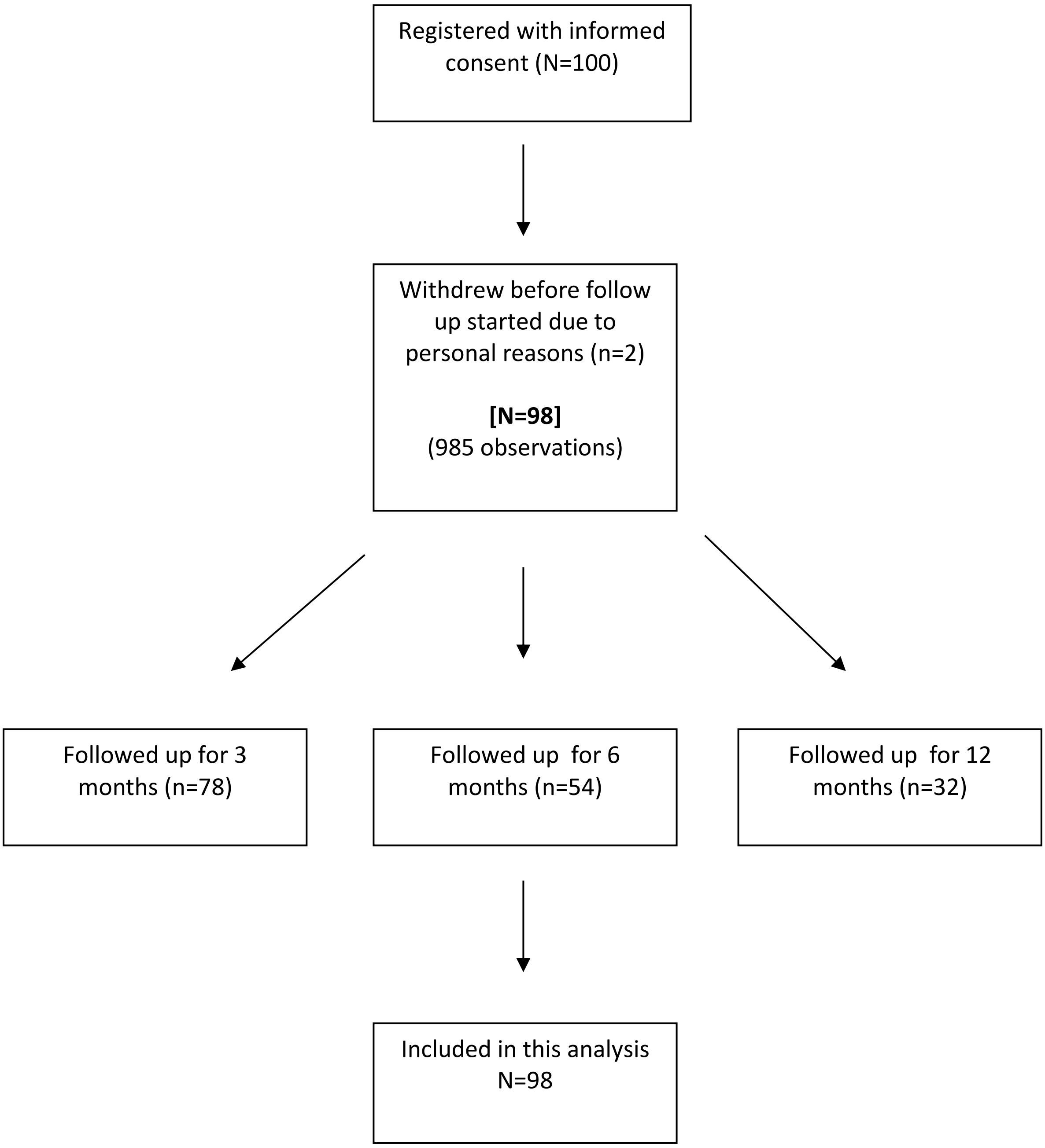

Registered with informed consent (N=100)
Withdrew before follow up started due to personal reasons (n=2)
[N=98]
(985 observations)
Followed up for 3 months (n=78)
Followed up for 6 months (n=54)
Followed up for 12 months (n=32)
Included in this analysis N=98

Figure 5.2: CONSORT for Study 3

There were 146 observed health states (5% of all possible health states) observed with EQ-5D-5L and 62 (26%) for EQ-5D-3L. The most frequent health states with the EQ-5D-5L were 11111 (6%), followed by 21222 (5%), 43533 (3%) and 31331 (3%). For EQ-5D-3L these health state values were 21222 (11%), followed by 22222 (10%), 22221 (7%), 22322 (6%) and 11111 (6%).

Demographics

The median age of the respondents was 69 years (range 39 to 86); 55/98 (56%) were male, 67/98 (68%) were ex-smokers and 19/98 (19%) current smokers. On Easter Co-operative Oncology Group (ECOG) performance status, there were 65/98 (64%) patients who were in grade 0-2 and the remaining were ECOG >2. ECOG is used as a measure of well-being (and prognosis), with higher values suggesting poorer prognosis; 26/98 (27%) were Stage I-II and 68/98 (69%) were Stage III and higher; Histology subtypes were 43/98 (44%) with adenocarcinoma and 36/98 (37%) with the squamous cell. The remainder were of varying subtypes (Table 5.1).

	(N=98)
Age (Median years, Range)	69 (39-86)
Gender	
Male	55 (56%)
Female	43 (44%)
Smoking Status	
Current Smoker	19 (19%)
Ex-Smoker	67 (68%)
Never	5 (5%)
Unknown	7 (7%)
Stage	
I -II	26 (27%)
III	31 (32%)
IV	37 (38%)
Unknown	4 (4%)
Histology	
Adenocarcinoma	43 (44%)
Squamous	36 (37%)
Mesothelioma	5 (5%)
Other	14 (14%)
ECOG:	
0: Normal activity	12 (12%)
1: Near full activity	23 (23%)
2: In bed < 50% of time	30 (31%)
3: In bed > 50% of time	27 (28%)
4: Totally confined to bed	4 (4%)

Table 5.1: Baseline and Demographics Characteristics

Performance of EQ-5D-5L and EQ-5D-3L Mapping Algorithms

The best performing model regardless of EQ-5D-3L or EQ-5D-5L was the BB model (Table 5.2). This had an AIC, R^2, RMSE, MAE and % predicted to within ±5% and ±10% of - 485, 75%, 0.092, 0.075, 29% and 59% for EQ-5D-3L and -385, 69%, 0.113, 0.099, 21% and 47% for EQ-5D-5L, respectively. The BB, therefore, had good model fit characteristics and predicted more utilities to within ±10% of the observed value compared to other models, particularly for the EQ-5D-5L.

	EQ-5D-5L			EQ-5D-3L		
	Random Effect	**Beta-Binomial**	**LVDM***	**Random Effect**	**Beta-Binomial**	**LVDM***
R^2	72%	75%	70%	67%	69%	67%
AIC	-365	-485	-383	-291	-385	-189
RMSE	0.152	0.092	0.153	0.183	0.113	0.179
MAE	0.114	0.075	0.115	0.141	0.099	0.139
Predicted Mean (SD)	0.577 (0.241)	0.575 (0.211)	0.569 (0.217)	0.523 (0.252)	0.518 (0.183)	0.532 (0.252)
Observed Mean (SD)	**0.572 (0.224)**	**0.572 (0.224)**	**0.572 (0.224)**	**0.515 (0.308)**	**0.515 (0.308)**	**0.515 (0.308)**
%predicted outside range	<1%	0	0	<1%	0	0
Predicted within +5%	19%	29%	20%	19%	21%	20%
Predicted within +5%	38%	59%	42%	37%	47%	35%

*Normal + Beta Mixture

Table 5.2: Comparison of Model Performance

Random Effects Model

The performance of the random effects model was comparable to the LDVMM. Table 5.3 shows the parameter estimates for the 15 QLQ-C30 coefficients.

*Statistically significant at the two-sided 5% level

	EQ-5D-5L			**EQ-5D-3L**		
	Estimate	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	0.2255	0.09157	0.0142	0.08046	0.08507	0.3450
Physical Functioning	**0.006718***	0.000676	<.0001	**0.005437***	0.000620	<.0001
Role Functioning	-0.00032	0.000591	0.5935	**0.001392***	0.000509	0.0066
Emotional Functioning	**0.001871***	0.000554	0.0008	**0.001949***	0.000481	<.0001
Cognitive Functioning	-0.00057	0.000491	0.2436	-0.00073	0.000448	0.1024
Social Functioning	0.000387	0.000530	0.4664	0.000516	0.000462	0.2652
Global Health Status / QoL	-**0.00109***	0.000409	0.0082	-0.00043	0.000401	0.2853
Fatigue	0.000324	0.000696	0.6420	0.000993	0.000647	0.1261
Nausea / Vomiting	-0.00041	0.000600	0.4990	0.000276	0.000524	0.5993
Pain	**-0.00290***	0.000495	<.0001	**-0.9**	0.000427	<.0001
Dyspnoea	0.000368	0.000464	0.4287	-0.00011	0.000421	0.7915
Insomnia	-0.00017	0.000338	0.6218	-0.00004	0.000313	0.9053
Appetite loss	-0.00030	0.000328	0.3673	0.000341	0.000295	0.2488
Constipation	-0.00013	0.000359	0.7139	0.000524	0.000306	0.0877
Diarrhoea	**0.001155***	0.000438	0.0087	0.000499	0.000425	0.2409
Financial Problems	0.000345	0.000334	0.3019	-0.00004	0.000297	0.9039

Table 5.3: Results from Statistical Modelling (Random effects Model)

If all the scores for the Functional, Global and Finance domain scores are assumed to be perfect (i.e. score of 100) and no signs and symptoms are present (i.e. score of 0), the predicted EQ-5D-3L and EQ-5D-5L mean scores are estimated to be about 0.89 and 0.96, respectively. In contrast, if symptom and functional scores are the worst possible (scores of 0 and 100 for function and symptoms, respectively), the predicted EQ-5D-3L and EQ-5D-5L results in nearly 0.10 and 0.09, respectively. EQ-5D-5L, therefore, predicts higher at both extremes (Table 5.4).

	QLQ-C30 Score		Predicted	
Model	**Symptom**	**Function**	**EQ-5D-3L**	**EQ-5D-5L**
Random effects	Best (0)	Best (100)	0.89	0.96
	Worst (100)	Worst (0)	0.10	0.019
Beta Binomial	Best (0)	Best (100)	0.901	0.983
	Worst (100)	Worst (0)	0.097	0.0094
LDVMM	Best (0)	Best (100)	0.884	0.972
	Worst (100)	Worst (0)	0.055	0.008

Table 5.4 Predicted Utilities from Three Scenarios

Beta-Binomial Model

The BB (Table 5.5) can be used to predict the EQ-5D using a standard logit link: P/1-P = exp (-α +$\boldsymbol{\beta}$**X**), such that P = 1/1+exp (-α +$\boldsymbol{\beta}$**X**), where P indicates the predicted EQ-5D and **X** are the QLQ-C30 scores. Hence, the predicted EQ-5D-5L are 0.983, approximating the value 1.00. Following a similar approach to the above, the first step is to predict the EQ-5D using the estimates in Table 5.5. Setting the functional scores of the EQ-5D-3L to perfect HRQoL for the two functions and symptom scores (score = 100 and 0 respectively), the predicted EQ-5D-5L is estimated as:

1/[1 + exp(-α +$\boldsymbol{\beta}$**X**) = exp[0.2255 + (100*PF+100*SF +……+0*FA ….+0*FI)] = 0.983.

	EQ-5D-5L			EQ-5D-3L		
	Estimate	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	-1.51144	0.00006	<0.001	-0.0123	0.003893	0.00248
Physical Functioning	**0.03644***	0.004666	<0.001	0.01918	0.00294	**<0.001***
Role Functioning	**0.009619***	0.00455	0.03867	0.00421	0.002685	0.12215
Emotional Functioning	**0.01904***	0.003192	<0.0001	0.00661	0.002007	**0.00166***
Cognitive Functioning	-0.00633	0.003312	0.06076	-0.00425	0.002111	**0.04858***
Social Functioning	-0.00013	0.002758	0.9712	-0.00035	0.001973	0.8598
Global Health Status / QoL	0.001652	0.002772	0.55344	-0.00197	0.001913	0.30724
Fatigue	0.003561	0.005282	0.50279	0.00443	0.002979	0.14223
Nausea / Vomiting	0.000452	0.004514	0.92057	-0.00146	0.0027	0.59069
Pain	**-0.03569***	0.003512	<0.001	-0.03278	0.00191	**<0.001***
Dyspnoea	**-0.00806***	0.0028	0.00553	0.00015	0.001759	0.93233
Insomnia	0.002047	0.002388	0.39474	0.00193	0.001491	0.20048
Appetite loss	**0.005383***	0.002446	0.03161	0.0002	0.001415	0.88807
Constipation	0.000454	0.002052	0.82565	0.0014	0.001386	0.3165
Diarrhoea	0.000353	0.00274	0.20705	0.00393	0.001841	**0.03688***
Financial Problems	-0.00432	0.002182	0.07174	-0.00113	0.001292	0.38527

Table 5.5: Results from Statistical Modelling (BB Model)

*Statistically significant at the two-sided 5% level

LDVM

The LDVM model estimates are more complicated to generate as they involve two distributions and two mixing probabilities. Consequently, more than 32 parameters are involved in determining predictions for the best-worst case scenarios (Table 5.6). The LDVMM also predicts well at extremes, despite similar R^2 and RMSE to the random effects model (Table 5.4). However, the LDVMM is much more complex to use as an algorithm. Users would also need to know details of the mixing probabilities, as well as make stronger assumptions about the mixed distribution. Other mixtures were also considered, but the Normal/Beta mixture offered the best (smallest AIC) fitting model (Table 5.7).

	EQ-5D-5L				**EQ-5D-3L**			
	Normal		**Beta**		**Normal**		**Beta**	
	Estimate	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	0.07353	0.05925	-0.7052	0.4046	0.1032	0.1008	0.1579	0.8373
Physical Functioning	**0.008668***	0.000515	**0.01394***	0.002851	**0.007667***	0.000771	-0.01009	0.005942
Role Functioning	0.00034	0.000439	**0.01271***	0.002239	0.000961	0.000943	**0.01046***	0.003785
Emotional Functioning	**0.002680***	0.000457	0.003145	0.002045	**0.001808***	0.000593	-0.00191	0.004717
Cognitive Functioning	**-0.00141***	0.000367	**-0.00521***	0.001998	**-0.00127***	0.000603	0.000919	0.003645
Social Functioning	-0.00085	0.000475	0.001153	0.001935	0.000355	0.000651	**0.009044***	0.003982
Global Health Status / QoL	0.00025	0.000236	-0.00051	0.00203	**-0.00151***	0.000424	0.001184	0.004713
Fatigue	0.000698	0.000519	-0.00074	0.002929	**0.002149***	0.000875	-0.01064	0.006315
Nausea / Vomiting	**-0.00063***	0.000368	0.001293	0.002378	0.000278	0.000649	-0.00853	0.00621
Pain	**-0.00662***	0.000343	-0.00094	0.001835	**-0.00584***	0.000568	**0.01325***	0.005386
Dyspnoea	**0.001407***	0.000476	**-0.00576***	0.001807	0.00064	0.000488	**-0.00791***	0.004004
Insomnia	0.00018	0.000239	-0.00156	0.001351	0.00029	0.000384	**-0.00656***	0.002875
Appetite loss	**-0.00085***	0.000253	**0.008535***	0.001406	**-0.00081***	0.000388	**0.009893***	0.002344
Constipation	**0.002190***	0.000261	-0.00215	0.001445	**0.001571***	0.000382	**-0.00631***	0.003363
Diarrhoea	**0.001377***	0.000289	-0.00265	0.001942	0.000749	0.000563	0.005638	0.00477
Financial Problems	**-0.00102***	0.00026	0.001778	0.001291	0.000539	0.000337	0.004688	0.00282

Table 5.6: Results from Statistical Modelling (LDVMM: Normal + Beta)

*Statistically significant at the two-sided 5% level

Mixture	**AIC**	
	EQ-5D-5L	**EQ-5D-3L**
Normal /Beta	-383.2	-189.1
Normal/Gamma[#]	-250.5	-250.5
Normal/Weibull[#]	-252.4	-128.4
Normal/Log Normal	-242.0	-124.4

Table 5.7: Comparison of Model Performance of other Mixture Models

[#] Model convergence problems resulted in some parameters not estimated and/or mixing probabilities not calculable.

Health States

EQ-5D-3L predictions by health state were generally as observed in the previous chapter - over-prediction at poorer health states. However, there appears to be some evidence that mapping algorithms based on EQ-5D-5L may yield improved predicted utilities at poorer health states. In particular, the BB model showed improved predictions, regardless of the instrument.

The predictions at poorer health states (Figures 5.3a (5L) & Figure 5.3b (3L)) present some interesting findings. Modelling with the LDVMM consisted of a BB and Normal distribution. Values >0.30 were modelled assuming a BB distribution. Predictions at poorer health states (assumed to be -0549 to 0.30) appear slightly worse. Moreover, better predictions with the LDVM after EQ-5D values >0.3 were observed. This supports a BB algorithm as a plausible model for developing a mapping algorithm for the EQ-5D-5L.

The predicted values are worse for the EQ-5D-3L. About 50% of predicted utilities were over-predictions (higher than the observed value by any amount) with the EQ-5D-5L; for EQ-5D-3L this value was 67% (Figure 5.3b).

a) EQ-5D-5L **b) EQ-5D-3L**

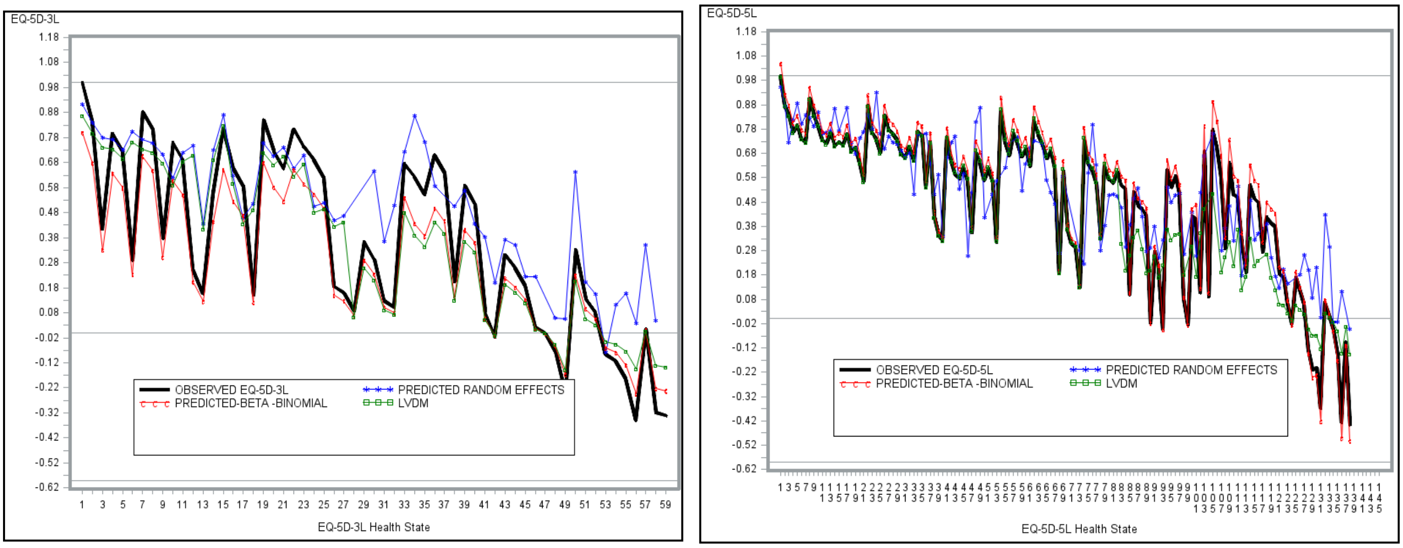


Figure 5.3: Observed vs. Predicted Values by the Health States

Simulation and Cross Validation

Each simulated data set of 985 observations for EQ-5D-5L and EQ-5D-3L were subject to a cross validation using a 50% random sample (about 492 observations each for EQ-5D-5L and EQ-5D-3L, respectively) for the BB model. Hence, a total of 1,000 R^2, RMSE and mean predicted values were observed (Table 5.8 and Figures 5.4 – 5.7). For EQ-5D-5L and EQ-5D-3L, respectively, the average (mean) R^2 from the BB model was 76% (range 51% to 89%) and 68% (range 38% to 79%); RMSEs averaged around 0.099 (range 0.069 to 0.155) and 0.113 (range 0.058 to 0.177). Simulations from the Random Effects and LDVM models showed similar performance but were both inferior compared to the BB.

Algorithm	Parameter	Mean	Lower 95%	Upper 95%	Range
EQ-5D-5L	**R^2**	0.76	0.69	0.82	(0.51, 0.89)
	RMSE	0.099	0.075	0.121	(0.069,0.155)
	Observed	0.572	-0.018	1.00	(-0.436,1.00)
	Predicted	0.575	0.198	0.950	(0, 1)
EQ-5D-3L	**R^2**	0.68	0.58	0.78	(0.38, 0.79)
	RMSE	0.113	0.103	0.120	(0.058, 0.177)
	Observed	0.515	-0.07	1.00	(-0.594, 1.00)
	Predicted	0.518	0.112	0.89	(0, 1)

Table 5.8: Results of Simulation and Cross Validation (BB Model)

a) EQ-5D-5L **b) EQ-5D-3L**

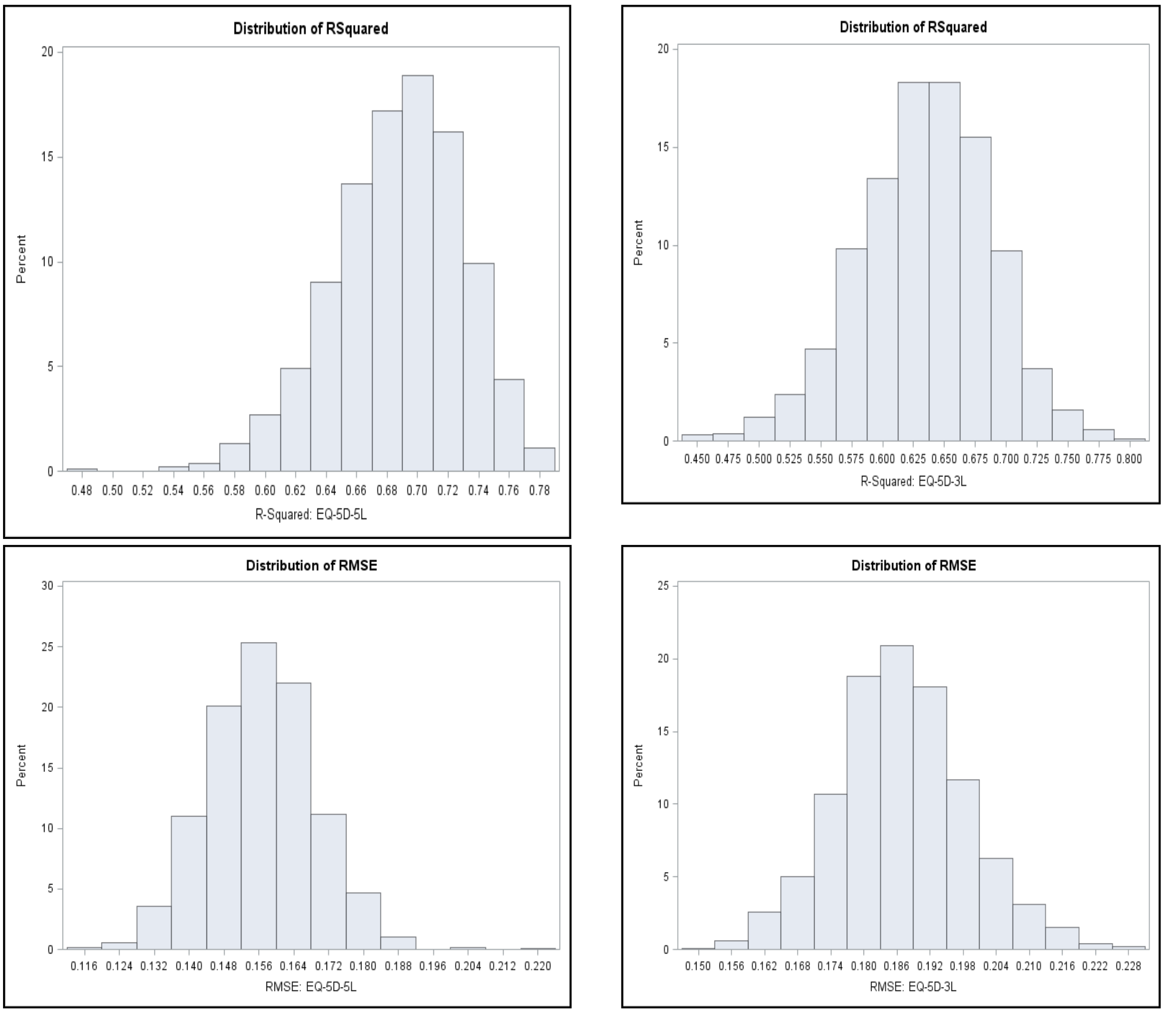
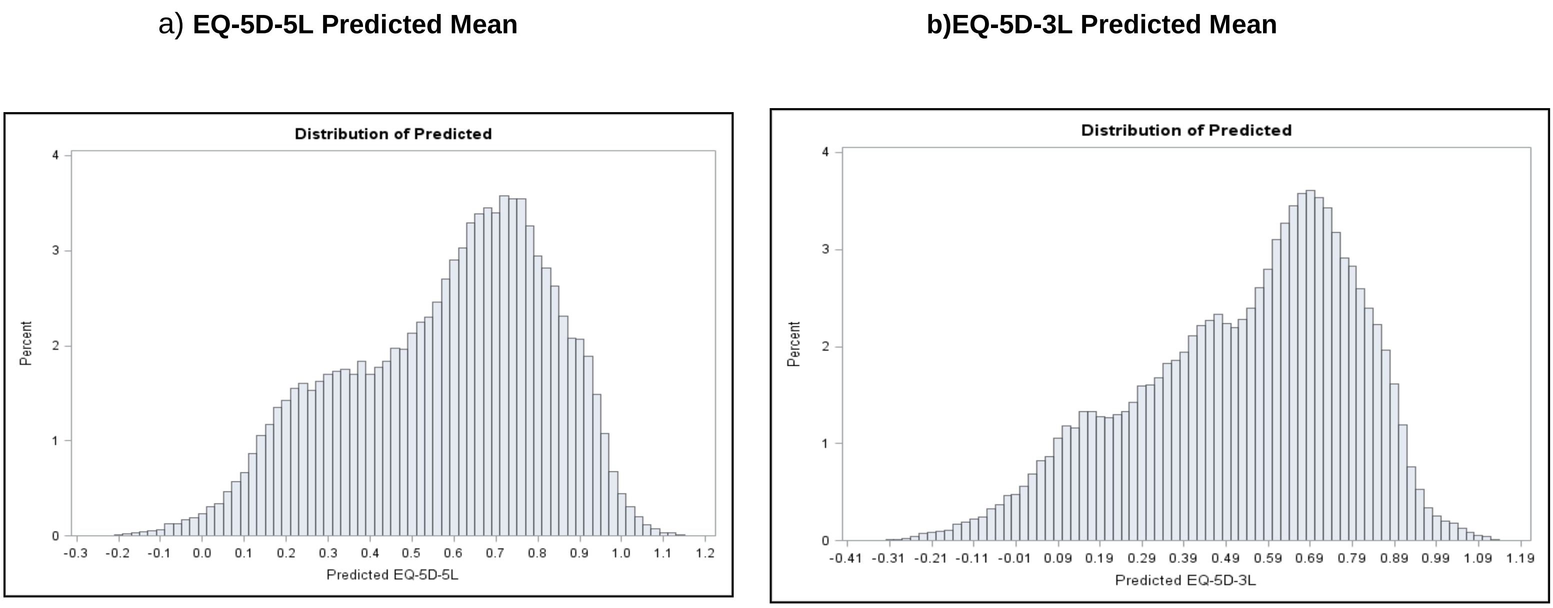


Figure 5.4: Distribution of R^2 and RMSE for Each of (a) EQ-5D-5L and (b) EQ-5D-3L after Cross Validation Models (50% Holdout Sample): Random Effects Model



a) **EQ-5D-5L Predicted Mean**

b)EQ-5D-3L Predicted Mean

Figure 5.5: Distribution of Predicted Means (a) EQ-5D-5L and (b) EQ-5D-3L after Cross Validation Models (50% Holdout Sample): Random Effects Model

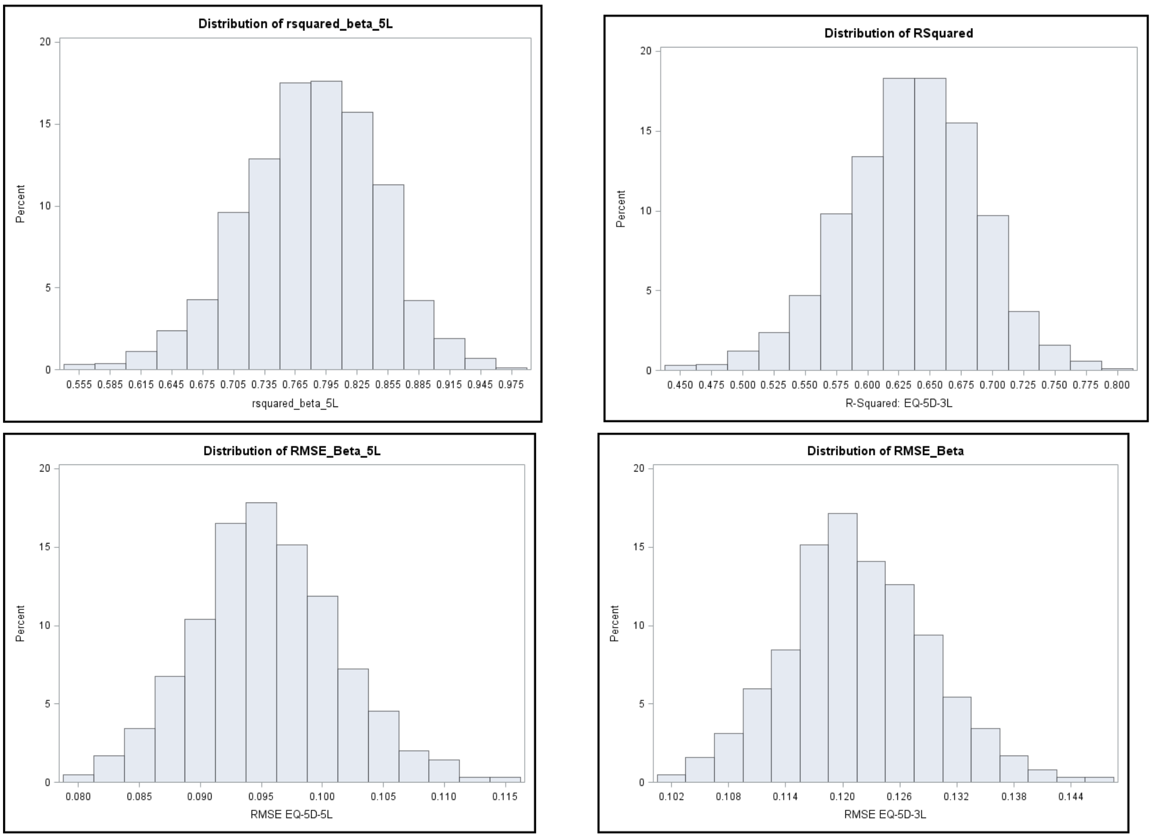


Figure 5.6: Distribution of R^2 and RMSE for Each of (a) EQ-5D-5L and (b) EQ-5D-3L after Cross Validation Models (50% Holdout Sample): BB Model

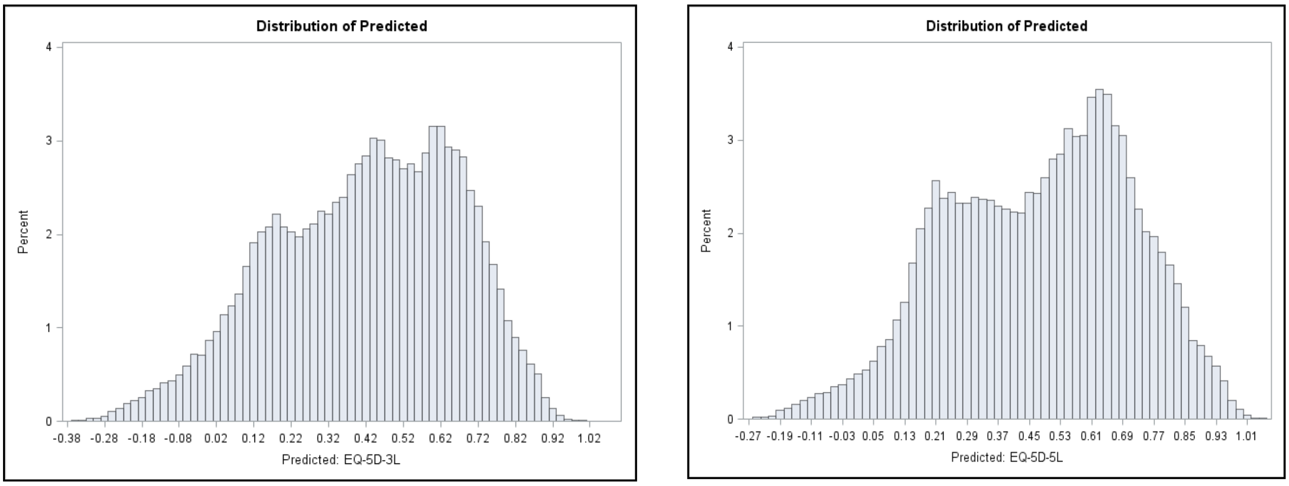
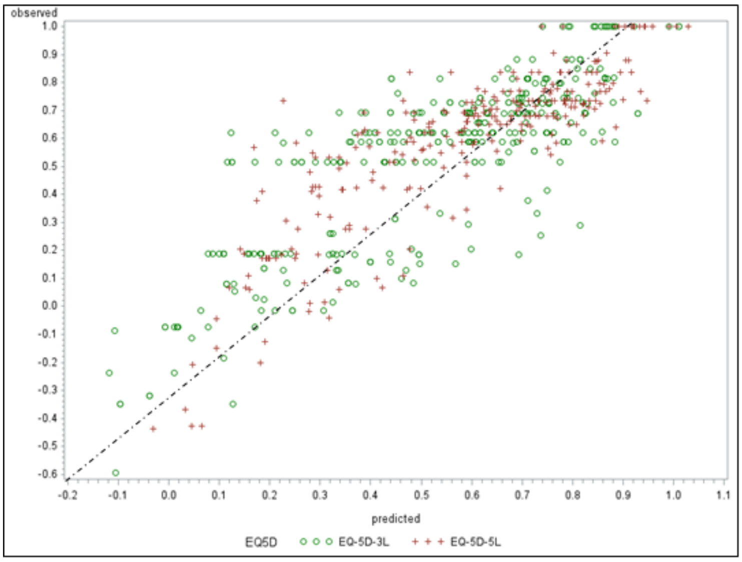


Figure 5.7: Distribution of Predicted for each of (a) EQ-5D-5L and (b) EQ-5D-3L after Cross Validation Models (50% Holdout Sample): LDVMM Model

Predicted mean utilities were closer to the observed for the EQ-5D-5L; 0.572 vs. 0.575, whereas, for the EQ-5D-3L, these were 0.515 vs. 0.518 (Table 5.8 and Figures 5.4 – 5.7). Hence, the sample predictions for the EQ-5D-5L appeared more accurate than those of the EQ-5D-3L, particularly with the BB model. When a different cut-off was used (e.g. 75% to model the data and 25% for prediction), there was no change was identified in the conclusion. A scatter plot of predicted versus observed EQ-5D-%L utilities are shown in Figure 5.8.



Note: At lower utility scores (poorer health), over –predictions is greater with the EQ-5D-3L, as compared to EQ-5D-5L.

Figure 5.8: Scatter Plot of Observed vs. Predicted Values (EQ-5D-5L, EQ-5D-3L) – BB Model

5.4 Discussion

Three mapping algorithms have been developed and compared for the EQ-5D-5L and EQ-5D-3L using contemporary and novel modelling methods. It has been shown that EQ-5D-5L may offer better prediction at poorer health states, where several previous algorithms with EQ-5D-3L have usually over-predicted. Modest improvements of an algorithm based on EQ-5D-5L over one based on EQ-5D-3L in terms of statistical metrics (e.g. R^2, percentage predicted) have been confirmed with a BB model in this and previous analysis [109]. Others have suggested that two-part models may offer a way to predict the different parts of the distribution in the context of mapping with improved performance for handling over-prediction [138]. More recently, the suitability of the BB type models over other models has been confirmed [192,193]. In this analysis, the bimodal nature of the EQ-5D-5L value sets noted earlier [188] (Figures 5.6 and 5.8) have been confirmed.

This appears to be the first time a mapping algorithm has been developed simultaneously from EQ-5D-5L and EQ-5D-3L in the same set of lung cancer patients, using EORTC-QLQ-C30 and compared to each other using data collected from lung cancer patients in a real world NHS setting. Previous works with the EQ-5D-5L highlighted some of the limitations of the EQ-5D-3L, relating to aspects like bimodality of utilities and a lack of sensitivity to detect differences between treatment groups [149,142,194]. Some earlier mapping models did not take this into account, where for instance, an algorithm using the FACT-B in a breast cancer population was reported with R^2 of nearly 48% (AIC was not reported) [195].

In this analysis, over-prediction at poorer health states still exist with EQ-5D-5L, although it is not as marked as EQ-5D-3L. The final value sets (Oppe et al., 2014) [188] were being developed at the time of writing this thesis, which may result in different predictions at poorer health states, compared to the final published ones. The reasons for over-prediction may be due to several factors, including the functional form of the model, the range of the scale (5 point vs. 3 point scale), the number of health states and other clinical characteristics. It was suggested in Chapter 4 that over-estimates at poorer health states may be linked to other factors like poorer prognosis. Preliminary evidence of this is shown by observing the relationship between ECOG performance and EQ-5D utilities (Table 5.9). It is possible that a further complexity is required in the modelling by using the joint distribution of utilities and other outcomes (e.g. adverse events) to model the QLQ-C30 scores.

ECOG	Mean EQ-5D-5L		Mean EQ-5D-3L	
	Observed	Predicted	Observed	Predicted
0	0.706	0.736	0.675	0.702
1	0.625	0.638	0.589	0.600
2	0.502	0.493	0.489	0.437
3	0.317	0.331	0.273	0.284
4	-0.024	0.237	0.067	0.199

Table 5.9: Utilities and ECOG *R*elationship

For the purpose of this thesis, the EQ-5D-5L and 3L assessments were taken in a narrow time window (e.g. 3 weeks apart). Therefore, there may be some concern about 'carry-over' or recall bias. To examine this, it was determined whether the health state responses were similarly recorded. For instance, if a response of 11112 was observed for EQ-5D-3L, it was checked whether this was also observed for EQ-5D-5L (responses >3 are not possible for EQ-5D-3L) or not. It was noted that for 15 out of the 146 (EQ-5D-5L) health states, the responses for EQ-5D-5L and EQ-5D-3L were the same. For instance, patients with responses of 11111 to both EQ-5D-5L and EQ-5D-3L in 18 of the 985 (pairs) of observations (<2%). In the vast majority of cases, the responses were different. This suggests that patients did not recall the previous responses and the presence of carry over may be unlikely.

However, there are several limitations of this research. Firstly, this is a small sample size with relatively few health states. Although the sample size is larger than the algorithm reported by Kontodimopoulous (2009) [131]. Secondly, inferences need to be restricted to a similar cancer population, until further evidence emerges of wider applicability across tumor types. Thirdly, external validity was not possible in an independent data set and therefore cross-validation was used as a 'second best,' accompanied by simulation for out of sample predictions. Fourth, the questionnaires could have been randomized in the order they were given, although as noted from above, the potential for an order type effect was likely to be minimal over a two week period. Finally, the values of the EQ-5D-5L are cross-walked from the EQ-5D-3L and are therefore subject to uncertainty. However, in the absence of a readily identified set of value sets, and given that the EQ-5D-5L is currently being used in clinical research and for economic evaluation in the interim, using the EQ-5D-3L cross-walk sets are considered acceptable.

Despite these limitations, this is the first mapping algorithm for the EQ-5D-5L using real world data with enhanced generalizability outside the RCT context. It is inevitable that further research is required, particularly through the use of exploring covariates and other clinical data in order to improve the mapping.

5.5 Conclusion

Mapping algorithms developed from EQ-5D-5L appear to provide improved estimates of utilities compared to EQ-5D-3L, specifically at poorer health states. Two-part models fit the data well and this result confirms earlier and more recent work. It is recommended that in studies where EQ-5D utilities have not been collected, an EQ-5D-5L mapping algorithm is used.

The next chapter (Chapter 6) will explore the nature of the relationship between toxicity, utility and other clinical data when generating a mapping algorithm from the QLQ-C30. This will assist in determining whether more complex functional forms are needed to improve the utility estimation.

Chapter 6

Chapter 6: Joint Modelling and Covariates to Improve Estimation of EQ-5D Utilities

Abstract

Introduction: In this chapter, the impact of covariates is examined on the over or under prediction of mapping algorithms, specifically at poorer health states. In Chapter 4, it was briefly noted how factors other than condition-specific measures can potentially influence utility estimation (e.g. the relationship between adverse events and mean utility). Although EQ-5D-5L is a more recent instrument, EQ-5D-3L still remains dominant in UK HTA submissions. Therefore, examination of the covariates' influence on prediction from both EQ-5D-3L and EQ-5D-5L are investigated.

Methods: The impact on the mapping with and without covariates from three models are considered: (i) Linear Random Effects Model (RE), which is a common type of algorithm; (ii) The Two-Part Beta-Binomial and (iii) a new joint model that considers a relationship between toxicity and EQ-5D to model condition-specific measures. Finally, an estimate of the utility increment/decrement is reported, following Nafees et al. (2008) [108]. Data from a real world study in NSCLC (Study 3) patients was considered to estimate utility for several important factors (e.g. response, gender, etc.).

Results: Mapping with EQ-5D-5L was better than EQ-5D-3L, irrespective of model or use of covariates. However, it was the joint model that performed the best - R^2, AIC, RMSE and percentage predicted within ±10% of observed were 81%, -4333, 0.069 and 81%, respectively for EQ-5D-5L with covariates (ECOG, Histology, Stage, and Smoking: $p<0.05$). The random effects model was the best fitting model: R^2, AIC, RMSE, and % predicted within ±10% of observed were 67%, -328, 0.177 and 36%, respectively for EQ-5D-5L.

Conclusion: Mapping based on the joint modelling of utility and toxicity in addition to covariates offers improvements in prediction of utilities over both the Random Effects and Beta-Binomial Model.

6.1 Introduction

Following on from Chapter 4, the possible reasons why mapping functions under/over predict, specifically for patients in poorer health states are considered by investigating the joint relationships between outcomes (e.g. EQ-5D) and factors (e.g. age, gender). Over/under prediction occurs when a predicted value differs from the observed utility score by *any* amount, for the purposes of this thesis. There is no formal definition of 'over' or 'under-prediction' in the literature on mapping. Although very 'small' departures might appear negligible, these 'small' differences can have a marked impact on the ICER. Where the numerator of the ICER equation (Chapter 1) is large (large difference in costs), even small differences between observed versus predicted estimates of EQ-5D can influence the ICER.

For example, in Table 6.1, an EQ-5D prediction error of about 2%, results in a similarly small (3%) change in QALY. However, the impact on ICER is marked, leading to an 8% change in the ICER (assume the QALY in the control group is unchanged) which changes a decision for a cost-effectiveness threshold of £30,000/QALY.

	Experimental	**Control**	**Difference**	**ICER**
Cost	100,000	96,000	4,000	
EQ-5D Observed	0.65	0.52	0.13	
QALY Observed	0.35	0.21	0.14	28,571
EQ-5D Predicted	0.66 [1.5%]	0.53 [1.9%]		
QALY Predicted	0.36	0.23	0.13	30,769 [+8%]

Table 6.1: Example Showing How Small Differences between Observed and Predicted EQ-5D Can Lead to Changes in ICER and the Cost-Effectiveness Decision

Where over-prediction occurs, QALYs may be higher (or lower) than expected, which may impact the consequent ICER. This might lead to some cancer treatments to be erroneously declared cost-effective (or vice versa). The example in Table 6.1 shows it is not difficult to see how cost-effectiveness decisions can be altered through over (or under prediction) of EQ-5D utilities. Under-prediction is just as critical, because if utilities are under-predicted in the group receiving the standard of care, the experimental group (new treatment) may unfairly bias against the standard of care (comparator) as noted in Table 6.2. Although the assumption that under/over prediction is similar between groups, it is conceivable that a treatment that is very toxic may result in poorer HRQoL (and therefore health states). If mapping algorithms are reported to over-predict at poorer health states, the degree of the over-prediction may be different for each treatment group.

Standard (S)	Experimental (E)	Potential Impact on QALY	Potential Impact on ICER*
Over-prediction	Over-prediction	Higher QALY both	No expected impact
Over-prediction	Under-prediction	Lower QALY on E or S	ICER biased against E
Under-prediction	Over-prediction	Higher QALY on E or S	ICER biased against S
Under-prediction	Under-prediction	Lower QALY both	No expected impact

Table 6.2: Potential Impact on ICER and QALY (assuming all else is the same between groups)

*For experimental vs. standard

6.2 Sources of Over/Under-Prediction from Mapping Algorithms

Over/Under-prediction (defined earlier as the difference between the observed and estimated utility) from mapping algorithms can happen for several important reasons:

(i) The patient populations are the same: For instance, when a mapping algorithm is developed from a lung cancer population and applied to a target cancer dataset, which is also a very similar population (e.g. similar age, same tumor type, similar ECOG, cancer stage). In this situation, over/under prediction might be due to some other unknown, unmeasured or 'hidden' factors, such as the timing of toxicity or onset of drug action.

(ii) The patient populations are different: The data used to develop the mapping algorithm and the data used for predicting are from different patients. For example, a mapping algorithm developed using breast cancer data may be applied to a prostate cancer data set. In this case, it is not only differences in cancer but also gender that may affect prediction (assuming a group by gender interaction).

(iii) Populations are 'similar' or the same, but differ in baseline characteristics: For instance, patients have lung cancer in both data sets, but consist of poor prognosis patients in one data set (from which the algorithm was developed) and good prognosis in the other data set (where the algorithm is tested); or advanced stage cancer in some and early stage cancer in others – such as differences in ECOG/morbidity, particularly if these change over time (time varying covariates)

(iv) Patient populations are different – they differ in baseline characteristics and differ in tumor types: For example, developing a mapping function from male patients with lung cancer and applying it to a female renal cell carcinoma data set. This is different to (ii) in that lung cancer patients and renal cell carcinoma patients can be male or female.

(v) The two HRQoL instruments are actually measuring different things, even if all other factors are similar. For instance, clinical and demographic characteristics are similar, but one instrument may be measuring a condition specific HRQoL feature and the other a generic HRQoL feature. The lack of overlap may result in poor prediction. The correlation between a generic and highly symptom specific measure may be poor for prediction purposes.

(vi) The inadequacy of the mathematical algorithm: The mathematical form of the model is not flexible enough to handle complex distributional properties like over-dispersion (ceiling effects), skewness and general non-normality, as well as bivariate relationships between multiple responses.

(vii) Different ranges: The range of the two measures is different. For instance, the data used to develop the algorithm is from -0.35 to 0.80 for EQ-5D-3L with 20 to 85 for QLQ-C30; the target data used to apply the algorithm has QLQ-C30 values that range from a wider scale of 0 to 100 leading to extrapolation and greater uncertainty.

(viii) Some important missing information or factors associated with both EQ-5D and QLQ-C30 are not included: For instance, high toxicity may result in poor HRQoL when measured with EQ-5D and with QLQ-C30. In this situation, a patient may have had a severe toxic event resulting in poor EQ-5D and also poor scores on the QLQ-C30. Conversely, a positive treatment effect may result, in good HRQoL and potentially less toxicity.

The above are likely to be the most common reasons, although there may be additional factors that explain the nature of utility prediction. If none of the issues in (i) to (viii) above explain how good/poor the predictions are, then alternative modelling approaches could be considered (e.g. Bayesian Networks, Chapter 7).

One hypothesis considered in this chapter is that mapping algorithms using clinical type measures like toxicity, which are collected in almost all clinical trials, can be used to form a joint relationship with EQ-5D responses to model a condition-specific measure over time. Therefore, a bivariate relationship of the form $(\mathbf{Y_1}, \mathbf{Y_2}) = f(\mathbf{X}\beta)$ is considered, where $\mathbf{X}$ is a matrix of responses from some condition-specific measure (e.g. the 15 scores from QLQ-C30) and β is a vector of coefficients associated with QLQ-C30; $\mathbf{Y_1}$, for instance, could be responses (utilities from EQ-5D), correlated with $\mathbf{Y_2}$, which might be adverse event grades. Moreover, the relationship between utilities and toxicities might be similar across cancer studies because one specific cancer drug may be used for several tumor types. This approach to mapping may address the over-prediction at poorer health states observed in chapter 4. Therefore, it is pertinent to first explore the potential reasons for over/under prediction.

6.3 Influence of Clinical and Demographic Factors on Utility Estimation

In some published algorithms [128], demographic characteristics are used to form part of the mapping algorithm; but these variables can be very specific to trials (e.g. gender is not a relevant factor for modelling in a mapping algorithm for prostate cancer patients); baseline clinical characteristics can also be quite specific to the trial or study in question and may not be generalizable. The use of such characteristics in statistical models of mapping is also referred to as covariate adjustment (where the included variable is called a covariate).

The use of covariate adjustments is not novel. Several published mapping algorithms have incorporated baseline covariates [111]. However, little research has been conducted on the impact of the covariates with two-part models and none exist for joint models. Moreover, the impact of covariates on prediction *at poorer health states* has not been considered.

If covariates are used to develop the mapping model, then ideally values for these factors should be present in the target data (i.e. the data set where the algorithm will be applied to generate the predicted utilities) set if the mapping is to be useful. For instance, if the mapping algorithm includes a coefficient for ECOG, but in the target data set, ECOG is not collected, the value of the algorithm becomes limited. In addition, if ECOG in one data set ranges from 0 to 1, whereas the target data set ranges from 0 to 3, the predictions are uncertain (or impossible to predict) for ECOG 4. Therefore, where covariates are used for developing a mapping algorithm, it may be better that they cover a wider range (e.g. ECOG 0 to 3 or an age range covering elderly patients).

The following covariates, in addition to the 15 domain scores of QLQ-C30, are often collected in (lung) cancer studies [165,166]

- ECOG categorical 0, 1, 2, 3 4.
- Age (as a continuous measure).
- Gender Categorical (Male /Female).
- Stage of Cancer (categorical I, II, III, IV, including sub-stages such as IIa)
- Histology (Adenocarcinoma, large cell, squamous for NSCLC only).
- Smoker status (smoker, ex-smoker, never smoked).

When covariates are modelled, the interpretation of coefficients is that the utility changes by a constant amount (if the model is linear in parameters) across the values the covariate assumes. For instance, the rate of change in utility across different ages may be assumed to be constant for each year a patient gets older. However, in reality, the rate of change in utility may be faster after some specific age or even slower, depending upon the particular disease (e.g. patients with Alzheimer's disease might deteriorate much more rapidly with age). It is possible to incorporate the covariates in a non-linear functional form to account for this e.g. Logarithm of age.

An alternative approach is to report average utilities for specific subgroups of patients or through certain clinical characteristics. Nafees et al. (2008) reported estimates of utility values from the general public for various clinical factors (e.g. a utility for each of ECOG 0, 1 or 3), using a standard gamble (SG) approach [108]. However, these were not from data in NSCLC patients, but from the community (general public) based estimates. Values of utilities are likely to be different between patients and the public and more so for specific clinical characteristics. For instance, a poor prognosis advanced (stage III or higher) NSCLC patient with an ECOG score of 3 may have a much lower utility value compared to a Stage I cancer patient with an ECOG of 0.

It has been acknowledged that estimates of utilities from *patients* belonging to each or combinations of these categories would have been informative, if available [108]. Moreover, using estimates of utility values based on SG methods from the general public are unlikely to yield similar values for patients, who have experienced toxicities while having important clinical characteristics (e.g. ECOG 2, Stage II) that influence preferences for health states. To summarize, if values of utilities are available from patients with this combination of clinical characteristics, then these ought to be used in an economic evaluation.

The data used in this chapter has a similar sample size compared to previously reported research [108], with the important distinction that utilities are estimated in cancer patients in a real world setting and will provide valid estimates of utilities for economic evaluation in cancer (NSCLC) studies. The results from this real world data offer unique estimates of both EQ-5D-5L and EQ-5D-3L utilities across several important demographic and clinical markers. Estimates of utilities will be compared and contrasted with those in a similar way to others [108] using a similar modelling approach.

Therefore, this chapter will compare several models with and without covariates, using data from NSCLC patients. These models include Linear, Non-Linear and Joint models. In addition, utility values for each of the clinical characteristics will also be reported and compared to those already published.

6.4 Methods

6.4.1 Study Design

Study 3 design (Observational study)

Details of the study design were provided in Chapter 3. Local ethics approval was given by the NHS research and ethics committee [(REC) Reference LH/56/2014].

Assessments

Described in Chapter 3

Adverse Events

Adverse events (AEs) and health resource use were collected as and when they occurred. The AEs were graded using National Cancer Institute's (NCI) Common Toxicity Criteria (CTC) Version 2.0 [196] from Grade 1 to Grade 5 (Death). See Chapter 3 for further details.

6.4.2 Modelling Approaches

Two approaches to modelling were considered. Firstly, jointly modelling of Grade 3 and higher adverse events and EQ-5D utilities as a function of the 15 QLQ-C30 domain scores, using a separate Random Effects (RE) model for comparison. Secondly, EQ-5D-3L utilities were modelled using clinical factors to estimate the (mean) utility values for given subsets of clinical and demographic characteristics, following Nafees et al. (2008). Model performance, in all the above modelling approaches, were compared as before using several criteria reported previously, including R^2, AIC, RMSE, % predicted within $\underline{+}5$ and $\underline{+}10\%$.

6.4.2.1 Linear Models with Covariate Adjustments

For any linear mapping model, a response from each patient at each time point Y_{ij} is required, where the subscript i refers to each patient and j to the time point at which the response was elicited. In order to predict a patient-level EQ-5D utility from a given condition-specific measure (e.g. the 15 QLQ-C30 domains), the linear form of mapping algorithms takes the familiar multiple regression forms:

$$Y_{ij} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n + \varepsilon_{ij,} \qquad \textbf{[6.1]}$$

where the ε_{ij} are identically and independently Normally Distributed $N(0,\sigma^2)$. Since each patient is likely to have repeated measures each block of observations (or each cluster) is independent (observations *within* each subject are unlikely to be independent). Consequently, a mixed model approach was used by adding a random effect (Random effects model).

The above approach is an example of how most mapping algorithms are reported and developed, whether linear, non-linear frequentist or Bayesian. The coefficients of these models are assessed for their 'significance', often by considering a p-value. If the p-value is statistically significant (e.g. $p<0.05$), the coefficient is retained in the model and the factor is considered 'useful' or 'important' for the purposes of prediction. The practicality or significance of a factor may also depend on its clinical relevance. For instance, in QLQ-C30, 'Financial' domain (patients have some fear of facing financial difficulties in paying for their cancer treatment) can be less significant if cancer treatment is paid by a country's national health service, despite being a statistically important predictor of the response (EQ-5D utility).

6.4.2.2 Joint Models

A joint model is an alternative approach for investigating over/under prediction. A joint model assumes the following (using an example):

- The utilities (first response, $\mathbf{Y_1}$) and independent factors (QLQ-C30) are related (Figure 6.1).
- The utilities ($\mathbf{Y_1}$) and 2nd response ($\mathbf{Y_2}$**,** where $\mathbf{Y_2}$ represents presence or absence of a grade 3 or higher adverse events) are related. The adverse events can typically be linked to the drug (e.g. Grade 3 or higher neutropenia)
- The adverse events ($\mathbf{Y_2}$) and independent factors (QLQ-C30) are related.

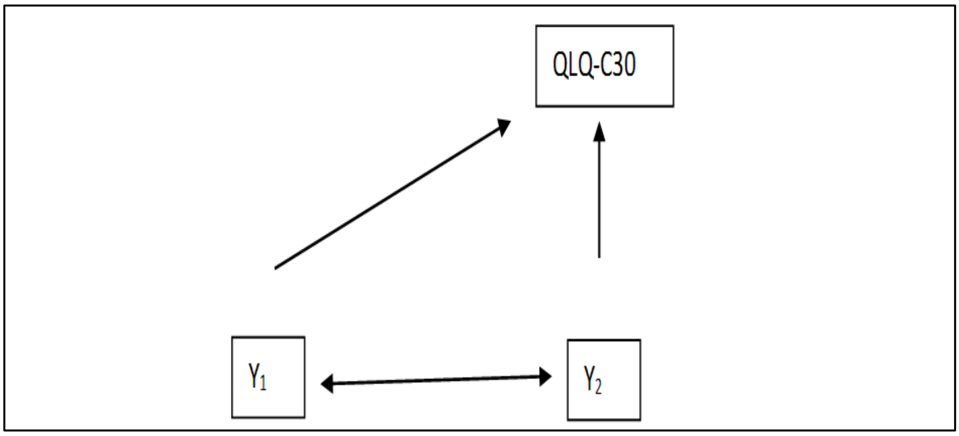


Figure 6.1: Description of Joint Modelling Approach

The heuristic reasoning behind the joint modelling approach is that as adverse events (**Y_2**) become worse, utility (**Y_1**) may also deteriorate. It is hypothesized that modelling the correlation between two (or more) outcomes and factors (QLQ-C30) might improve predictive performance (and smaller standard errors). In Figure 6.1, the arrows move from **Y_1** and **Y_2** to illustrate the (independent) predictive relationship between each of **Y_1** and **Y_2**. However, **Y_1** and **Y_2** themselves are correlated. The benefits of jointly modelling **Y_1** and **Y_2** are that since both **Y_1** and **Y_2** are, in fact, *outcomes* measured after treatment, modelling this correlation is likely to result in better prediction, smaller standard errors of coefficients and improved model fit.

6.4.2.3 Joint Modelling Notation

The joint modelling approach will now be formalized using standard notation [197,198]. Assume that adverse event is a binary response (**Y_2**), such that any AE grade $\geq$ 3 is categorized as 1, otherwise 0. A consequence of **Y_2** is its impact on **Y_1**, the EQ-5D utility. Variables that are assumed to impact **$Y = (Y_1, Y_2)$** are the 15 QLQ-C30 domain scores (explanatory variables) and other potential baseline covariates, **Z**. If **Z** consists of the 15 QLQ-C30 domain scores and also additional baseline covariate, age, then **Z** will consist of the 15 coefficients of the 15 QLQ-C30 scores and additional coefficients such as covariates.

One may model each **Y_k** separately by structuring the mean, E[**Y_k**|**Z_k**] and var[**Y_k**|**Z_k**], k=1,2. The values of k refer to the response outcomes (AE for k=1 and utility for k=2). For instance, if modelled separately, this would be:

$$Y_1 = f(Z_1)^{k=1} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_{15} X_{15} \qquad \textbf{[6.2]}$$

$$\text{and } Y_2 = f(Z_2)^{k=2} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_{15} X_{15} \qquad \textbf{[6.3]}$$

The covariates $\mathbf{Z_1, Z_2}$ need not be the same. Although, in this instance, they will be assumed to influence $\mathbf{Y_1}$ and $\mathbf{Y_2}$ in equal importance. A parameter $\boldsymbol{\theta}$ is introduced to link the two outcomes (utility and AE) through a shared random effect when estimating the expected value of **Y**, i.e. E[$\mathbf{Y}|\mathbf{Z}$, θ] or by structuring the covariance matrix var ($\mathbf{Y}|\mathbf{Z}$).

The joint effect of $\mathbf{Y_1}$ (utilities) and presence or absence of a grade 3 adverse event presence, has the joint distribution:

$$f(\mathbf{Y_1},\mathbf{Y_2}|\mathbf{Z}) = f(\mathbf{Y_1}|\mathbf{Z})\, f(\mathbf{Y_2}|\mathbf{Y_1},\mathbf{Z}) \qquad \textbf{[6.4]}$$

The generalized linear model $g_k(E(Y_{ik}|Z_i)) = Z'_{ik}\beta_k$, k=1,2 is the framework for undertaking the modelling. The function g_k is a link function used to model the specific distribution of each outcome: g_1 models utilities and assume these as normally distributed (for simplicity) with a link function $g_1(u) = u$ (the identity link); g_2 models the adverse event where

$$g_2(u) = \phi^{-1}(u) \text{ a probit link.}$$

The procedure GLIMMIX in SAS® is used to estimate the two marginal models jointly.

6.4.2.4 Two-Part Models

In certain situations, the outcome of interest has a large number of zero or one outcomes and a group of non-zero outcomes that are discrete or highly skewed, such as in health care costs, where some patients have zero costs or the distribution of positive costs are often extremely skewed. An example of such a model was shown using data in chapter 4. The results of chapter 4 are already published [109], where a zero-one (three-part) model was used. It is called a 3-part model, because of the need to model the over-dispersed zeros (i.e. lots of zero's), modelling the over-dispersed ones and then modelling the values between 0 and 1. It has been suggested that two-part models may offer a way to predict the different parts of the distribution in the context of mapping with improved performance for handling over-prediction [138]. This was shown to be the case with the use of the BB model in Chapter 4.

One limitation of the BB model used in Chapter 4 was the restriction of negative values to zero (because there were so few), which necessitated a 3-part model. For this chapter, a transformation is carried out so that a two-part model can be used (i.e. overdispersion of values around 1 or 0). Hence, the two-part model in this chapter will be similar to the one used in Chapter 4 and a non-linear BB model will be used (rather than a TOBIT type two-part model [101]).

Therefore, the following set of models will be considered (Table 6.3)

Model	QLQ-C30	Covariates*	Structure	Model Details
1	Yes	No	Linear	Random Effects
2	Yes	Yes	Linear	Random Effects
3	Yes	No	Non-Linear	Beta-Binomial
4	Yes	Yes	Non-Linear	Beta-Binomial
5	Yes	No	Joint	Two responses
6	Yes	Yes	Joint	Two responses

*Covariates: Age, gender, ECOG, Stage, histology, smoking status

Table 6.3: Summary of Models Compared

6.4.3 Utility Estimates for Various Clinical Factors

Using data from Study 3, EQ-5D-3L and EQ-5D-5L utilities are modelled, including several demographic and clinical factors. A Random Effects (RE) model was used previously [108,199]. The intercept was considered as the overall mean and utility increments/decrements from this intercept were used to provide estimates of utilities for each clinical or demographic characteristic. Therefore, a similar approach is used in order to facilitate comparisons with previously reported results.

6.5 Results

The details of demographics from Study 3 were provided in Chapter 5. The relationship between EQ-5D, ECOG and Toxicity are shown in Tables 6.4 and 6.5. A clear relationship between EQ-5D and toxicity is shown where EQ-5D HRQoL becomes worse as the AE severity increases. It appears to be this relationship at the patient-level, which has been exploited in an attempt to improve model performance. Hence, the proposed model appears plausible, because utility worsens with worsening toxicity.

ECOG	Mean EQ-5D-5L		Mean EQ-5D-3L	
	Observed	Predicted	Observed	Predicted
0	0.706	0.736	0.675	0.702
1	0.625	0.638	0.589	0.600
2	0.502	0.493	0.489	0.437
3	0.317	0.331	0.273	0.284
4	-0.024	0.237	0.067	0.199

Table 6.4 ECOG in Relation to EQ-5D Utility

AE Grade	Mean EQ-5D-5L	Mean EQ-5D-3L
0	0.52	0.56
1	0.52	0.55
2	0.54	0.53
3	0.49	0.46
4	0.31	0.39

Table 6.5 Relationship between AE and Utility

Adverse Events

There were a total of 56/98 (57%) of patients who reported grade 3 or higher AEs; all patients reported at least one AE of any grade; 20%, 12%, 10% and 36% and 14%, respectively reported Grade 0 to Grade 1 AEs (Table 6.6). The majority of the Grade 3 and above toxicities appeared to be related to routine treatment or the underlying disease and consisted of weight loss (19%), Dyspnoea (10%), Pain (14%) and Chest Infection (8%). The maximum grade was used if the AE occurred more than once.

AE Grade	N=98 (n,%)
0	20 (20%)
1	12 (12%)
2	10 (10%)
3	36 (36%)
4	14 (14%)
5	6 (6%)
Grade 0-2	42 (43%)
Grade ≥3	56 (57%)
Five Most Common Grade ≥3 AEs	
Weight Loss	19 (19%)
Dyspnoea	16 (16%)
Pain	10 (19%)
Chest Infection	14 (14%)

Table 6.6: Toxicities by Maximal Grade

Model Performance

Tables 6.7 and 6.8 illustrate the results of model comparisons. Amongst all models, for both EQ-5D-3L and EQ-5D-5L, the Joint modelling approach based on R^2, and AIC indicated the 'best' model fit, either with or without covariates. Without covariates, the RE, BB, and Joint models reported R^2 of 67%, 69%, and 80% respectively, with AIC's of -291, -385 and -3701 for EQ-5D-3L (Table 6.7, Table 6.8). For EQ-5D-5L, these were 72%, 75% and 85% for R^2; and AICs were -365, -475 and -3757, respectively. Also, adjusted R^2 were slightly lower but similar. Moreover, in the presence of covariates, ten QLQ-C30 predictors were statistically significant compared to only four for both BB and RE models when modelling EQ-5D-3L utilities. For EQ-5D-5L, a similar pattern emerged (Table 6.7, 6.8 and 6.9), with improvements to the BB model. Therefore, modelling the correlation between toxicity and utility appears to have contributed towards a more powerful statistical test of rejecting H_0:β_i=0 in favor of H_1:β_i≠0. This hypothesis relates to the coefficients or predictors of the utility. That is, at least one of these predictors is non zero (β_i≠0) and influences the prediction of the EQ-5D utility.

	1: RE	2: RE Cov	3: BB	4: BB Cov	5: Joint	6: Joint Cov
Intercept	0.08046	0.5118	-0.0123	0.1448	**0.141174**$^{\#}$	**-0.159258**$^{\#}$
Physical Functioning	**0.005437***	**0.3598**$^{\#}$	0.01918	**1.8351**$^{\#}$	**0.005358**$^{\#}$	**0.005165**$^{\#}$
Role Functioning	**0.001392***	0.02073	0.00421	0.4006	**-0.001852**$^{\#}$	**-0.003122**$^{\#}$
Emotional Functioning	**0.001949***	0.1294	0.00661	**0.6921**$^{\#}$	**0.003562**$^{\#}$	**0.004074**$^{\#}$
Cognitive Functioning	-0.00073	-0.06618	-0.00425	**-0.4856**$^{\#}$	**-0.001235**$^{\#}$	**-0.001533**$^{\#}$
Social Functioning	0.000516	0.04849	-0.00035	0.07463	0.000182	0.000293
Global health status / QoL	-0.00043	**-0.08086**$^{\#}$	-0.00197	-0.3012	**0.002020**$^{\#}$	**0.002501**$^{\#}$
Fatigue	0.000993	0.06067	0.00443	0.4700	0.000257	0.000338
Nausea / Vomiting	0.000276	-0.05768	-0.00146	0.02201	**0.001152**$^{\#}$	**0.000408**$^{\#}$
Pain	**-0.00219**	**-0.2327**$^{\#}$	-0.03278	**-1.0981**$^{\#}$	-0.001556	**-0.002496**$^{\#}$
Dyspnoea	-0.00011	-0.02561	0.00015	-0.08724	-0.000092679	0.000117
Insomnia	-0.00004	-0.00274	0.00193	0.05503	-0.000470	-0.000389
Appetite loss	0.000341	0.009468	0.0002	-0.05375	0.000791	0.000706
Constipation	0.000524	0.03047	0.0014	0.1260	-0.002282	**-0.001911**$^{\#}$
Diarrhea	0.000499	0.03633	0.00393	0.2142	**0.001098**$^{\#}$	**0.000622**$^{\#}$
Financial Problems	-0.00004	**0.04696**$^{\#}$	-0.00113	0.1526	**0.141174**$^{\#}$	**0.000395**$^{\#}$
Model Statistics						
Predicted Mean* (SD)	**0.523 (0.252)**	**0.522 (0.201)**	**0.518 (0.183)**	**0.517 (0.199)**	**0.519 (0.112)**	**0.513 (0.099)**
Observed Mean (SD)	0.515 (0.308)	**0.515 (0.308)**	**0.515 (0.308)**	**0.515 (0.308)**	0.515 (0.308)	0.515 (0.308)
R^{2}	**67%**	**74%**	**69%**	**78%**	**80%**	**81%**
RMSE	**0.183**	**0.177**	**0.113**	**0.101**	**0.079**	**0.069**
% predicted +5%	**19%**	**21%**	**21%**	**43%**	**67%**	**73%**
% predicted +10%	**37%**	**39%**	**47%**	**57%**	**77%**	**81%**
AIC (smaller better)	**-291**	**-328**	**-385**	**-396**	**-3701**	**-4333**
Covariates of significance*	**N/A**	**ECOG, Histology**	**N/A**	**ECOG, Histology Stage Smoking**	**N/A**	**ECOG Histology Stage Smoking**

*predicted from model

[1] RE: Random Effects (no covariates); 2: RE (covariates); 3: BB (no covariates); 4: BB (covariates); 5: Joint (no covariates); 6: Joint (covariates); ~ Mean Absolute Error (calculated as 1/n Σ |(predicted – observed)|; Statistically significant at two-sided 5% level; *statistically significant covariates at 2 sided level 5% level; $^{+}$computed by modelling observed vs. predicted

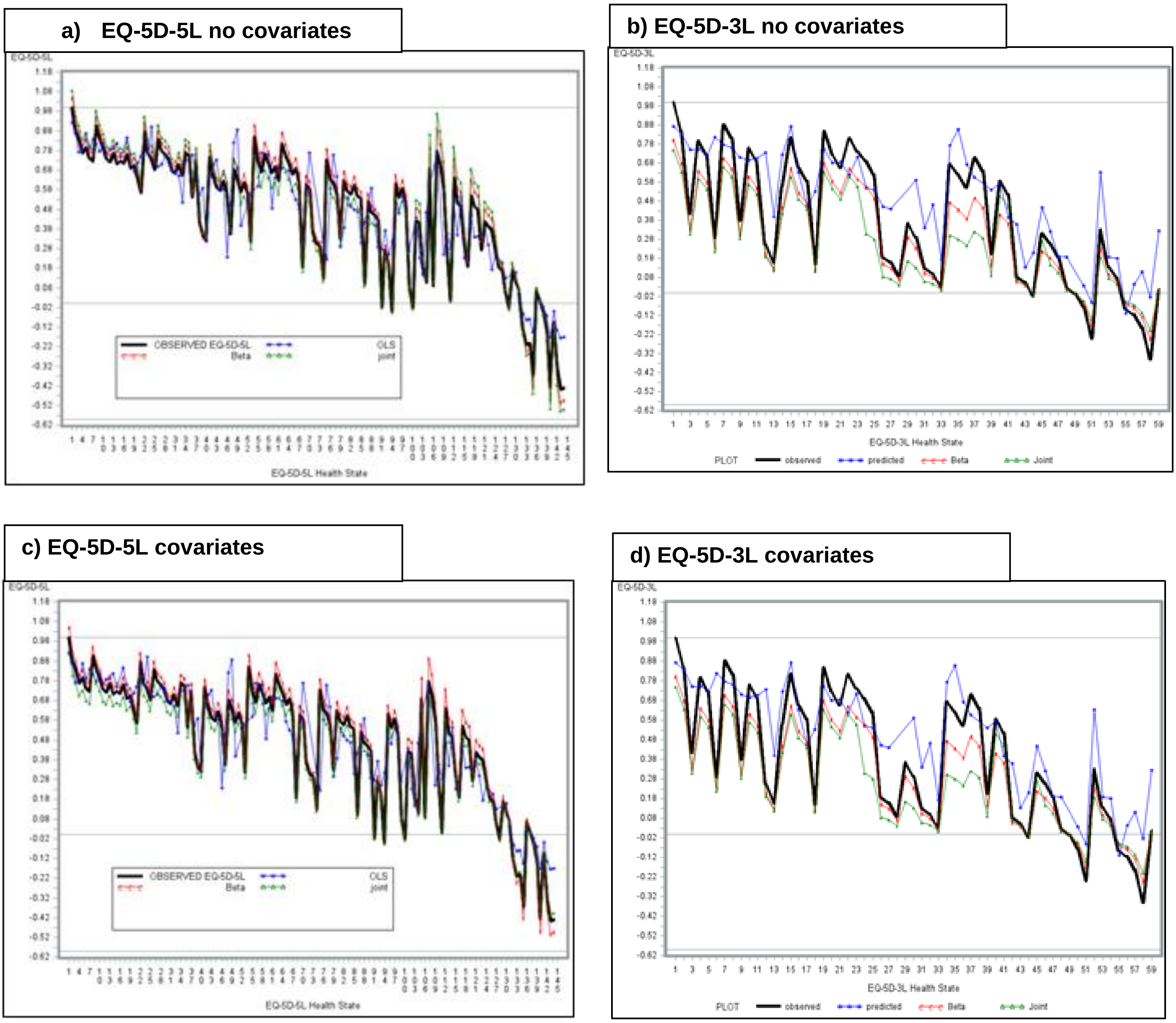
Table 6.7: Summary of Results Comparing Joint Modelling Approach – EQ-5D-3L

	1: RE	**2: RE Cov**	**3: BB**	**4: BB Cov**	**5: Joint**	**6: Joint Cov**
Intercept	0.2255	0.41878	-1.51144	-0.5679	**0.13512**#	**-0.16668**#
Physical Functioning	**0.006718***	**0.41208**#	**0.03644***	**2.3147**#	**0.00623**#	**0.003943**#
Role Functioning	-0.00032	**0.06898**#	**0.009619***	0.2258	**-0.00199**#	**-0.002922**#
Emotional Functioning	**0.001871***	**0.12975**#	**0.01904***	**0.6208**#	**0.001233**#	**0.003774**#
Cognitive Functioning	-0.00057	-0.03798	-0.00633	-0.1863	**-0.00023**#	**-0.00200**#
Social Functioning	0.000387	0.034609	-0.00013	**0.3489**#	0.000211	0.000493
Global health status / QoL	**-0.00109***	-0.02527	0.001652	-0.1560	**0.00432**#	**0.001999**#
Fatigue	0.000324	**0.09045**#	0.003561	**0.7556**#	0.000199	0.000229
Nausea / Vomiting	-0.00041	-0.00175	0.000452	**0.4261**#	**0.00221**#	**0.000511**#
Pain	**-0.00290***	**-0.13667**#	**-0.03569***	**-1.0404**#	-0.000888	**-0.001999**#
Dyspnoea	0.000368	-0.00408	**-0.00806***	-0.04231	-0.000123	0.000161
Insomnia	-0.00017	0.001200	0.002047	-0.07279	-0.000460	-0.000334
Appetite loss	-0.00030	0.021845	**0.005383***	0.1075	0.000823	0.000834
Constipation	-0.00013	0.020708	0.000454	**0.2775**#	-0.001992	**-0.001884**#
Diarrhoea	**0.001155***	0.014075	0.000353	0.1022	**0.009398**#	**0.0005975**#
Financial Problems	0.000345	-0.002281	-0.00432	0.02042	**0.19394**#	**0.0004221**#
Model Statistics						
Predicted Mean* (SD)	0.577 (0.241)	0.576 (0.299)	0.575 (0.211)	0.573 (0.223)	0.571 (0.113)	0.571 (0.100)
Observed Mean (SD)	0.572 (0.211)	0.572 (0.211)	0.572 (0.224)	0.572 (0.211)	0.572 (0.211)	0.572 (0.211)
R^2	72%	73%	75%	77%	85%	87%
RMSE	0.152	0.149	0.092	0.087	0.059	0.053
% predicted ±5%	19%	31%	29%	56%	63%	69%
% predicted ±10%	38%	44%	59%	72%	81%	84%
AIC (smaller better)	-365.3	-377.4	-485.3	-486.1	-3757	-4432
Covariates of significance#	N/A	Histology	N/A	ECOG, Histology, Stage	N/A	ECOG Histology Stage Smoking

*predicted from model; [1] RE (no covariates); 2: RE (covariates); 3: BB (no covariates); 4: BB (covariates); 5: Joint (no covariates); 6: Joint (covariates); ~ Mean Absolute Error (calculated as 1/n Σ |(predicted – observed)|; #Statistically significant at two sided 5% level; #statistically significant at 2 sided level 5% level; +computed by modelling observed vs. predicted

Table 6.8: Summary of Results Comparing Joint Modelling Approach – EQ-5D-5L

The predictions at poorer health states were also improved for the Joint model (Figure 6.2). The results were slightly better for EQ-5D-5L in comparison to EQ-5D-3L. The estimates of mean EQ-5D predictions were slightly improved with covariates - the percentage predicted within + 5% and +10% were higher when models included covariates (Tables 6.4 and 6.5). Hence, there appears to be a combination of a wider scale for EQ-5D-5L and information from covariates and toxicity data, which inform expected utility behaviour. Results from the cross-validation simulation (Table 6.9) suggesting the Joint models appear to be a more useful mapping algorithm with less uncertainty (slightly shorter confidence intervals for parameters).


a) EQ-5D-5L no covariates

b) EQ-5D-3L no covariates

c) EQ-5D-5L covariates

d) EQ-5D-3L covariates

Figure 6.2: Comparing by Health States (a,b) without and (c,d) with Covariates for Each Model (RE, BB, Joint)

Model	Algorithm	Parameter	Mean	Lower 5%	Upper 95%	Range
RE	EQ-5D-5L	R^2	0.69	0.61	0.75	(0.43, 0.79)
		RMSE	0.116	0.138	0.177	(0.113,0.218)
		Observed	0.572	-0.018	1.00	(-0.436,1.00)
		Predicted	0.573	0.0037	0.924	(-0.305, 1.205)
	EQ-5D-3L	R^2	0.63	0.54	0.72	(0.45, 0.80)
		RMSE	0.187	0.167	0.207	(0.14, 0.23)
		Observed	0.515	-0.07	1.00	(-0.594, 1.00)
		Predicted	0.582	0.05	0.88	(-0.41, 1.19)
BB	EQ-5D-5L	R^2	0.76	0.69	0.82	(0.51, 0.89)
		RMSE	0.099	0.075	0.121	(0.069,0.155)
		Observed	0.572	-0.018	1.00	(-0.436,1.00)
		Predicted	0.575	0.198	0.950	(0, 1)
	EQ-5D-3L	R^2	0.68	0.58	0.78	(0.38, 0.79)
		RMSE	0.113	0.103	0.120	(0.058, 0.177)
		Observed	0.515	-0.07	1.00	(-0.594, 1.00)
		Predicted	0.518	0.112	0.89	(0, 1)
Joint	EQ-5D-5L	R^2	0.85	0.63	0.96	(0.58, 0.99)
		RMSE	0.058	0.051	0.069	(0.049, 0.072)
		Observed	0.572	-0.018	1.00	(-0.436,1.00)
		Predicted	0.570	-0.0161	0.998	(-0.436, 1.00)
	EQ-5D-3L	R^2	0.81	0.61	0.94	(0.572, 0.975)
		RMSE	0.072	0.063	0.089	(0.059, 0.101)
		Observed	0.515	-0.07	1.00	(-0.594, 1.00)
		Predicted	0.512	-0.059	0.99	(-0.594, 1.00)

Table 6.9: Results of Simulation and Cross Validation

Example of an application of the Joint Model

An example of how the joint model algorithm is applied is shown for the EQ-5D-3L:
For a joint model, two sets of equations are needed. One set through modelling the probability of a grade 3 or higher adverse event with the 15 QLQ-C30 variables and one set through modelling EQ-5D. Table 6.10 shows the coefficients for the Binary and Linear predictors.

	Linear (EQ-5D)	**Binary (AE)**
Intercept	0.08046	0.5118
Physical Functioning	**0.005437***	**0.3598**#
Role Functioning	**0.001392***	0.02073
Emotional Functioning	**0.001949***	0.1294
Cognitive Functioning	-0.00073	-0.06618
Social Functioning	0.000516	0.04849
Global health status / QoL	-0.00043	**-0.08086**#
Fatigue	0.000993	0.06067
Nausea / Vomiting	0.000276	-0.05768
Pain	**-0.00219**	**-0.2327**#
Dyspnoea	-0.00011	-0.02561
Insomnia	-0.00004	-0.00274
Appetite loss	0.000341	0.009468
Constipation	0.000524	0.03047
Diarrhea	0.000499	0.03633
Financial Problems	-0.00004	**0.04696**#
Scale σ_2	0.7051	
ρ_{12}	-0.623	

Table 6.10: Coefficients from linear and logit parts of the model for prediction

From Table 6.10, as noted previously the expected EQ-5D utility increases with some factors (e.g. Physical function) while decreasing with pain. Similarly, the incidence of grade 3 or higher adverse events in the Binary column increases for factors like Pain (a negative coefficient) and the risk of an adverse event falls with improving physical function or symptoms.

The expected value of EQ-5D after taking into account the 3-way relationship between toxicity, AE, and QLQ-C30, is determined by:

$$E[Y_2|Y_1, \text{QLQ-C30}] = Z_2\beta_2 + \alpha Y_1 + 1/2\ \sigma^2_2 \text{ x } \{[Z_1\beta_1 + \rho_{12}\ \sigma_2]/Z_2\beta_1\} \quad \textbf{[6.5]}$$

where:

$Z_2\beta_2$ are the 16 coefficients (including the intercept) relating he EQ-5D-3L with the QLQ-C30,

Y_1 is the binary outcome related to the presence or absence of a Grade 3 AE

$Z_1\beta_1$ are the 16 coefficients (including the intercept) for the relationship between the probability of AE and QLQ-C30

α is a parameter such that the ratio of the sample means for Y_2 when (Y_1=1), and Y_2 when (Y_1=0), is e^{α} -1

σ_2 is a scale parameter estimated from the modelling

ρ_{12} is the correlation between grade 3 AE and EQ-5D

In equation **[6.5],** the EQ-5D is first estimated when Y_1 (Grade 3 AE outcome is set to 1) and then repeated when Y_1 =0 (no AE). Then the EQ-5D is estimated by computed an average (mean) over all patients in the sample.

Hence, the estimated mean EQ-5D-3L is determined as follows where Y_1=1:

[(0.08046 +0.005437*PF +0.001392*RF +………-0.00004*FI) +

0.8*1 + ½*0.7051)] x

{((0.5118 + 0.3598*PF + 0.02073*RF +…….+0.0469*FI) + (-0.623) x (0.7051))/
(0.5118 + 0.3598*PF + 0.02073*RF +…….+0.0469*FI))}

=0.413

In equation **[6.5],** when covariates are added, the mean EQ-5D-3L utility is 0.413 when AEs are present and 0.613 without AEs, resulting in an overall mean of (0.413+0.613)/2 = 0.513

Relationship between Utility and Covariates

Tables 6.10 and 6.11 show the utility decrements associated with covariates and AE grades. Following Nafees et al. (2008), Lloyd et al. (2006) and Sturza (2010) [108,159,199], the estimates of EQ-5D-3L utilities from Study 3 slightly vary from those reported earlier [69,138,168]. Model-based least squares mean estimates yield EQ-5D-3L mean utilities of 0.5882 (Study 3) vs. 0.653 (Nafees et al., 2008); and 0.5882 vs. 0.673 (Lloyd et al., 2006). The differences in these results may reflect differences in populations between those in Nafees et al. (2008), Lloyd et al. (2006) and Study 3.

Nafees et al. (2008) reported estimates based on community approximations (and not actual patients), as were those in Lloyd et al. (2006). Some differences may be due to the fact that Study 3 is a real-world study in NSCLC patients. Only EQ-5D-3L was modelled (response) against covariates because EQ-5D-5L data were not available at the time. The responses were modelled using a multivariate regression model. There is a clear presence of multicollinearity (Variance Inflation Factor =1.9) because patients with comorbidities also presented with adverse events (e.g. Hemoptysis, the coefficient of -0.1825) are likely to occur with Cough (-0.0706). The multicollinearity does not impact the direction of the effect (i.e. EQ-5D worsens in the presence of the AEs). However, the magnitude of the effect is notable. For instance, Hemoptysis results in nearly three times a greater decrement in utility compared to having a cough. Both of these symptoms are simultaneously present in late-stage NSCLC patients.

Table 6.10 shows the utility decrements (Study 3) resulting from patients who experience adverse events (Grade 1 or higher). For instance, patients with rash are considered to have a utility decrement by 0.146 (value of -0.146). The AEs with the largest impact on utilities were Pleural Effusion (-0.398), Dysphagia (-0.331), Decubitis (-0.327) and Vomiting (-0.311). The AEs with the least impact were Alopecia (-0.071), Double Vision (-0.071), Fracture (-0.071) and High Platelet Count (-0.076). Interestingly, QLQ-C30 also suggested Alopecia had a minimal impact on HRQoL.

Adverse Event	EQ-5D-3L
Pneumonia	-0.162636
Rash	**-0.146111***
Vomiting	**-0.311000***
ALP	-0.086734
Alopecia	**-0.071000***
Back Pain	-0.003421
Cough	**-0.070600***
Creatinine	-0.086787
Cystitis	-0.294000
Decubitus	-0.327000
Desquamation	-0.094521
Dizziness	-0.098738
Double vision	-0.071000
Dry Skin	**-0.101831***
Dysarthria	-0.294000
Dysgeusia	-0.294000
Dysphagia	-0.330500
Fracture	-0.071000
Haemoptysis	**-0.182500***
Hoarseness	-0.234500
Hot flushes	-0.294000
Hot sweats	-0.071000
Hypoalbuminemia	-0.110666
Infection, normal ANC	-0.140818
Leucocytes	-0.082675
Nail changes	-0.294000
Neurology other Blackout	-0.294000

Ocular other hemianopia	-0.294000
Platelets	-0.076259
Pleural effusion	-0.398000
Skin Changes	-0.132186
Hypoalbuminemia	-0.294000

*Statistically significant at 5% level

Table 6.11: Mean EQ-5D-3L Utilities Decrement for Each Adverse Event

With respect to the interpretation of Tables 6.10 & 6.11, one consideration is that utility was measured every month, whereas AEs are reported at the time of occurrence (irrespective of when EQ-5D is collected). Since the precise time of when an AE occurs is difficult to predict, the assessment of EQ-5D is not necessarily close to the window of assessment (this is true for most studies or trials). Some AEs tend to last for a longer period of time or increase in intensity (e.g. pain and dyspnoea). Therefore, some utility measurements would be taken during and some after the start of an AE. Consequently, the results presented can be interpreted as depicting an average impact (decrement) on EQ-5D utility.

The mean utility for patients who had a complete or partial response was estimated as 0.6532 + 0.0193 = 0.6725 (extracted from Table 6.11 of Nafees et al., 2008); Lloyd et al. reports these as 0.673. In Study 3, the utility in those patients who had a response was: 0.5882 + 0.0539 = 0.642 (complete or partial response). However, for progressive disease (PD), this was 0.6532 – 0.1798 = 0.4734 (Nafees et al., 2008); for Lloyd et al. (2006) this was: 0.473 and for Study 3 this was: 0.5882 – 0.133 = 0.4552 (Table 6.11). Similarly, for patients who experienced fatigue, the estimate is about: 0.6532-0.07346 = 0.579 (Table 6.11); compare this with 0.599 (Complete Response) or 0.580 (Partial response), depending upon whether the tumor response was a complete or partial response or stable disease (SD). These utility values are well within the range of earlier reported utility estimates [200, 159], despite differences in methodological approaches to obtain the utilities.

	EQ-5D-3L Study 3	**Nafees et al. (2008)**	**Lloyd et al. (2006)[#]**
Intercept (Mean)	0.5882	0.6532	NR
Gender			
Male		NR	NR
Female	-0.0256	NR	NR
ECOG			
0	0.1847	NR	NR
1	0.1474	NR	NR
2	0.1024	NR	NR
3	-0.0010	NR	NR
4	-0.0145	NR	NR
Stage			
I	0.0613	NR	NR
II	0.0492	NR	NR
III	-0.0632	NR	NR
IV	-0.0614	NR	NR
Histology			
Squamous	-0.0009	NR	NR
Non-Squamous	0.0170	NR	NR
Smoking Status		NR	NR
Current Smoker	-0.0192	NR	NR
Ex-Smoker	0.0155	NR	NR
Never Smoked	0.0182	NR	NR
Any Adverse Event* (any grade)			
Grade 3+ AE	-0.0470	NR	NR
Grade 3+ Serious AE	-0.0519	NR	NR
Tumour Response			
CR/PR	0.0539 (CR/PR)	0.0193 (CR/PR)	0.673 (CR/PR)
Stable Disease	0.0012	NR	0.653
Progressive Disease	-0.133	-0.1798	0.473
Type of Grade 3+ SAE			
Rash	-0.2461	-0.03248	0.640
Neutropenia	-0.0718	-0.08973	0.582
Nausea Vomiting	-0.3114	-0.04802	0.624
Fatigue	-0.1235	-0.07346	0.599
Hair Loss	-0.0012	-0.04495	0.628
Diarrhoea	-0.1043	-0.0468	0.623

*Maximum Grades for AEs; # reported as combinations of patients who respond

NR: Not Reported

Table 6.12: Mean EQ-5D-3L Utilities Increments / Decrements for Each Covariate/ Factor

6.6 Discussion

In this chapter, data from a prospectively designed real-world observational study were collected for investigating the impact of modelling covariates and other factors on estimating utilities. It was indicated that a model, which incorporates covariates and toxicity has an important impact on estimating utilities. The joint model, in combination with the enhanced EQ-5D-5L scale, may offer an underlying explanation for why existing mapping models have performed poorly (in similar patients). This model would be a marked improvement over two-part non-linear models advocated earlier [109] if it is confirmed in a larger dataset.

This chapter also reported estimates of utilities for a range of toxicities and clinical characteristics associated with common treatments (chemotherapy) for NSCLC patients. These estimates of utilities are not reported elsewhere in a similar population. This may help future economic modelling, wherever appropriate estimates of utility inputs into health economic models for NSCLC can be utilized. Moreover, in the review of HTAs (Chapters 2 & 3), it was indicated that evidence for utilities appeared to be dependent on the use of estimates in Nafees et al. (2008). Adding to the evidence base from real world actual patient data is likely to be very informative for future economic evaluation of cancer treatments.

However, further research is still needed. Although the joint modelling improved the fit over the RE and BB models, the over-dispersion of EQ-5D utility may not have been optimally modelled. Moreover, there is still potential for prediction outside the range. Therefore, a combined or joint model, assuming EQ-5D responses following a BB type distribution and toxicity a Binomial or Multinomial, may be an extension of this approach. The mathematics of this is likely to be highly complex since the joint distribution of these mixed distributions will be mathematically challenging and would involve either Copula or Finite Mixture Models Techniques [201, [202].

A limitation of complex models (including joint models) is how they can be practically utilized. A RE model is reasonably straight forward to use and the BB model raises the complexity but is still usable. However, the joint modelling approach would require additional assumptions: (i) that there is a relationship between AEs and utility and (ii) an estimate of the correlation between the two. For instance, in this study, the mean EQ-5D-3L utility was about 0.54 for AE grade of 2, 0.49 for AE grade 3 and 0.31 for AE grade 4 (grade 5 is death) – Tables 6.4 & 6.5. A Williams test for trend showed this to be a statistical trend (utility becomes significantly worse as toxicity grade rises; p=0.00267). This degree of correlation needs to be ascertained in future studies if a joint mapping model is to be used successfully.

6.7 Conclusion

Mapping based on the joint modelling of utility and toxicity, in addition to covariates, offers improvements in prediction of utilities over both the RE and Beta-Binomial Model. It is also possible that the QLQ-C30 is not very good at identifying condition-specific factors such as toxicity, adverse events, and clinical features, and the EQ-5D is possibly more sensitive to. The utility increments for specific clinical characteristics from NSCLC patients in a real world NHS setting have been reported. Further research and application in a larger data set are, hence, warranted. The results recommend a complex approach that may offer improved models for mapping. The next chapter discusses this complexity further by considering a Bayesian approach to mapping, which has been reported to offer advantages over some other models discussed in the previous chapters.

Chapter 7

Chapter 7: Mapping using Bayesian Networks

Abstract

Introduction: Several mapping algorithms exist to map EQ-5D-3L from disease-specific measures, while very few algorithms attempt to map EQ-5D-5L. The aim of this chapter is to map from the QLQ-C30, a cancer-specific measure of HRQoL, on to EQ-5D-5L, using Bayesian methods. Literature appears to support the use of Bayesian methods for mapping. Hence, a Bayesian Network (BN) model is used for the purpose of this study. No BN model exists for the QLQ-C30, using either EQ-5D-3L or EQ-5D-5L.

Methods: Using data from a sample of 100 non-small cell lung cancer patients, a comparison between a Random Effects, Beta-Binomial (BB) and a BN model was carried out using five separate networks for each of the EQ-5D-5L and EQ-5D-3L domains. The R^2, AIC, MAE, and RMSE were computed to compare the model performance.

Results: Mapping based on EQ-5D-5L was superior, irrespective of the functional form of the model. However, the BN performed the worst; MAE, RMSE, AIC, R^2 and %predicted to within ±5% and ±10% for EQ-5D-5L were: 0.115, 0.140, -320, 67%, 17% and 36%, respectively. For the BB model, with EQ-5D-5L, these were respectively, 0.075, 0.092, -485, 75%, 29% and 59%.

Conclusion: Probabilistic mapping using a BN is a novel approach for the QLQ-C30 and EQ-5D-5L mapping. The overall performance of the model was complex, but not did perform as well as the Random effects or BB model with the QLQ-C30, although the BN appears to under predict (instead of over predicting) at poorer health states.

7.1 Introduction

Most mapping methods have used linear ordinary least squares (OLS) and non-linear, censored (e.g. TOBIT) type models to predict EQ-5D utilities from the QLQ-C30. In the previous chapter, a more complex joint function reported encouraging results, while very few mapping algorithms using Bayesian approaches are reported. Khan (2014) [146] and Le (2013) [147] report that a Bayesian mapping approach may have improved prediction abilities over other mapping methods. Kharroubi et al. (2015) [139] attribute the better performance of Bayesian mapping to the fact that a Bayesian model describes the uncertainty in the estimated EQ-5D-3L utility scores and is, therefore, a more appropriate method for estimating utility inputs for cost-utility analyses. A Bayesian mapping algorithm using a Bayesian Network (BN) uses the probabilistic dependencies among variables, which makes it more challenging, but potentially a more powerful predictive model. It also allows complete health state profiles to be predicted and not just utility values.

Bayesian modelling techniques can be useful in many real-life data analysis and management questions. They provide a natural way to handle missing data, they allow combining of data with expert judgments, they facilitate learning about causal relationships between variables, provide a method for avoiding overfitting of data [205] and can show good prediction accuracy even with small sample sizes. Several advantages of using Bayesian Networks for mapping have been reported [147,204]. Importantly, one can predict the health state profile as well as the utility which may have a richer source of information. No Bayesian mapping algorithm could be identified from the QLQ-C30 using a BN approach with the EQ-5D-5L from any CSM. Hence, the aim of this chapter is to map the EQ-5D-5L from EORTC-QLQ-C30 using a Bayesian Network approach and compare its results to those of the Beta Binomial (BB) mapping algorithm presented in Chapter 4. In addition, a comparison is conducted with the EQ-5D-3L.

7.2 Methods

Study Design and Data Collection

The data used in this chapter is from the Study 3 data, discussed in Chapter 5, where monthly EQ-5D-5L, EQ-5D-3L, and QLQ-C30 data were collected prospectively from NSCLC patients. The EQ-5D-5L utility scores used in this chapter are obtained using cross-walked values (available at the time) to be consistent and facilitate interpretation of results reported in Chapter 5 [188]. The EQ-5D-3L were determined using the Dolan (1997) tariff [75].

QLQ-C30 scores

QLQ-C30 domain scores were treated as categorical variables to facilitate calculations. One reason for this categorization is because a discrete form of the BN uses the groupings described in

Table 7.1 (below), following a similar method to that reported earlier [147,148]. Five categories for each QLQ-C30 domain based on a preliminary review of the distribution were used. These were 1: scores between 0 and 20; 2: for scores between 20 and 40; 3: scores between 40 and 60; 4: scores between 60 and 80 and 5: scores between 80 and 100. Khan et al., (2015) [155] have suggested that QLQ-C30 scores appear to fall into categorical groupings in practice (i.e. if the score is 25 then this falls into category 2). The QLQ-C30 was also used as a continuous value in addition to the discrete form in the BN for comparison.

QLQ-C30 Category	*Value of Domain*	*Probability of response*
1	**0 ≤** QLQ-C30 **≤ 20**	θ_1
2	**20 <** QLQ-C30 **≤ 40**	θ_2
3	**40 <** QLQ-C30 **≤ 60**	θ_3
4	**60 <** QLQ-C30 **≤ 80**	θ_4
5	**80 <** QLQ-C30 **≤ 100**	θ_5

Table 7.1: QLQ-C30 Categorization

The EQ-5D-3L uses only three possible levels (no problem, moderate problem, and extreme problem) to assess generic HRQoL the patient's health state. Although NICE considers EQ-5D-3L as an appropriate instrument to assess patient utility, the recent EQ-5D-5L allows for five levels of measuring generic HRQoL (no problem, slight problem, moderate problem, severe problem, and extreme problem), which may arguably make it a more sensitive measure (see Chapter 8). A BN for both EQ-5D-5L and EQ-5D-3L was used for the purpose of this study.

7.2.1 Modelling Approach

A Random Effects, BB (as discussed in Chapter 4) and BN model were compared.

Bayesian Network Approach

The idea of a BN approach is to estimate the probability of each of the response categories of the EQ-5D-5L (or EQ-5D-3L) conditional on the 15 QLQ-C30 responses (categories) by forming a network. A network shows the plausible relationships between several factors (e.g. Anxiety from the EQ-5D-5L with Symptom Scores from the QLQ-C30). This allows a specification of the structural form of the network based on prior (expert) judgment or belief with respect to the EQ-5D responses. The network is expressed using probability and graph theory through a visual representation of the joint probability distributions between EQ-5D-5L and the QLQC-30 in a directed acyclic graph (DAG) [147,148] as shown in Figure 7.1.

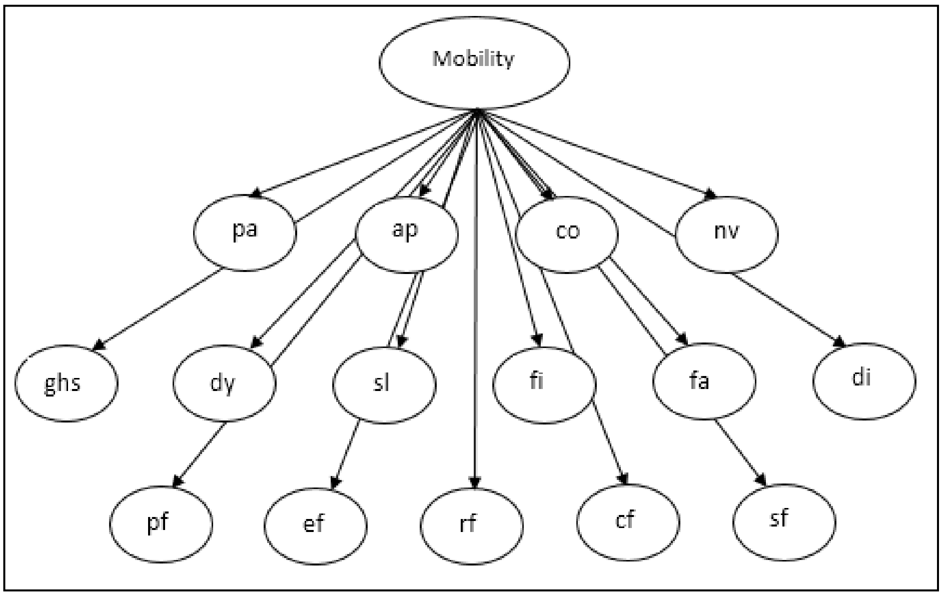


Figure 7.1: DAG Graph between EQ-5D-5L Mobility scale and QLQ-C30 15 domain scores

The DAG contains a node for every variable x_i in the domain D with n variables$(x_1, x_2, \ldots, x_n)$, with a finite set of arrows or edges between nodes, which represent the probabilistic dependencies among variables [147]. In this context, a node represents a health domain (e.g. anxiety from the EQ-5D-5L), while the state of the node reflects the possible responses to that particular domain (score of 1 to 5). Thus, a Bayesian Network $BN = (G, P)$ consists of the DAG, G-nodes and P links (or vertices) with a set of conditional probability distributions for all variables x_i in the BN. Child nodes x_i are those whose probability distribution depends on other nodes, known as the parent nodes $\pi(x_i)$ [148]. As an example, the Bayesian network that maps QLQ-C30 insomnia (SL) domain on to the EQ-5D-5L anxiety domain (AX), the structure and conditional probabilities are shown in Figure 7.2.

Parent Node: EORTC-QLQ-C30 sl → ***Child Node: EQ-5D-5L ax***

- P(sl=1) = 0.254
- P(sl=2) = 0.285
- P(sl=3) = 0.00
- P(sl=4) = 0.268
- P(sl=5) = 0.163

- P(ax=1|sl=1) = 0.698
- P(ax=2|sl=1) = 0.397
- P(ax=3|sl=1) = 0.063
- P(ax=4|sl=1) = 0.016
- P(ax=5|sl=1) =0.016

Figure 7.2: Description of Joint Probabilities

The joint probability distribution of *BN* is:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P\big(x_i | \, \pi(x_i)\big)$$

These conditional probabilities are based on Bayes rule, which states the following in the context of Anxiety (ax) from the EQ-5D-5L and Sleep (sl) from the QLQ-C30:

$$P(ax \mid sl) \;=\; \frac{P(sl\,, ax)}{P(sl)} \;=\; \frac{P(sl \mid ax)\;P(ax)}{P(sl)}$$

where $P(ax)$ is the prior probability of a response on the Anxiety domain of the EQ-5D-5L.

For the data used in this chapter, five separate Bayesian networks, one for each EQ-5D-5L and EQ-5D-3L domain are developed (Figure 7.3). The networks are graphical models describing the probabilistic relationships between EQ-5D-5L and QLQ-C30, where parameters (i.e. predicted probabilities) are obtained through the following three steps [147]:

(i) PC (Peter Spirtes and Clark Glymour) Algorithm,

(ii) EM Algorithm,

(ii) Probabilistic Inference,

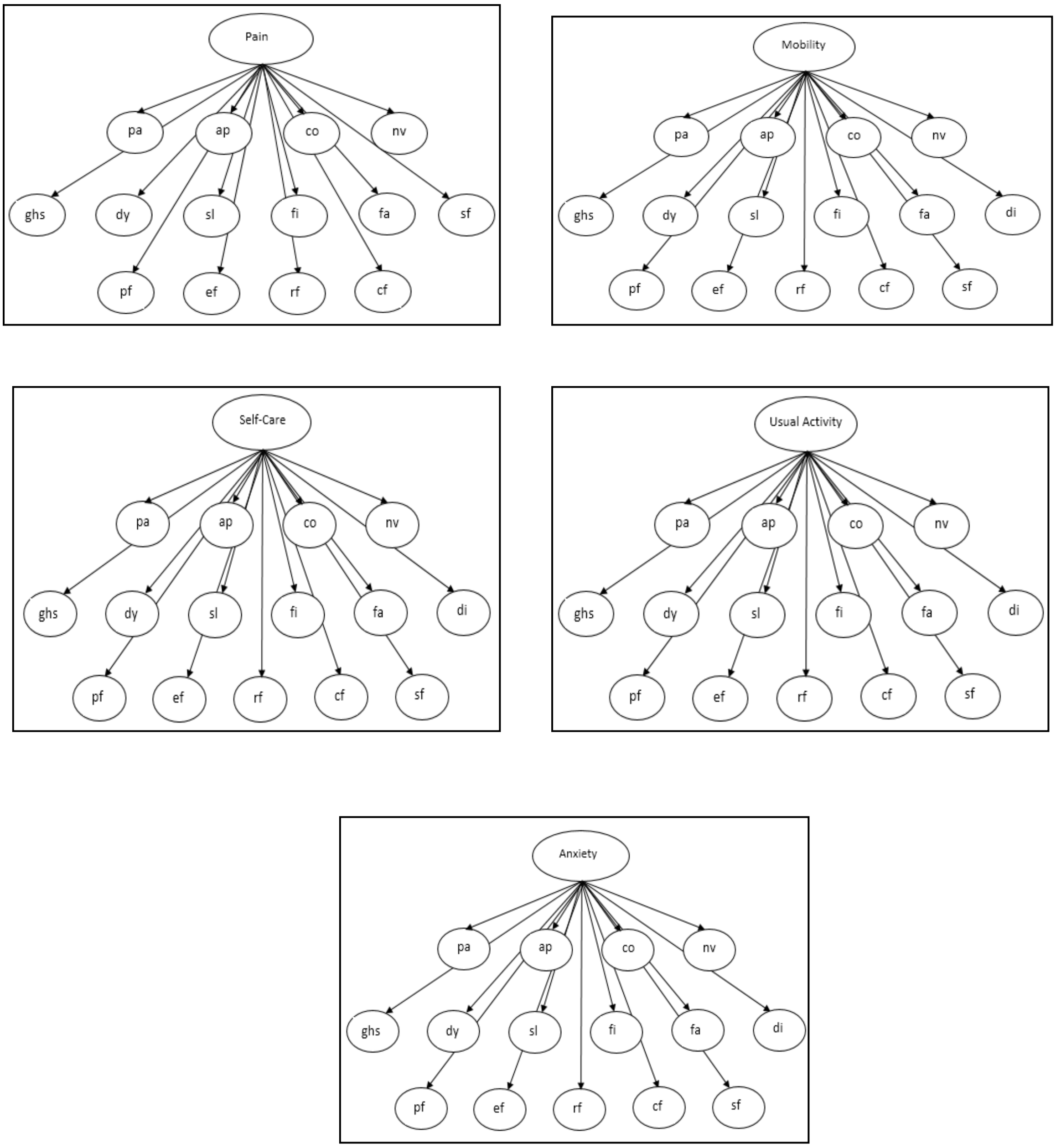


Figure 7.3: DAG for BN using Study 3 Data (EQ-5D-5L)

All networks were obtained and evaluated using the software Bayes Server version 7.8 [205].

The PC algorithm is utilized to learn the structure of the Bayesian network through testing conditional independence among each pair of variables. If tests of independence are rejected (i.e. if there is no conditional independence among the variables, it is rejected) the links between the nodes are removed (i.e. there is no point in developing a probabilistic relationship between that pair of variables because statistical tests would show a lack of dependence). Once the relationships among variables are known, the links between the nodes are formed to indicate probabilistic relationships among them. Links are identified through arcs (edges or arrows) to indicate the relationship, thus forming the graph of the network. The DAG for the BN in this setting for EQ-5D-5L and EQ-5D-3L are shown in Figure 7.3 and 7.4.

The EM algorithm is used to estimate the value of parameters in the suggested network through two steps, the E-step, finds the expected values of *Q* with respect to θ (θ is the probability of response in each of the categories) and the M-step, which maximizes Q in θ*:

$$Q\left(\theta|\theta^{(i-1)}\right) = E\left\{logP(X,Y|\theta)|X,\theta^{(i-1)}\right\} \qquad \textbf{[7.1]}$$

Where *X* is the known or observed data, *Y* is the unknown or missing data, θ^{i-1} is the known parameter estimates used to evaluate the expected values, and θ is the new parameter used to optimize *Q* (Bilmes, 1998). With some adjustments, this equation could be written as:

$$E\left\{logP(X,Y|\theta)|X,\theta^{(i-1)}\right\} = \int logP(X,y|\theta)f(y|X,\theta^{(i-1)})\,dy, \qquad \textbf{[7.2]}$$

$$\theta^{i} = argmax_{\theta}Q(\theta,\theta^{i-1}) \qquad \textbf{[7.3]}$$

After learning the structure of the BN and estimating the parameters (observed probabilities), the third step of Bayesian mapping algorithm is probabilistic inference. This step involves finding predicted probabilities of the response levels for every EQ-5D-5L domain (Le, 2013). Probabilistic inference involves computing the probability of an EQ-5D-5L response conditional on the EORTC-QLQ-C30 responses and involves using an estimate of the prior probability of response for each EQ-5D-5L response category. An example of how the probabilistic inference structure set up using categorized responses from the QLQ-C30 was demonstrated in Table 7.1.

The goal is to predict for each level of every EQ-5D-5L domain for a given patient. Hence an important difference in the mapping approach here is that EQ-5D descriptive health states rather than utility values are predicted. The predicted health states are subsequently converted to 'predicted' utilities. As an example, for Mobility. The equations below compute the probability of a Mobility score = 1 (for the EQ-5D) given good Function and reasonably low symptoms. If the

Mobility is good and the Functional domain scores are high, then the chance of Mobility score =1 (i.e. good mobility) should be high. A prior probability of 0.2 is used.

$$P(mobility = 1 \mid QLQC30) = \prod_{i=1}^{15} P(QLQC30_i \mid mobility = 1\ ) * 0.2$$

$P(mobility = 1 \mid pf = 5, rf = 5, ef = 5, cf = 5, sf = 4, ghs = 5, fa = 1, nv = 1, pa = 1, dy = 2, sl = 1, ap = 2, co = 1, di = 1, fi = 1) = P(pf = 5 \mid mobility = 1) * P(rf = 5 \mid mobility = 1) * P(ef = 5 \mid mobility = 1) * P(cf = 5 \mid mobility = 1) * P(sf = 4 \mid mobility = 1) * P(ghs = 5 \mid mobility = 1) * P(fa = 1 \mid mobility = 1) * P(nv = 1 \mid mobility = 1) * P(pa = 1 \mid mobility = 1) * P(dy = 2 \mid mobility = 1) * P(sl = 1 \mid mobility = 1\ ) * P(ap = 2 \mid mobility = 1) * P(co = 1 \mid mobility = 1) * P(di = 1 \mid mobility = 1) * P(fi = 1 \mid mobility = 1) * P(mobility)$ = 0.8,

where $P(Mobility)$ is the prior probability and equal to 0.2.

The same calculation is performed for the remaining four levels of mobility. The domain with the highest value (termed *argmax*) from equation **[7.3]** is the most likely level (score) of mobility for this patient. In this example, the predicted probability of a mobility score =1 is about 0.8. Not only does this prediction match the observed level for mobility for this patient, but it is also logical because for a patient with good physical functioning (PF = 1, i.e. score for PF >80) it seems sensible for these patients to have a high level of mobility.

The above step is then repeated for every EQ-5D-5L domain at all five levels, using the same prior of 0.2, until the entire health state of the patient is predicted (for each of Anxiety, Mobility, Self-Care, Usual Activities and Mobility).

After computing predicted probabilities of EQ-5D-5L response probabilities, a Monte Carlo simulation is used to determine the utilities in the following manner:

$$\left\{\begin{array}{c} 1\ if\ 0 < P_{1i}(X) \le 0.2 \\ 2\ if\ 0.2 < P_{2i}(X) \le 0.4 \\ 3\ if\ 0.4 < P_{3i}(X) \le 0.6 \\ 4\ if\ 0.6 < P_{4i}(X) \le 0.8 \\ 5\ if\ 0.8 < P_{5i}(X) \le 1 \end{array}\right\} u_i \sim \mathit{Uniform}\ (0,1)$$

That is, given the predicted probabilities of the response levels for all EQ-5D-5L domains obtained from the BN, estimated EQ-5D utility scores are generated using the Monte Carlo simulation method. Responses for each EQ-5D level (for each domain, separately) are determined by comparing predicted probabilities with a random number from a uniform distribution. The most likely predicted probability is used to determine utility.

In summary, the above methodology is applied as follows:
(i) Conduct statistical tests of independence between EQ-5D and QLQ-C30 to determine the statistically independence with the 15 QLQ-C30 domains. These may be excluded from the BN.

(ii) Compute the highest chance of response for a given EQ-5D domain category, for a given value (or category) of the QLQ-C30. For example, for the Anxiety score, there are five conditional probabilities associated with the PF of the QLQ-C30 category given specific values of EQ-5D Anxiety scores.

Prob[QLQ-C30$_{PF}$ = 1| Anxiety$_{EQ\text{-}5D\text{-}5L}$=1]
Prob[QLQ-C30$_{PF}$ = 1| Anxiety$_{EQ\text{-}5D\text{-}5L}$=2] etc.

However, since it is of interest to compute the chance of EQ-5D response conditional on QLQ-C30, the inference is reversed and estimation is computed through Bayes Theorem.

(iii) Derive the maximal probability associated with the most likely value of response or the highest probability across the response categories.

(iv) From the response categories, the predicted health state is generated and is compared to the observed health state (using the utilities).

(v) Monte Carlo simulation is used to repeat the process so that the average of the maximal predicted probabilities is used to determine response and subsequent utilities.

(vi) The above are repeated for each of EQ-5D-5L and EQ-5D-3L individually.

(vii) The approach was repeated by assuming that the QLQ-C30 domain scores are a continuous distribution (normally distributed).

7.3 Results

Overall Summary of Findings

Details of the study design are reported in Table 5.11 (Chapter 5).

Structure of the Network

The PC Algorithm applied in the software Bayes Server® to learn the structure of the networks indicates significant associations between almost all QLQ-C30 domains and each EQ-5D-5L domain. Only diarrhoea (DI) of QLQ-C30 had a weak association with the Pain domain of the EQ-

5D-5L (p=0.056). Consequently, five networks were considered suitable for this mapping model (Figure 7.3). However, for the EQ-5D-3L, some QLQ-C30 domains were statistically independent. For instance, Global Score (GHS), Sleep (SL), Nausea & Vomiting (NV), Diarrhoea (DI) and Constipation (CO) for Anxiety had p-values well above the 5% level and were consequently dropped from the network. An interesting conclusion that may be drawn from greater statistical independence between EQ-5D-3L compared to EQ-5D-5L is the potential that EQ-5D-5L may result in a better mapping and may be more sensitive than the EQ-5D-3L. The resulting BN for EQ-5D-3L is shown in Figure 7.4 after the similar approach was applied for other EQ-5D-3L domains.

Performance of Mapping Algorithms

Tables 7.2 and 7.3 report the results of each algorithm, where QLQ-C30 is treated as a discrete outcome and as a continuous outcome. The model using the EQ-5D-5L performed better, regardless of the model. The best fitting algorithm for the discrete form of the QLQ-C30 was based on the BB model: MAE, R^2, RMSE, AIC, %predicted with $\underline{+}$5% and % predicted with $\underline{+}$10% were 0.075, 75%, 0.092, -365, 29% and 59%, respectively for EQ-5D-5L. The BN was the worst performing algorithm (Table 7.2).

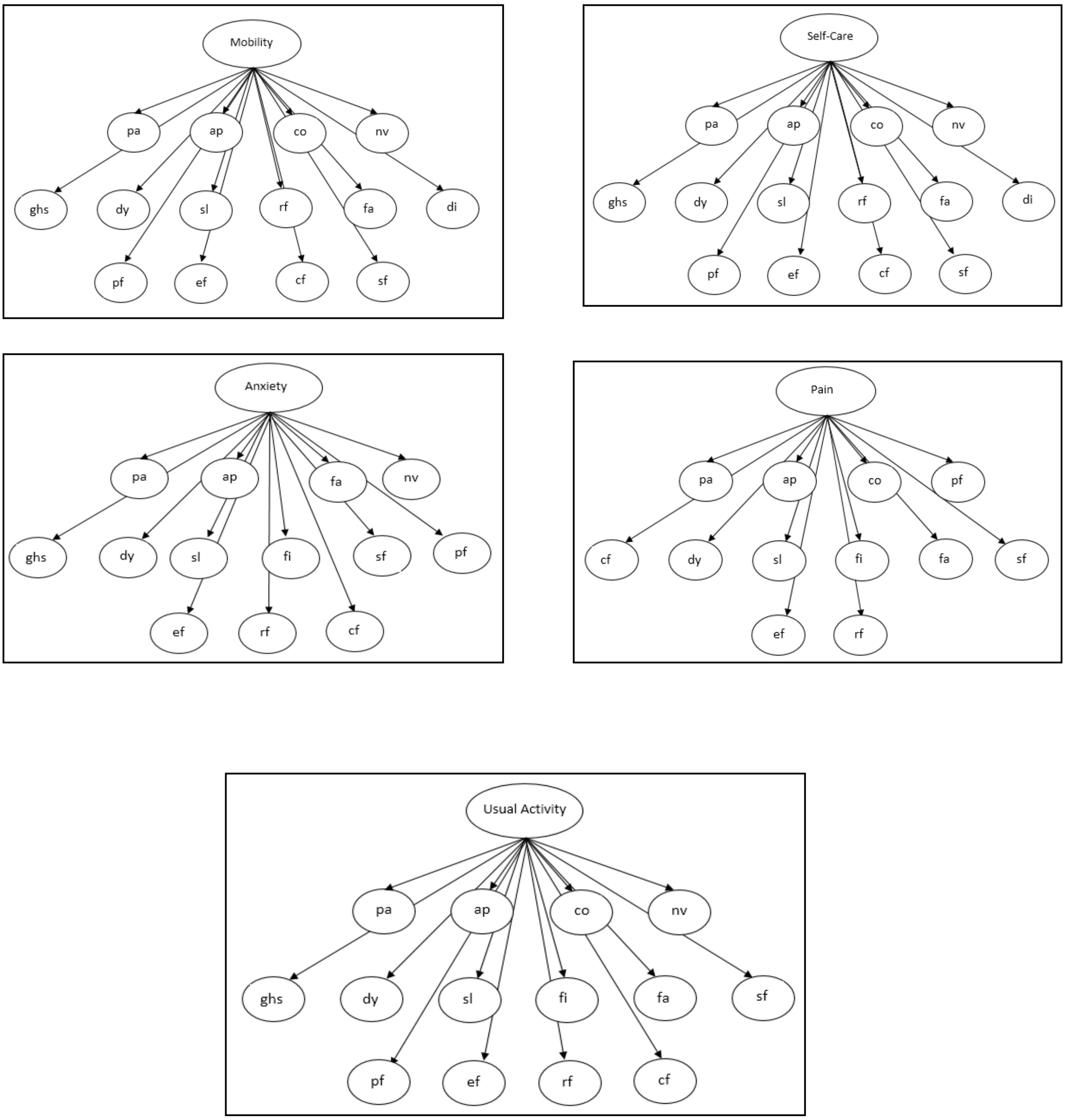


Figure 7.4: DAG for BN using Study 3 Data (EQ-5D-3L)

For the EQ-5D-3L, these were 0.220, 58%, 0.20, -107, 12% and 28%, respectively. The findings confirm the results of Chapter 5 that mapping from the EQ-5D-5L is superior. For the continuous case of the QLQ-C30, the BN performed slightly worse (Table 7.3) – the above results altered by a few points for the worst (e.g. R^2 fell from 67% to 63%). Consequently, in this application of a BN, the algorithm performed worst when compared to the existing methods.

	EQ-5D-5L			***EQ-5D-3L***		
	Random Effect	***Bayesian Model***	***Beta-Binomial Model***	***Random Effect***	***Bayesian Model***	***Beta-Binomial Model***
MAE	0.114	0.11507	0.075	0.141	0.220	0.099
RMSE	0.152	0.14024	0.092	0.183	0.200	0.113
R^2	72%	67%	75%	67%	58%	69%
AIC	-365	-320	-485	-291	-107	-385
Predicted Mean (SD)	0.577 (0.241)	0.568 (0.0182)	0.569 (0.217)	0.523 (0.252)	0.505 (0.0271)	0.532 (0.252)
Observed Mean (SD)	0.572 (0.224)	0.572 (0.224)	0.572 (0.224)	0.515 (0.308)	0.515 (0.308)	0.515 (0.308)
%predicted with ±5%	19%	17%	29%	15%	12%	23%
%predicted with ±10%	38%	36%	59%	31%	28%	33%

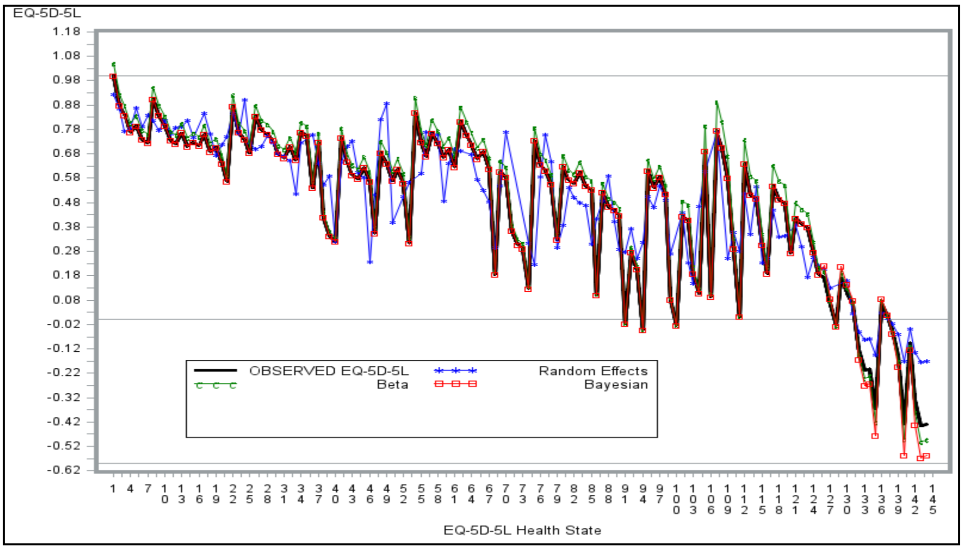
Table 7.2: Model Comparison with Bayesian using discrete form for QLQ-C30

	EQ-5D-5L			***EQ-5D-3L***		
	Random Effect	***Bayesian Model***	***Beta-Binomial Model***	***Random Effect***	***Bayesian Model***	***Beta-Binomial Model***
MAE	0.114	0.1488	0.075	0.141	0.255	0.099
RMSE	0.152	0.1482	0.092	0.183	0.210	0.113
R^2	72%	63%	75%	67%	53%	69%
AIC	-365	-287	-485	-291	-79	-385
Predicted Mean (SD)	0.577 (0.241)	0.6475 (0.020)	0.569 (0.217)	0.523 (0.252)	0.5135 (0.0283)	0.532 (0.252)
Observed Mean (SD)	0.572 (0.224)	0.6854 (0.011)	0.572 (0.224)	0.515 (0.308)	0.5206 (0.0179)	0.515 (0.308)
%predicted with ±5%	19%	17%	29%	15%	10%	23%
%predicted with ±10%	38%	33%	59%	31%	25%	33%

Table 7.3: Model Comparison with Bayesian using continuous form for QLQ-C30

Prediction by Health States

Figure 7.5 compares the observed and the predicted EQ-5D-5L utilities for each observed health state and the plot indicates good prediction at most health states. Interestingly, Figure 7.5 shows the BN model under predicting utilities at poorer health states. This is contrary to what has been observed elsewhere, where mapping algorithms have over predicted utilities at poorer health states. The BN appears to be a more conservative mapping algorithm which may imply that QALYs may be underestimated (or lower), particularly after disease progression.



Note: x-axis is health state and y-axis is predicted EQ-5D-5L utility

Figure 7.5 Observed vs. Predicted EQ-5D-5L by Health State

Cross Validation

Bootstrap (non-parametric) samples of size 100 were taken from 50% of the Study 3 data sets; 200 bootstrap samples were taken due to the computing time involved. A plot of the RMSE, R^2 and predicted utilities were generated for the BN and compared. The results confirm the bimodal distribution of the EQ-5D-5L observed earlier in Chapter 5. Interestingly, there was slightly less variability in the predicted utilities as well as RMSE. Therefore, although the BN did not perform as well as the BB model, it appears to have less uncertainty (Figure 7.6 & 7.7).

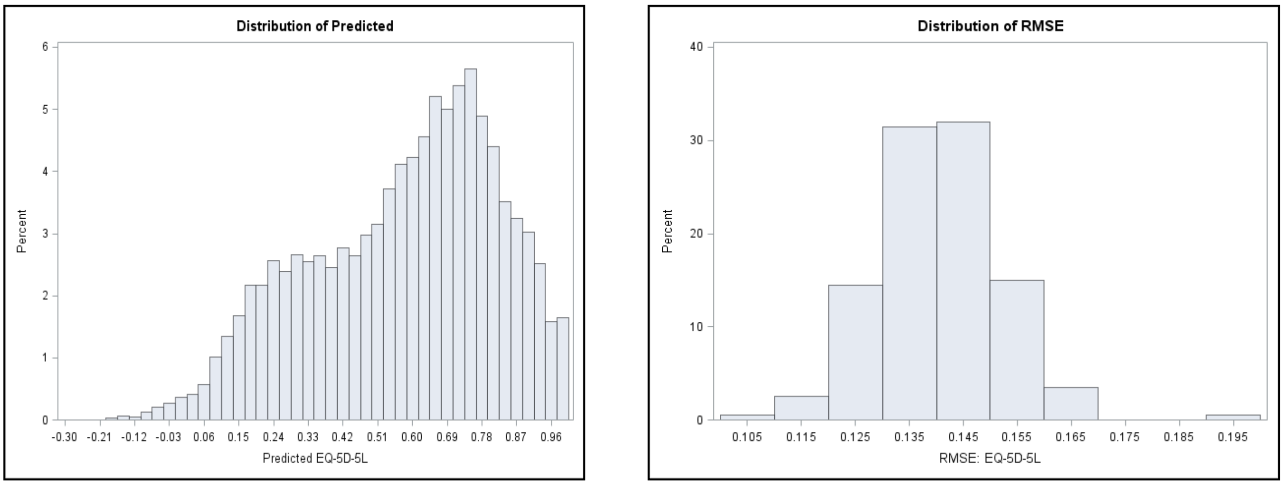


Figure 7.6: Distribution of Predicted and RMSE EQ-5D-5L from Cross Validation for BN Model

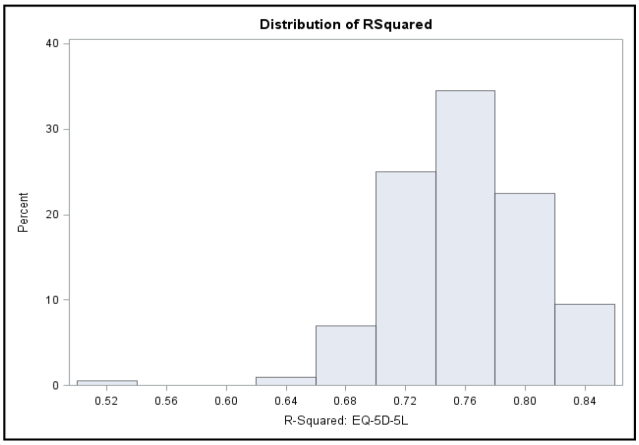


Figure 7.7: Distribution of R^2 for EQ-5D-5L from Cross Validation for BN Model

7.4 Discussion

In this ‘first time’ simultaneous mapping of the EQ-5D-5L and EQ-5D-3L from the QLQ-C30 in an NSCLC population, it has been shown that a BN algorithm performs worst among algorithms, whether using QLQ-C30 as a discrete variable or as a continuous variable. Algorithms using the EQ-5D-5L perform better than the EQ-5D-3L, regardless of the model. Interestingly, the BN model under predicts utility in poorer health states and could be considered as a conservative approach to estimating utilities.

One reason for this may be the limited number of poor health states. Less than 3% of health states were worse than health states of 5523, corresponding to a utility of about 0.181 in this study (based on cross-walked data). Therefore, estimating probabilities might be unreliable since the proportions in each health state are highly uncertain. When estimating probabilities of discrete categories, especially with the discretization of the QLQ-C30 (not carried out for other mapping models), there is likely to be a loss of information. The categorization of the QLQ-C30 is arbitrary, even if they are based on observed frequencies. In any case, using QLQ-C30 as a continuous form demonstrated worse (but potentially more realistic) model performance.

Previous research has not reported any BN with the QLQ-C30 or with the EQ-5D-5L, hence, comparisons with previous research are difficult to make. Therefore, several limitations of this modelling approach are highlighted. Firstly, a BN is computationally intensive. Even with specialist software such as Bayes server® [205], the execution time can be very lengthy (several hours). Secondly, the prior estimates of the EQ-5D were set at 0.2. These are based on a simple concept that the probability of any EQ-5D response is equally likely, which may be far from reality. If reasonably informative prior data were available, an alternative set of posterior probabilities for EQ-5D responses would be determined. Thirdly, the sample size was small; hence the results of validation approaches (cross validation) will be uncertain. The performance of an algorithm is better evaluated on its application and validity in an independent data set, for which this was not possible. Lastly, the modelling approach can become complex. The number of QLQ-C30 domains might have played a factor in the poorer performance of the Bayesian approach. In previous applications of BNs, fewer networks were used with the SF-12.

However, despite these limitations, if sufficient prior information can be gathered to propagate informative priors, the use of BN may show promise, specifically for predictions at poorer health states. The network could be further expanded by creating links between the EQ-5D-5L domains in addition to the QLQ-C30. The success of these predictive algorithms using BN has been reported in the literature in the context of mapping [147,148]. If BN can resolve the issue of over (under) - prediction at poorer health states, this would be a significant improvement over the existing models.

7.5 Conclusion

The BN mapping algorithm performs well, however, other models outperform it. With further research using other prior distributions and alternative network structures, the BN algorithm can be useful if the overprediction at poorer health states can be resolved.

In the previous chapters, more complex approaches to mapping were considered, which specifically use the EQ-5D-3L and EQ-5D-5L. In all the cases, using the EQ-5D-5L appears to demonstrate superior mapping and particularly so with the BB model. Given that the EQ-5D-5L has a different scale, a related question is what role the EQ-5D-5L (5 points) scale in itself plays in measuring response that influences mapping and the ability to detect HRQoL benefit, and impact on QALYs. This will be further investigated in the next chapter.

Chapter 8: Comparing Sensitivity between EQ-5D-5L, EQ-5D-3L, and EORTC-QLQ-C30

Published: Interpreting small treatment differences from quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30 (Khan I, Bashir Z and Forster M; Health and Quality of Life Outcomes, 2015 13:180)

Abstract

Introduction: The EORTC QLQ-C30 is a commonly used Health-Related Quality of Life (HRQoL) instrument in cancer patients. For economic evaluations, the more generic EQ-5D-3L and EQ-5D-5L are used. However, the comparative sensitivity of these instruments to detect treatment benefits remains uncertain. This research compares treatment effects amongst EQ-5D-5L, EQ-5D-3L, and QLQ-C30 within the same set of patients. Effects using odds ratios (OR) are considered, rather than the differences (MD) so that HRQoL effects can be measured on a similar scale to determine if EQ-5D is underestimating HRQoL benefits compared to the condition specific QLQ-C30.

Methods: Data from a prospective observational cohort of 100 NSCLC patients were used. Patients were followed up for at least 12 months. HRQoL was assessed at baseline and monthly thereafter, during routine hospital visits. Treatment effects were compared using both MDs and ORs from a linear and non-linear mixed effects models, respectively.

Results: EQ-5D-5L appeared to be more sensitive than EQ-5D-3L. The improvement from baseline in HRQoL was: 35% vs.25% for EQ-5D-5L vs. EQ-5D-3L; OR=1.35 (95% CI : 1.01, 1.36; p<0.001) and OR=1.25 (95% CI : 0.94, 1.66; p=0.0915) respectively; about 50% of QLQ-C30 scores showed improvements relative to baseline with ORs ranging from 1.04 (Physical Function) to 1.40 (Pain). Most (87%) QLQ-C30 effect sizes were smaller than EQ-5D-5L. The mean change from baseline for EQ-5D-3L was: 0.057 (95% CI: 0.008, 0.105; p=0.0224); for EQ-5D-5L this was: 0.0457 (95% CI: -0.0073, 0.0986; p=0.0907).

Conclusion: EQ-5D-5L appears to be more sensitive than the EQ-5D-3L and shows comparable effect sizes (ORs) to QLQ-C30. There is no evidence to suggest EQ-5D-5L is less sensitive to detecting treatment effects than either EQ-5D-3L or QLQ-C30. QALYs derived from the EQ-5D-5L are more likely to reflect HRQoL effects obtained from condition-specific measures.

8.1 Introduction

A key factor for assessing the sensitivity of a HRQoL instrument is the metric used to assess it. Related to the metric (whether a generic or CSM) is whether it can be understood and interpreted by patients and clinicians [155]. A distribution-based approach [148, 211] has been suggested where a 'moderate' effect size is defined as one-half of the SD of baseline and a 'small' effect size as 20% of the SD at baseline [212]. Other measures, like anchor-based approaches, or presenting effect sizes by disease severity, so that the treatment effects can be demonstrated have also been used [213-215]. There is no consensus as to which is the best HRQoL measure for establishing an important difference.

Recently, Khan et al. (2105) [155] highlight instances, where small but potentially important HRQoL effects can be missed by comparing odds ratios (OR) with a mean difference (MD). The OR can facilitate aligning patient and clinical understanding of HRQoL [155]. It is plausible that this is also true for EQ-5D (both EQ-5D-5L and EQ-5D-3L), as they have similar distributional properties (i.e. skewed, censored and over-dispersed data). This makes the OR an appropriate measure to compare the sensitivity of both EQ-5D and QLQ-C30; and a potential metric for determining the minimally important clinically significant difference. The OR may be a potentially suitable measure for defining a minimal difference, as its interpretation is similar to the hazard ratio (HR), which is familiar to many oncologists (and patients participating in clinical trials).

Although the EQ-5D instruments are intended for health economic evaluations, it is essential to understand, compare and interpret these effects in the context and background of CSMs such as QLQ-C30 (most widely used and considered as a 'Gold Standard'). When a condition-specific and generic measure offers conflicting interpretations over the true size of the HRQoL benefit, the Quality Adjusted Life Year (QALY) can be uncertain. Therefore, understanding the extent of HRQoL differences between the two measures is critical to determine whether the measures of effectiveness reflect higher or lower QALYs and consequently, whether a new treatment offers more or less value for money than originally believed.

In this chapter, previous findings [208, 155, 255] are further investigated to compare the sensitivity between EQ-5D and QLQ-C30 using ORs alongside MDs as a measure of sensitivity. Sensitivity refers to whether an instrument is able to detect changes in health states over time [216]. To the best of my knowledge, no direct comparison of effects between EQ-5D and QLQ-C30 within the same group of patients to investigate sensitivity have been reported in the literature so far. In cancer studies (e.g. clinical trials), the QLQ-C30 is the most widely used in Europe and can be considered as a 'gold standard'. In the USA, the FACT (discussed earlier) is considered to be the 'gold standard' for assessing HRQoL in cancer patients. The focus will be on comparing effect sizes in terms of ORs, MDs, and standardized effect sizes.

8.2 Methods

8.2.1 Study Design

Data from Study 3 were used for the purpose of designing this study. The details have already been provided in earlier chapters on the design, HRQoL instruments, and assessments. Patients in the study were newly diagnosed, good performance (ECOG 0 - 1) NSCLC patients and received their first line platinum based chemotherapy (maximum of 6 cycles, where 1 cycle is 21 days).

8.2.2 Scoring HRQoL

Scoring of the QLQ-C30 has already been discussed in detail in earlier chapters. For EQ-5D-3L, the raw responses (on a 3 point scale) were converted into a single index on a scale of -0.549 to 1.0, using the UK (Dolan) tariff [75]. Similarly, for the EQ-5D-5L (5 point scale), utilities were provided on a score between -0.549 to 1.00, described previously [217,218] mapping algorithm for the UK tariff. All responses were transformed (so that effects from all three instruments can be compared) to a 0 to 1 scale using ***Y-a/b-a,*** where ***Y*** is the HRQoL score, ***a*** is the minimum value and ***b*** is the maximum value. This transformation allows all measures of effects to be compared on the same scale. For instance, a score of 80 on a scale from 0-100 was transformed as 80 – 0/ 100-0 = 0.8. Values close to zero or one are considered to be indicative of poorer health states or nearer to ‘Full’ health for a given domain, respectively. Negative values are not possible for the QLQ-C30, unlike EQ-5D, where the above transformation can be applied so that the analysis can be conducted using a BB model.

8.2.3 Statistical Analysis

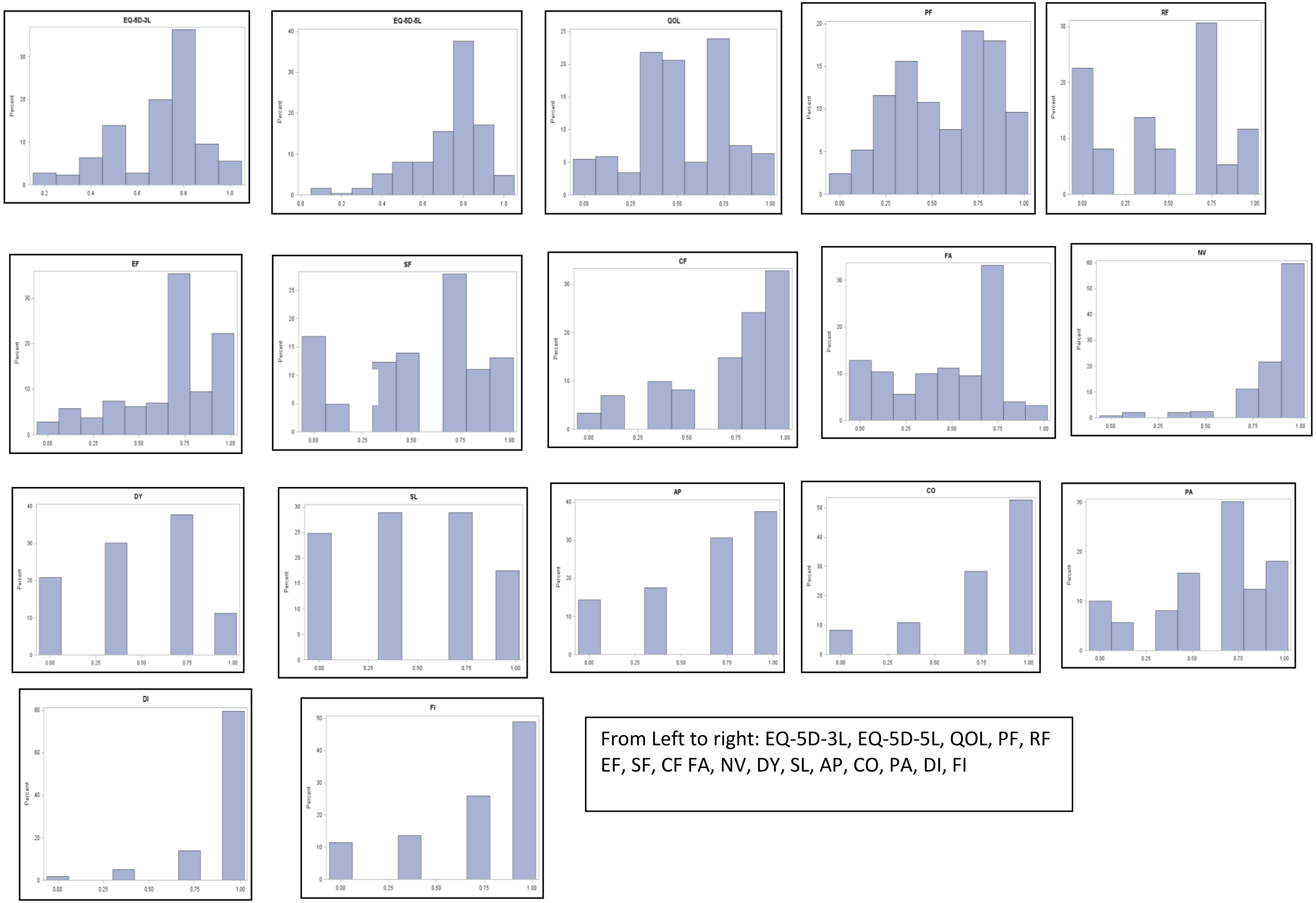
MDs and ORs were derived from comparisons between baseline and all post-baseline measures (collected monthly). A test of average differences over time was used to determine if post baseline measures could be combined. A (zero-one inflated) non-linear, repeated mixed effects Beta-Binomial (BB) model was used to present treatment effects in terms of odds ratios (ORs) [109,177]. A repeated measures linear mixed effects model was used for determining differences in terms of means. Also, both models take into account clustering (multiple measurements per subject). However, the BB model was used for this study as it handles the distributional properties (over-dispersion and skewness) particularly well (Khan and Morris, 2014; Khan et al., 2015 [155,109]). Raw mean differences and standard deviations (SD) were computed to calculate the effect size (Mean difference between post-baseline and baseline/pooled SD). Three effect size measures were [213-214]: Effect Size (ES) =Mean of within patient differences between post-baseline and baseline divided by SD of baseline responses; Standardized Response Mean (SRM) = Mean of within patient differences between post-baseline and baseline divided by SD of changes; and half SD (HSD) = 0.5 multiplied by SD of within patient differences between post-baseline and baseline (0.5*SD).

For the purpose of this analysis, the evaluable patient group (for statistical analysis) were defined as all the patients registered, who fulfilled study inclusion criteria: patients aged >18 with histologically confirmed NSCLC, gave informed consent, had a baseline HRQoL on any one of the HRQoL instruments and at least one post-baseline measure for EQ-5D-5L or EQ-5D-3L or QLQ-C30. All analyses were conducted on the transformed scale (0 to 1) unless stated otherwise. Covariates were included in the modelling (age, gender, and ECOG).

8.3 Results

The details of Study 3, along with demographic and clinical characteristics were described earlier (Chapter 5).

EQ-5D-5L and EQ-5D-3L were unsurprisingly highly correlated (r=0.904; p<0.001); EQ-5D were also correlated (Pearson's Correlation) with QLQ-C30: correlations ranged from r = 0.286 (DI) to 0.799 (PF) for EQ-5D-3L and -0.148 (Global QoL) to 0.839 (PF) for EQ-5D-5L, respectively. Responses from QLQ-C30 and EQ-5D were highly skewed (non-normally distributed); suggesting that effects based on MDs may be unreliable (Figure 8.1).

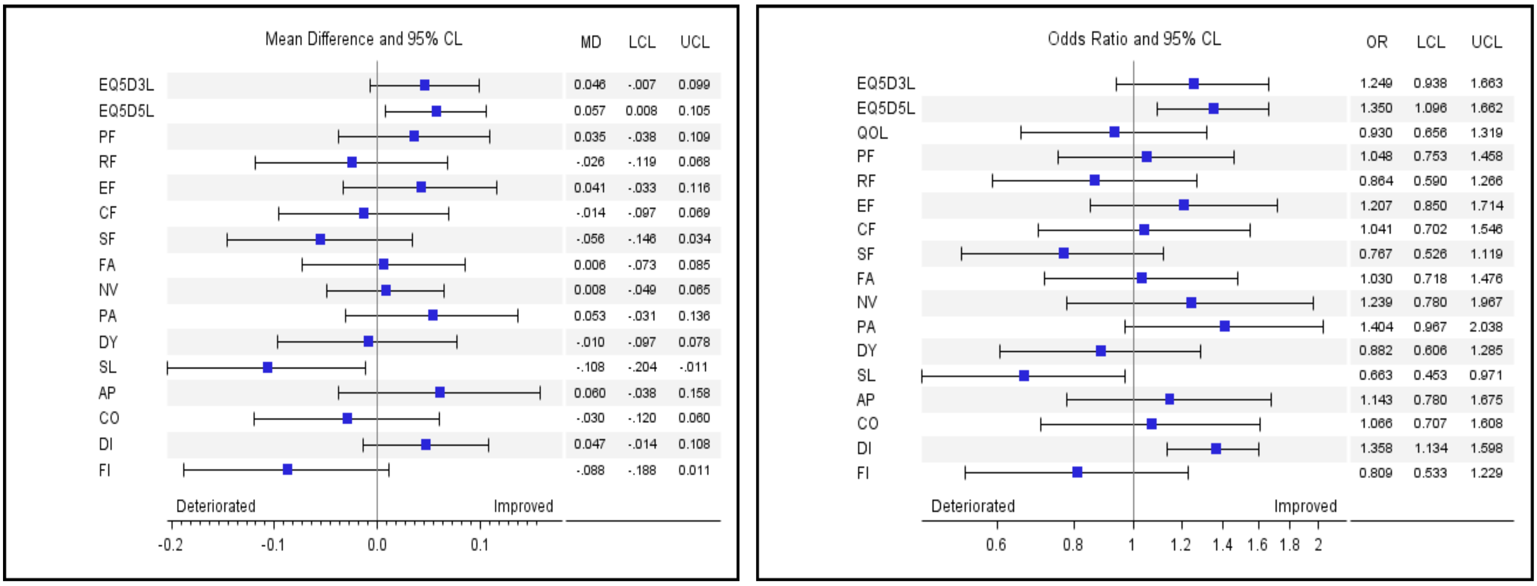


From Left to right: EQ-5D-3L, EQ-5D-5L, QOL, PF, RF
EF, SF, CF FA, NV, DY, SL, AP, CO, PA, DI, FI

Figure 8.1: Distribution of EQ-5D-5L, EQ-5D-3L and QLQ-C30

8.3.1 Comparing Effects between EQ-5D-5L, EQ-5D-3L: Changes over Time

EQ-5D-5L appears more sensitive than EQ-5D-3L, while demonstrating comparable effect sizes to QLQ-C30. Patients are 35% more likely to improve their HRQoL (utility) after treatment (post baseline) compared to baseline (OR=1.35; 95% CI: 1.096 - 1.662; p<0.001) for EQ-5D-5L (Forest Plot Figure 8.2, 8.3 and Table 8.1); for EQ-5D-3L this was 25% (OR=1.25; 95% CI: 0.938 - 1.66; p=0.0915).



a) Odds Ratios (Relative to Baseline)

b) Mean Differences (Relative to Baseline)

Figure 8.2: Forest Plot of HRQoL Effects for All Instruments

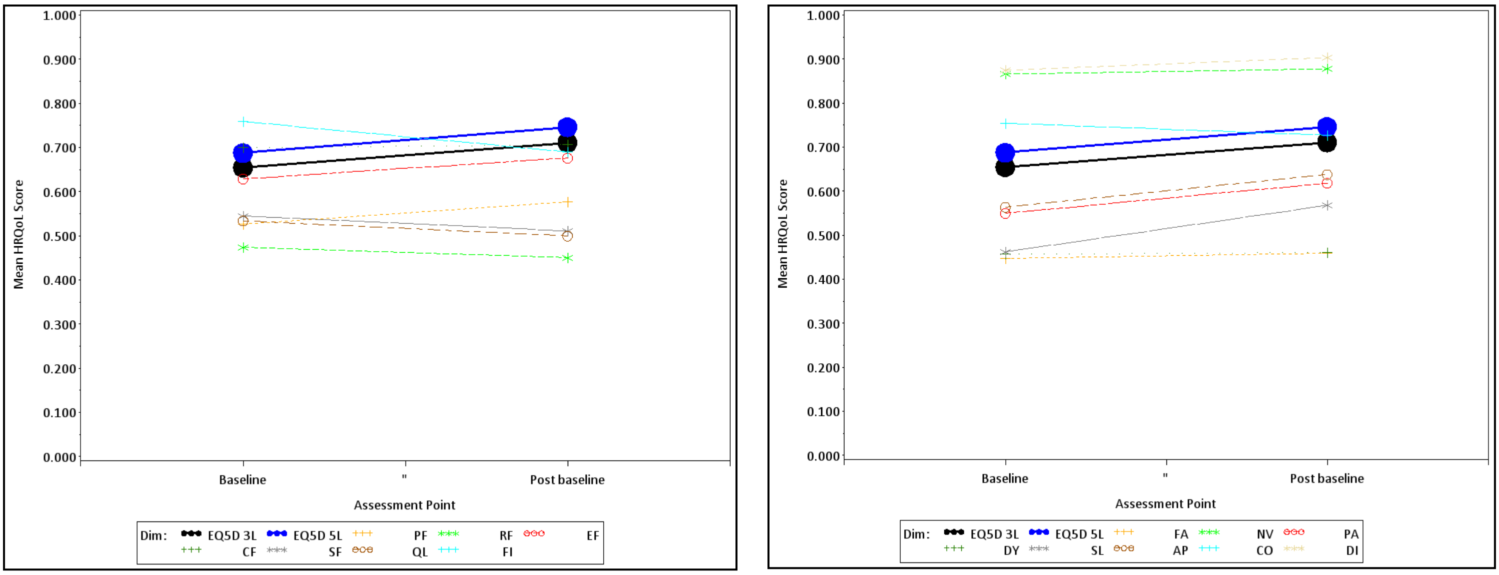
Global Health Status Score (QL), Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF); Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (SL), Appetite Loss (AP), Constipation (CO), Diarrhoea (DI), Financial Problems (FI) and an overall score

	OR	L95%	U95%	P-value	MD	L95%	U95%	P-value
EQ-5D-3L	1.249	0.938	1.663	**0.0915***	0.04570	-0.00729	0.09869	**0.0907**
EQ-5D-5L	1.350	1.096	1.662	**<0.001***	0.05666	0.008106	0.1052	**0.0224**
QOL	0.930	0.656	1.319	0.6825	-0.02146	-0.09098	0.04806	0.5437
PF	1.048	0.753	1.458	0.7809	0.03540	-0.03827	0.1091	0.3448
RF	0.864	0.590	1.266	0.4526	-0.02583	-0.1193	0.06764	0.5867
EF	1.207	0.850	1.714	0.2920	0.04135	-0.03334	0.1160	0.2766
CF	1.041	0.702	1.546	0.8399	-0.01371	-0.09651	0.06909	0.7447
SF	0.767	0.526	1.119	0.1676	-0.05623	-0.1465	0.03403	0.2210
FA	1.030	0.718	1.476	0.8728	0.005743	-0.07310	0.08459	0.8860
NV	1.239	0.780	1.967	0.3631	0.007769	-0.04899	0.06453	0.7877
PA	1.404	0.967	2.038	**0.0739***	0.05267	-0.03110	0.1364	0.2167
DY	0.882	0.606	1.285	0.5117	-0.00954	-0.09671	0.07762	0.8295
SL	0.663	0.453	0.971	**0.0351***	-0.1078	-0.2041	-0.01144	**0.0285**
AP	1.143	0.780	1.675	0.4913	0.05991	-0.03814	0.1580	0.2299
CO	1.066	0.707	1.608	0.7588	-0.02983	-0.1201	0.06047	0.5158
DI	1.358	1.134	1.598	0.0006	0.04681	-0.01404	0.1077	0.1310
FI	0.809	0.533	1.229	0.3190	-0.08820	-0.1879	0.01149	**0.0826**

Table 8.1: HRQoL Changes Relative to Baseline: Odds Ratio and MD

Note: Statistically significant at 10% level. Shaded values show the direction of effects to be the same. MD: Difference between post-baseline versus baseline

In comparison, model-based estimates (Linear Mixed Model) of HRQoL MDs improved by only 8% and 7% for EQ-5D-5L and EQ-5D-3L respectively. This discrepancy may reflect the strong skewness in the data, a feature also noted in previous analyses of similar HRQoL data (Khan, 2015) [155]. EQ-5D utilities were moderately skewed (p-value <0.001, for Kolmogorov-Smirnov test of normality, Figure 8.1). The estimates of MDs (post-baseline vs. baseline) were similar: 0.0461 (observed) vs. 0.0457 (Linear mixed) for EQ-5D-3L; and 0.0573 (observed) vs. 0.0567 (Linear mixed) for EQ-5D-5L. This suggests that an improvement in HRQoL from baseline by about 0.046 points translates into about a 24% (Figures 8.2, 8.3 and Table 8.1) improvement in HRQoL for EQ-5D-3L; and for EQ-5D-5L, a 0.057 point improvement in MD corresponds to a 36% improvement on an OR scale. The improvement may partially be due to the fact that these are newly diagnosed patients, who are starting treatment and have better HRQoL prior to disease progression.



a) **Function Domains of QLQ-C30, EQ-5D-5L and EQ-5D-3L**

b) Symptom Domains of QLQ-C30, EQ-5D-5L and EQ-5D-3L

Figure 8.3: Mean HRQoL Scores (Transformed Scale) Relative to Baseline

8.3.2 Comparing Effects over Time with QLQ-C30

About 8/15 (53%) of the QLQ-C30 effects (ORs) showed improvements relative to baseline, ranging from OR=1.04 (Physical Function) to OR=1.40 (Pain). On an average, patients are 40% more likely to show improved pain symptoms after treatment (post-baseline), compared to before starting the treatment (baseline). Most (87%) QLQ-C30 effect sizes were smaller than those of EQ-5D-5L (Figure 8.2 & 8.4). In contrast, 7/15 (47%) of QLQ-C30 effects based on MDs showed improvements (below the commonly stated target of 10 points, often cited [219,220]).

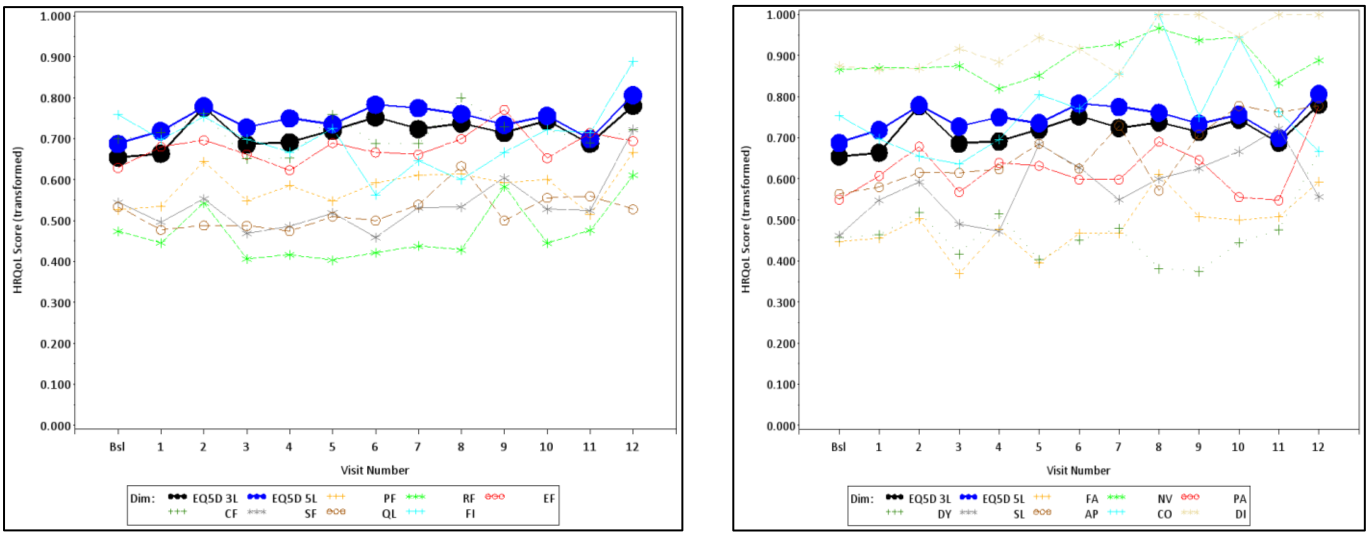


Figure 8.4: Changes over Time for All Instruments (Transformed Scale)

Differences between apparently larger effects with ORs and ‘smaller’ ones with MDs are likely due to the presence of skewness in the data. For instance, pain (PA) improved by 0.053 points (equivalent to a 5 point difference on the original scale) and 40% (OR=1.4) with an OR (Table 8.1 and Fig. 8.2 & 8.3) and a sizeable proportion of patients scored highly (better pain control). The left skewness influences the mean and consequently MD. Although for health economic evaluations, the mean is the statistic of choice, from a clinical perspective, this may not necessarily be the case, a theme which is discussed in section 8.4.

8.3.3 Standardized Effect Sizes

The standardized effect sizes were generally larger for EQ-5D-5L compared to EQ-5D-3L, which is consistent with the above findings (Table 8.2). The ES, SRM and 0.5*SD were 0.413, 0.314 and 0.090 for EQ-5D-5L and 0.315, 0.230 and 0.100 for EQ-5D-3L respectively. For QLQ-C30, the values ranged from: -0.031 (DY) to 0.302 (SL) for ES, 0.0216 (FA) to 0.285 (SL) for SRM and 0.115 (NV) to 0.218 (FI) 0.5*SD. These measures of effect sizes appeared to indicate the sensitivity of EQ-5D-5L, were comparable to the QLQ-C30 and were similar to those previously reported [221] in lung cancer patients.

	MD*	ES[1]	SRM[2]	0.5*SD[3]
EQ5D3L	0.046	0.315	0.230	0.100
EQ5D5L	0.057	0.413	0.314	0.090
QOL	-0.021	-0.076	-0.240	0.157
PF	0.035	0.114	0.128	0.135
RF	-0.026	-0.070	-0.069	0.186
EF	0.041	0.148	0.117	0.174
CF	-0.014	0.048	0.144	0.155
SF	-0.056	-0.159	-0.142	0.196
FA	0.006	0.206	0.021	0.138
NV	0.008	0.037	0.034	0.115
PA	0.053	0.148	0.184	0.143
DY	-0.010	-0.031	-0.025	0.143
SL	-0.108	0.302	0.285	0.197
AP	0.060	0.171	0.150	0.189
CO	-0.030	0.093	0.088	0.199
DI	0.047	0.179	0.207	0.168
FI	-0.088	-0.256	-0.201	0.218

Table 8.2: Comparison of Effect Sizes between EQ-5D and QLQ-C30

*Within patient mean difference divided by pooled SD

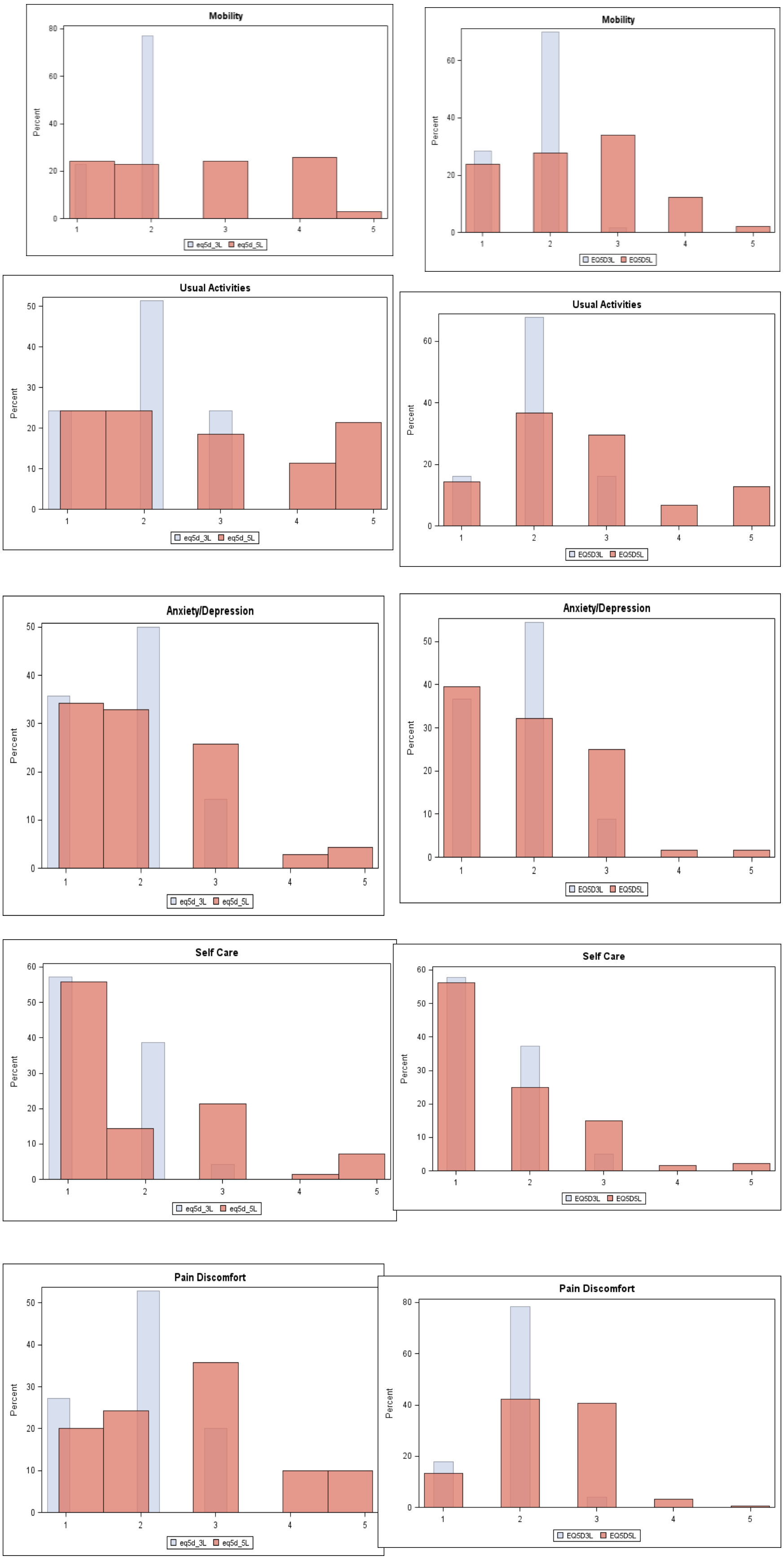
[1] MD / SD of baseline

[2] MD / SD of changes between post-baseline and baseline

[3] 0.5*SD of changes between post-baseline and baseline

Comparisons Based on Response Categories for EQ-5D-3L and EQ-5D-5L

The responses at baseline and post-baseline are shown for each domain of the EQ-5D-5L and 3L. Inspection of Figure 8.4 reveals that while the EQ-5D-3L distribution looks relatively stable from baseline to post-baseline, the EQ-5D-5L distributions appear to shift. For instance, for Mobility with EQ-5D-3L in the first three response categories (1:'No Problems', 2:'Slight Problems' and 3:'Moderate), the proportion of baseline vs. post-baseline responses were 22% vs. 26% and 78% vs. 74% (Figure 8.5) and for EQ-5D-5L, these were 22% vs. 24%, 20% vs. 25%, and 22% vs. 32%, respectively. For EQ-5D-5L, post baseline, the proportion of responses in the first three categories appear to be increasing, which suggests that the expanded scale has improved sensitivity and responsiveness.



Key:

EQ-5D-5L:

For Mobility, Self-Care and Usual Activities;

1:“No problem”
2: “Slight problems”,
3: “Moderate problems”
4: “Severe problems”
5:”Unable”

For Pain/Discomfort and Anxiety/Depression,
1:”No problem”
2: Slight problem”
3: “Moderate problem”
4; “Severe problem” and
5:”Extreme problem”

EQ-5D-3L:

For Mobility, Self-Care and Usual Activities;

1:“No problem”
2: “Some problems”
3: “Unable”

For Pain/Discomfort and Anxiety/Depression
1:“None”
2: “Moderate”
3: “Extreme”

Figure 8.5: Distribution of Responses for EQ-5D-5L and EQ-5D-3L Pre (left) and Post-Baseline (right)

8.3.4 Adjustments for Covariates

Interactions between pre and post-baseline with some potential predictors of HRQoL (age, gender, ECOG, and Stage) were tested. Both EQ-5D-5L and EQ-5D-3L MDs (post-baseline minus baseline) were larger for males vs. females (0.062 vs. 0.027 and 0.065 vs. 0.047) respectively and statistically significant with p-values of <0.001. This indicates the ability of both the instruments to detect HRQoL changes (where they exist) in subgroups. For QLQ-C30, differential changes between males and females were observed only for PF (p-value=0.0338), SL (p-value=0.0285). A similar result was identified for ECOG and also with ORs. All instruments were able to detect differences across ECOG, yielding varying effect sizes, and depending upon ECOG. For instance, MDs of 0.056, 0.031, 0.0256 and 0.118 were observed for EQ-5D-5L for ECOG 0, 1, 2, 3 and 4, respectively.

8.4 Discussion

Treatment effects from three important and commonly used HRQoL instruments have been presented and it has been shown that EQ-5D-5L is more sensitive than EQ-5D-3L. Moreover, the treatment effects from EQ-5D-5L (relative to baseline) are comparable to those of QLQ-C30. EQ-5D-5L also appeared to be more sensitive than the EQ5D-3L for worsening severity (e.g. poorer ECOG status). This is true whether effects are reported as MDs, ORs or standardized effect sizes. ORs were used because distributions of QLQ-C30 and EQ-5D were skewed. Using ORs allows researchers to interpret small effects, which can be missed (or dismissed) as irrelevant or uninterpretable.

ORs have been used previously in the analysis of HRQoL. Feddern et al. (2015) [222] report them for assessment of pain. Others [223] use a propensity score (logistic regression) approach to report odds of HRQoL deterioration. ORs with the QLQ-C30 in renal impaired patients have also been reported [224]. In these analyses, scores were dichotomized in order to generate the ORs. However, in this analysis, no such dichotomization (and consequent loss of information) was required, due to the flexibility of the Beta-Binomial regression approach. Given the acknowledged difficulties in interpreting standard effect sizes (Norman et al., 2003) [211] in a clinically meaningful context, these findings offer an approach to detect effects that can be interpreted clinically and judged against the background of an economic evaluation context because utilities and clinical effects can be assessed on the same scale.

A useful property of the OR in situations, where a generic instrument is considered to lack sensitivity (e.g. EQ-5D-3L shows a small effect) is that it can contextualize these effects. Therefore, it can be concluded in a more informative way whether a QALY was unduly underestimated or not. As noted earlier, an apparently small MD in NV (from QLQ-C30) of about 0.008 points (Figure 8.2) can be interpreted as a 24% improvement in nausea and vomiting symptoms, rather than a mean change of 0.008 points. Consequently, EQ-5D utilities used to adjust efficacy outcomes (e.g. for QALYs) may not necessarily yield low QALYs because of a lack of sensitivity because it is a generic HRQoL measure. The plausible hypothesis that if an instrument is generic it may not capture condition specific HRQoL features, does not appear tenable from these findings. It is likely that EQ-5D provides adequate estimates of HRQoL. Any potential loss of information from the condition-specific QLQ-C30 is likely to be compensated by the expanded EQ-5D-5L scale (because EQ-5D effects are sometimes larger than QLQ-C30). Most of the items of QLQ-C30 are on a four-point scale; hence, EQ-5D-5L, with a five-point scale, may be expected to show greater sensitivity. To emphasize this point, from Table 8.1 we observed differences that are likely to be perceived as small; yet when these differences are considered on a relative scale, they suggest a more meaningful impact on patient HRQoL that might lead us to believe using a MD. In general, generic measures are reported considered to have less sensitivity. Clinically relevant effect sizes from generic measures such as the EQ-5D have been suggested to range from 0.03 [157,158]. In other studies too, these effect sizes are considered to be clinically important [225,226]. The minimally important difference for the EQ-5D (0.074) was almost double that for the SF-6D (0.041) – proportionally equivalent to the range of utility scores for each scale [225,226].

In this analyses, whereas both EQ-5D-3l and 5L report effect sizes greater than 0.03, the suggested minimum for the EQ-5D, none of the effect sizes for the QLQ-C30 satisfied what is considered to be a clinically relevant difference of 15 points (or even 10 points). On the basis of the data in this chapter, the effect sizes from the EQ-5D are commensurate with clinical relevance in the context of other similar expectations for these instruments, whereas for the QLQ-C30, this is not the case. Consequently, EQ-5D does appear to be more sensitive.

The findings here confirm earlier work [208,221] that EQ-5D-3L has comparable effect sizes to QLQ-C30, using measures of standardized effect sizes. The findings presented extend this research for both EQ-5D-3L and EQ-5D-5L in a larger study. In addition, this is shown with several metrics, which take into account the skewed distribution of the data. Values of 0.5*SD have been reported ranging from 0.07 to 0.11, which is comparable to what is reported here [221]. However, Krahn et al. (2007) [152], conclude that generic measures are less sensitive and should be complemented by a CSM. However, in both of these earlier and later studies the use of the MD statistic or its variants (e.g. ‘standardized effect size’, ‘standardized response mean’) makes interpretation of the effects challenging. Others [211] conclude that EQ-5D-3L is not as sensitive as the 15D or SF6D; 15D has a wider scale than EQ-5D-3L, which, therefore, supports the finding that EQ-5D-5L is more sensitive than EQ-5D-3L and possibly QLQ-C30 [211,224,227]. Based on MDs, Le et al. (2013) [182] also conclude that EQ-5D-5L was as sensitive to the condition-specific FACT-B in breast cancer patients.

The interpretation of HRQoL effects, whether as an ES, MD, SRM, ½ SD, ORs or variants of these, remains a challenge. For instance, it is often unclear how a 2 point MD change in PF (which is statistically significant) is clinically important for patients and clinicians. Conclusions are often relegated to statistical interpretations, especially when the observed differences are small.

Although the OR is not a direct metric used in economic evaluation, the BB model is mathematically tractable to estimate the mean EQ-5D from a given OR. It is important that when an apparently small (or large) mean utility or QALY is observed, it can be elucidated with a statistic such as an OR. Interestingly, a way to overcome the difficulty of interpreting the standardized effects size would be to convert the effect size scale, so as to interpret effect sizes as a proportion (of patients), who benefit from treatment [212]. The approach presented here follows a similar line of thought but is more direct. It allows one to subsequently compare these proportions through ORs using a BB modelling approach while quantifying the uncertainty (using confidence intervals). Quantifying uncertainty of the ES is not as simple as using confidence intervals. A simulation is an alternative approach; in initial simulations, it was noted that the potential for considerable uncertainty exists in (standardized) effect sizes. This is an interesting area for further research.

This study was an observational study with the main objective of assessing HRQoL over time and therefore has limitations. Firstly, the study has a relatively small sample size, although it is larger than the sample size utilized in earlier similar studies [208,221]. Revicki (2006, 2008) [213-214] observes that for assessing the sensitivity of HRQoL instruments, an observational study is adequate. For the purposes of deriving QALYs, ORs are not a statistic useful as an input into a health economic model; however, the mean can be estimated from the Logit function. The use of the OR may be considered a complex metric to interpret, but is not any more difficult than the standardized effect sizes and hazard ratios. Markers of disease evolution were collected (progression free survival): the results relating to treatment effects are consistent with what is observed with pre and post progression utilities. The mean pre and post progression utilities over time were 0.566 and 0.474 respectively with EQ-5D-3L; with EQ-5D-5L these were 0.573 and 0.432. Hence the 5L also seems to be more sensitive in the sense that higher utilities were observed at each of the two clinical states (pre- and post progression) which are consistent with the descriptive states from the EQ-5D.

However, further research must be conducted to quantify the loss of information between generic and CSMs using the relationship between MDs, ES, SRM, ORs and impact on QALYs

8.5 Conclusion

To conclude, EQ-5D-5L is more sensitive than EQ-5D-3L and can yield more precise HRQoL effects. In this chapter, EQ-5D-5L has shown comparable HRQoL benefits to QLQ-C30. Effects sizes based on ORs may contextualize small (or large effects) from EQ-5D effects compared to QLQ-C30; and can elucidate the interpretation of QALYs, especially in borderline decisions of cost-effectiveness. This is the first study that has examined the sensitivity of EQ-5D-5L and EQ-5D-3L with the QLQ-C30 in this manner.

The sensitivity of the instruments is vital to determine whether the utilities reflect estimates of HRQoL benefit from disease-specific HRQoL measures. The availability of different approaches to estimation of utility, particularly through several published mapping algorithms creates a difficulty in ascertaining which if any algorithm is optimal or more useful. Hence, in the next chapter, methods for selecting an optimal or more 'useful' mapping algorithm amongst those published are developed and evaluated.

Chapter 9

Chapter 9: Deciding between Published Mapping Algorithms to Predict EQ-5D Utility Scores from QLQ-C30 in Lung Cancer Patients

Abstract

Introduction: There are several published algorithms for mapping QLQ-C30 onto EQ-5D-3L. However, there is an absence of robust validity of these algorithms and in particular, a method to select an optimal algorithm. The performances of these algorithms are compared using commonly reported metrics. A criterion of algorithm selection using a ‘traffic light’ system is based on a simulation approach.

Methods: Data from three studies in NSCLC were used to predict and compare observed EQ-5D-3L utility scores from nine published algorithms. Cut-off values for R^2, root mean squared error (RMSE) and percentage of utilities predicted within ±10% of the observed value were determined from bootstrap simulated mean values in order to compare and select algorithms.

Results: Mapping algorithms were classified into three groups: ‘Green: Good/Useful’, ‘Amber: Useful but with caution’ and ‘Red: Poor’. The simulated R^2 values ranged between 0.14 to 0.61, RMSE from 0.17 to 0.21 and percentage predicted within ±10% of the target from 19% to 55%. The cut-off values from bootstrap simulations for these were 0.45, 0.182 and 36% respectively. Three algorithms were classified as ‘Green: Useful’. These had ≥50% chance of R^2 and % predicted exceeding 0.45 and 36%, respectively. There was also at least 50% chance of RMSE < 0.182; five algorithms were considered as ‘Red: Not Useful’ (>50% chance of worse than average values for RMSE, $R^{2,}$ and % predicted) and one algorithm was classified as ‘Amber: Useful but with caution’.

Conclusion: For estimating predicted utilities in patients with NSCLC from the QLQ-C30 instrument, a ‘useful’ set of mapping algorithms have been identified. Using bootstrap simulations, a criterion for model selection using a basic multi-criteria selection approach is offered. Given the increasing number of algorithms available, further research into multi-criteria approaches for deciding between algorithms is required.

9.1 Introduction

In the previous chapters, alternative functional forms of algorithms and approaches to predicting patient level utilities were discussed. Previous reviews of mapping functions [112] were restricted to developing algorithms with emphasis on their performance based on various metrics. The emphasis was on model structure and complexity and less about the practical consequences of models – such as whether the model is usable, and how to select from an increasing number of available algorithms. A multi-criteria approach for selecting an algorithm may, therefore, be required to compare their performance (as there are several measures that are used to assess usefulness in practice).

Current approaches for testing the validity and usefulness of published algorithms are based on using similar statistical metrics [163], but are less formalized and determined from a single instance of prediction. Crott (2014) [161] and Arnold et al. (2015) [162] investigated the practical value of algorithms by applying published mapping algorithms to independent data sets. Arnold et al. (2015) tested several algorithms on a small data set of n=73 in patients with malignant mesothelioma. Similarly, Crott (2014) [161] also tested some of these algorithms on 219 breast cancer and 172 non-small cell lung cancer (NSCLC) patients. Both Crott (2014) and Arnold et al. (2015) [161,162] tested algorithms in a way similar to Doble and Lorgelly (2016) by using various statistical criteria, without developing any selection criteria. Also, no formalized approach to selection had been considered so far [72,163].

In this chapter, a similar approach to previous methods [161,162] is adopted, but on a larger number of NSCLC patients. In addition, comparisons between observed and predicted health states are performed. A decision chart based on common criteria (such as R^2, RMSE, % predicted) is proposed for the feasibility of the selection decision between algorithms using bootstrap simulation.

9.2 Methods

9.2.1 General Methods

Literature Review of Published Algorithms

Based on the review presented earlier (Chapter 2), several algorithms based on the following criteria for selection were identified in this chapter:

Criteria for Selecting Published Mapping Algorithms

The criteria for selection of algorithms were based on:

a) Ensuring that QLQ-C30 was used as a part of the process of developing the algorithm.

b) At least 2 coefficients in the mapping function were reported (the intercept and at least one of the 15 domains (these domains consist of 5 functional domains: Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF); 8 symptom domains: Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI); a domain for Financial Problems (FI) and an overall Global Health Status Score (QL)).

c) Algorithms were included irrespective of the tumour type. Moreover, most algorithms were reported in a manner that there were no restrictions on use in any specific tumour type.

d) Algorithms were included, whether developed from the randomized clinical trial (RCT) data or other studies (e.g. surveys, observational studies).

e) Algorithms were used irrespective of which country the data were generated from (or which language the studies were published in).

Selected Mapping Algorithms

Nine algorithms were selected based on the above criteria. The selected studies are described and summarized in Table 9.1 below:

Algorithm	Published Equation or Formulae for EQ-5D-3L prediction
I: Mckenzie (2009)	EQ_{I} =0.2376 – 0.0006*FI-0.0003*DI+0.0001*CO+0.0003*AP+0.00004*DY-0.0024*PA-0.0005*NV - 0.0021*FA+0.0002*SF+0.0009*CF+ 0.0028*EF+ 0.0022*RF+0.0004*PF+ 0.0016*QL+0.00004*SL
II:Kontodimopoulos (2009)	EQ_{II} =-0.1814+ 0.00546*QL + 0.00313*EF +0.00508*PF
III: Jang et al. (2010)	EQ_{III} = 0.3381 +0.0035*PF+0.0007*RF+0.0011*EF+0.0007*CF-0.0007*SF+0.0009*QL+0.0003*FA-0.0002*NV-0.0021*PA-0.0001*DY -0.0001*SL-0.0001*AP+0.0005*CO+0.0004*DI-0.0001*FI
IV: Crott et al. (2012)	EQ_{IV} = 1-[0.85927- 0.006963*PF – 0.008734*EF – 0.003993*SF +0.0000355*PF^2 +0.0000552*EF^2 +0.0000290*SF^2+ 0.001145*CO+0.003561*PA-0.0003678*SL-0.0000540*DI^2 + 0.0000117*SL^2
V: Kim-K (2012)	EQ_{V} = 0.5534+0.0007*QL+ 0.0032*PF+0.0005*EF+0.0005*SF-0.0013*PA +0.0008*DY-0.0006*AL+0.0005*DI
VI: Versteegh et al. (2012)	EQ_{VI} = 0.130 +0.0007*QL+ 0.0032*PF+0.0005*EF+0.0005*SF-0.0013*PA +0.0008*DY-0.0006*AL+0.0005*DI
VII: Kim-A(2013)	EQ_{VII} =0.53897+0.002*PF+0.002*RF+0.003*EF+0.001*SF+0.001*QL+ 0.001*FA-0.001*PA-0.001*CO
VIII: Khan (2014)*	EQ_{VIII} = 1.318+0.260*PF+0.2340*RF+0.379*EF+0.257*SF+0.061*CF-0.012*FA+0.062*NV-0.235*PA+0.0088*DY-0.102*SL +0.0017*AP +0.026*CO+0.051*DI+0.062*FI+0.224*QL
IX: Proskorovsky (2014)	EQ_{IX} =0.1197+0.000961*FI-0.0003*DI-0.0005519*CO- 0.000479*AP +0.000376*DY-0.00229*PA+0.000573*NV - 0.0000575*FA+0.0000447*SF-0.000378*CF+ 0.00104*EF+ 0.000731*RF+0.00471*PF+ 0.00161*QL+0.000966*SL

Table 9.1: Formulae for Estimating EQ-5D-3L Utilities

Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF), Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI), Financial Problems (FI); Global Health Status Score (QL); *using model developed from SOCCAR data to test on TOPICAL data.

The algorithm details were provided in section 2.4.5 (Chapter 2).

HRQoL Instruments

The EORTC QLQ-C30 and EQ-5D-3L were used for this simulation. EQ-5D-5L were not available for the TOPICAL and SOCCAR studies. These instruments were discussed in detail in earlier chapters.

Data

Each of the published algorithms was applied to three data sets from NSCLC studies: The TOPICAL trial (N=670); SOCCAR trial (N=130) [165,166] and Study 3 (N=100). Details of each study are in Chapter 3. ,

	Observed			Simulated	
	TOPICAL [N=2038]*	SOCCAR [N=1002]	Study 3 [N=985]	TOPICAL [N=2038]	SOCCAR [N=1002]
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
EQ-5D	**0.61 (0.29)**	**0.75 (0.23)**	**0.52 (0.31)**	**0.61 (0.28)**	**0.75 (0.22)**
PF	54.15 (26.30)	78.60 (36.53)	56.25(26.81)	54.15 (25.10)	77.90 (37.88)
RF	49.18 (34.67)	71.28 (43.04)	46.00 (33.42)	49.11 (33.87)	70.99 (42.11)
EF	73.89 (24.41)	77.06 (42.23)	66.34 (27.40)	72.99 (23.91)	77.16 (44.19)
SF	66.90 (32.40)	74.55 (45.17)	52.54 (32.50)	66.10 (32.30)	74.35 (46.10)
CF	10.00 (21.02)	20.44 (49.53)	71.25 (28.92)	10.05 (20.92)	20.99 (50.13)
NV	10.65 (18.81)	10.58 (37.26)	13.25 (20.96)	9.99 (19.01)	9.98 (36.18)
PA	26.00 (29.70)	20.66 (41.47)	38.90 (30.97)	25.90 (28.90)	20.05 (43.45)
DY	49.76 (32.92)	31.57 (45.98)	52.27 (31.75)	49.10 (31.62)	31.98 (47.13)
SL	29.03 (32.67)	24.00 (42.70)	46.39 (34.64)	29.66 (32.00)	23.98 (41.66)
AP	36.78 (35.57)	20.47 (43.61)	38.20 (36.45)	36.12 (35.11)	19.97 (43.53)
CO	20.33 (28.21)	18.64 (46.78)	26.81 (33.28)	20.83 (27.91)	18.66 (45.99)
DI	17.14 (27.54)	6.06 (40.62)	11.64 (23.30)	17.14 (26.14)	6.59 (43.92)
FA	48.52 (28.89)	33.94 (42.20)	53.78 (28.47)	47.92 (28.89)	32.99 (41.44)
FI	77.17 (24.41)	83.13 (42.32)	28.48 (34.26)	78.17 (29.30)	84.09 (46.18)
QL	52.27 (23.25)	63.75 (29.65)	51.19 (24.16)	52.17 (25.11)	63.22 (28.89)

Table 9.2: Resultant Simulated Mean (SD) Compared to Observed Values

Physical Function (PF), Role Function (RF), Emotional Function (EF), Cognitive Function (CF), Social Functioning (SF, Fatigue (FA), Nausea & Vomiting (NV), Pain (PA), Dyspnoea (DY), Insomnia (IN), Appetite Loss (AL), Constipation (CO), Diarrhoea (DI), Financial Problems (FI), Global Health Status Score (QL).

9.2.2 Testing each of the Published Algorithms on Independent lung cancer Data

For each of the (nine) published algorithms, the observed QLQ-C30 (patient level) domain scores from the 3 lung studies (separately) were substituted into each of the algorithms (Table 9.1). The observed EQ-5D-3L values were then compared to the predicted values (by regressing observed versus predicted) to generate R^2, RMSE and percentage predicted within + 10% of the observed EQ-5D (i.e. if the observed EQ-5D utility is 0.50, then a predicted value to within +10% would be 0.45 to 0.55). Since most predicted values are likely to lie within +20% and higher and very few to be within +5%, It was felt that +10% was an acceptable criterion because several one and two-way sensitivity analyses in economic evaluations, utility increments of higher than +10% have been used [206].

All available coefficients were used for prediction, regardless of statistical significance. A non-statistically significant predictor does not necessarily mean irrelevance or it is not important (e.g. due to lack of statistical power) [183]. There appear to be no caveats, which suggest that non-statistically significant coefficients should be excluded for the purposes of predictions [183]. In practice, when users apply an algorithm, they may not be able to use their observed QLQ-C30 data if some coefficients are not reported (loss of information).

Comparing by Health States

The mean predicted EQ-5D was computed for each (ordered) health states (from 11111 to 33333). The predicted values from each algorithm were compared with the observed EQ-5D-3L values from each of the 3 studies.

9.2.3 Classification of Algorithms

Algorithms were classified using a 'traffic light' system and a decision chart to guide the algorithm selection. Since each set of predicted utilities corresponds to one data set (TOPICAL, SOCCAR, Study 3), the use of a single estimate of a statistic such as R^2, or percentage of predicted values within $\underline{+}$10% of the observed will be uncertain. Therefore, in order to classify algorithms based on values of R^2, RMSE, and percentage predicting within $\underline{+}$10%, 1,000 bootstrap simulations were used to generate a distribution of these statistics. Each (simple) bootstrap sample consists of the same number of observations as the original study (e.g. the SOCCAR study had n=130 patients, and therefore a bootstrap size would be 130). For each bootstrap sample, the (QLQ-C30) domain scores were used as inputs into each of the published algorithms to generate patient level predicted EQ-5D-3L utilities. The observed and predicted utilities from each bootstrap sample were then used to derive R^2, RMSE and % predicted to within $\underline{+}$10% and further used to classify published algorithms as:

(i) Green: 'Useful'.

(ii) Amber: 'Useful, but caution needed'.

(iii) Red: 'Poor'.

The cut-off values for (i) to (iii) were determined by bootstrap simulation [228,229]. For each study (data set), 1,000 estimates of the above statistics (R^2, MSE, and % predicted) were generated, along with their (bootstrap) mean. An estimate of the overall mean (μ) based on ($\hat{\mu}$) across all algorithms (and data sets) was derived for each metric: R^2, RMSE and %predicted.

For each algorithm, for each of the three categories/classifications (i) to (iii), the following general conditions were used:

$\theta^*_{.j} > \hat{\mu}$ for R^2, % predicted [**Useful**] **[9.1]**

$\theta^*_{.j} < \hat{\mu}$ for RMSE : **[Useful]** **[9.2]**

or

$\theta^*_{.j} < \hat{\mu}$ for R^2, % predicted [**Poor]** **[9.3]**

$\theta^*_{.j} > \hat{\mu}$ for RMSE : **[Poor]** **[9.4]**

Where $\theta^*_{.j}$ represents the true mean parameter value across the 3 studies, for each of the j=1…9 algorithms, for a given metric (R^2, RMSE, %predicted). Hence there are, $\theta^*_{.1}, \theta^*_{.2}, \dots\dots..\theta^*_{.9}$ values, one for each algorithm. The table (Table 9.3) below clarifies further.

Algorithm (j)		**R^2**				**RMSE**			**%Pred**	
Dataset	**1**	**2**	**3**	**Average**	**1**	**2**	**3**	**1**	**2**	**3**
1	θ_{11}	θ_{12}	θ_{13}	$\boldsymbol{\theta^*_{.1}}$	etc.	etc	Etc	Etc	etc	etc
2	θ_{21}	θ_{22}	θ_{23}	$\boldsymbol{\theta^*_{.2}}$	etc	etc	Etc	Etc	etc	etc
3	θ_{31}	θ_{32}	θ_{33}	$\boldsymbol{\theta^*_{.3}}$	etc	etc	Etc	Etc	etc	etc
4	θ_{41}	θ_{42}	θ_{43}	$\boldsymbol{\theta^*_{.4}}$	etc	etc	Etc	Etc	etc	etc
5	θ_{51}	θ_{52}	θ_{53}	$\boldsymbol{\theta^*_{.5}}$	etc	etc	Etc	Etc	etc	etc
6	θ_{61}	θ_{62}	θ_{63}	$\boldsymbol{\theta^*_{.6}}$	etc	etc	Etc	Etc	etc	etc
7	θ_{71}	θ_{72}	θ_{73}	$\boldsymbol{\theta^*_{.7}}$	etc	etc	Etc	etc	etc	etc
8	θ_{81}	θ_{82}	θ_{83}	$\boldsymbol{\theta^*_{.8}}$	etc	etc	Etc	etc	etc	etc
9	θ_{91}	θ_{92}	θ_{93}	$\boldsymbol{\theta^*_{.9}}$	etc	etc	Etc	etc	etc	etc
				$\boldsymbol{\hat{\mu}}$	etc	etc	Etc	etc	etc	etc

Table 9.3: Example of simulated data structure

In Table 9.3, θ_{ij} is the statistic derived for data set i for algorithm j, $\theta^*_{.j}$ is the average value of the statistic across the 3 data sets for a given algorithm j, and $\hat{\mu}$ is an estimate of the true overall mean across all algorithms and data sets. All estimates are based across the bootstrap values. The differences between ($\theta^*_{.j}$ - $\hat{\mu}$) in **[9.1]** and **[9.4],** provides an indication of how far each estimate from each algorithm across all studies/datasets ($\theta^*_{.j}$) is compared to the overall mean value ($\hat{\mu}$) across all studies and algorithms.

Define $\Omega_j = (\theta^*_{.j} - \hat{\mu})$, then, one would desire $\Omega_j \geq 0$ for metrics such as R^2 and %predicted (the algorithm j would be above average performance for such a metric). Conversely, one may desire $\Omega_j < 0$ for metrics such as RMSE. One could also choose values Ω_j of other than 0. In addition, one would desire the (empirical) probability of $\Omega_j \geq 0$ to be high or at least some value Δ.

More formally,

$$\Pr[(\Omega_j \geq 0) | \Lambda, A_i] \geq \Delta \text{ , for } \Lambda \in \{R^2 \text{ , \% predicted}\} \quad \textbf{[9.5]}$$

$$\Pr[(\Omega_j < 0) | \Lambda, A_j] \geq \Delta \text{ , for } \Lambda \in \{MSE\} \quad \textbf{[9.6]}$$

In this example, demonstration of the selection method, Δ is set to 0.5. This corresponds to at least a 50% chance (a fifty-fifty chance in the absence of additional knowledge) the selected algorithm based on the statistic of interest has above average performance. One could have a classification of algorithms such that there was at least 80% chance a given algorithm is classified as ‘Useful’. However, the higher the probability, the fewer algorithms are likely to be classified. Table 9.4 (below) outlines the possible outcomes from equations **[9.1] to [9.4].** The criteria attempts to compare each algorithm compared to the overall mean across all 8 algorithms and all 3 data sets. Hence the criteria is set against the current mean of any particular statistic. Further details for the derivation of these criteria are now developed.

R^2	**RMSE**	**%Predicted**	**Overall**
Useful [10.1]	Useful [10.2]	Useful [10.1]	Useful (Green)
Any 2 are ‘Useful’ (or any 1 Poor)			Useful/Caution (Amber)
Any 2 are ‘Poor’ (only 1 ‘Good’)			Poor (Red)
Poor [10.3]	Poor [10.4]	Poor [10.3]	Poor

Table 9.4: Summary of Decision Outcomes Based on Model Statistics

9.2.4 Methodology for Algorithm Selection

The criteria in Table 9.4 essentially classify the algorithms based on the bootstrap means across all three studies. If the individual parameters ($\theta^*_{.j}$) from each algorithm based on observed data are above or below the expected true parameter ($\hat{\mu}$) value, using data across all 3 (bootstrapped) studies, the algorithms are appropriately classified depending on whether the relative frequency (probability) the statistic (e.g. RMSE) is above or below the bootstrap average. The criteria may be correlated (i.e. if RMSE is low and R^2 is high, then the %predicted within ±10% may also be high).

Technical Details of Selection Criteria

Assume A_1, A_2…….A_n (j= 1 to n) algorithms for testing (in this case, n=9). Assume there are S_1, S_2...S_m data sets (with equal or unequal sample sizes) to test each algorithm (m= 3 in this example, as there are 3 data sets).

After each algorithm (A_j) has been tested on each data set (S_i), a set of parameters θ_{ij} are estimated. Each θ_{ij} could refer to true values of model fit statistics (**Λ),** R^2, RMSE, and % predicted (in this instance). The matrix **Z**, is a matrix of dimension m x n parameter values (in this case for n=9 algorithms and m= 3 data sets, there would be a 3x9 matrix for a single metric, with each element θ_{ij}.

$$\mathbf{Z} = \begin{pmatrix} \theta_{11} & \theta_{12} \dots \theta_{19} \\ \theta_{21} & \theta_{22} \dots \theta_{29} \\ \theta_{31} & \theta_{32} \dots \theta_{39} \end{pmatrix}$$

However, since each θ_{ij} is an estimate for each of (R^2, RMSE and %predicted), this becomes a 3 x 9 x 3 dimension data structure. We now compute an estimate of each parameter for each algorithm across the m data sets.

$$\theta^*_{.j} = 1/m(\textstyle\sum_{j=1}^{m} \theta_{.j}) \qquad \textbf{[9.7]}$$

Therefore, $\theta^*_{.j}$ is an estimate of the true mean value of the metric of interest across the m (m=3) studies for a given algorithm j. Hence, there are several estimates, $\theta^*_{.1}, \theta^*_{.2}, \dots\dots..\theta^*_{.9}$, one for each algorithm, in this case.

Simulate from each study, S_m, (m=1…3), z bootstrap samples (e.g. z=1,000). For each bootstrap sample, use the algorithm A_j to predict the patient level utilities and consequently, an estimate of model fit statistic of interest (e.g. R^2). For example, with 1,000 bootstrap samples from each study, this would yield 3,000 R^2 values for each algorithm (hence, 3,000 x 9 estimates of R^2).

Define each bootstrap estimate of each parameter:

$$\omega_{ijk} \text{ (k= 1 ….z)} \qquad \textbf{[9.8]}$$

Compute the mean across all bootstrap estimates:

$$\hat{\mu} = \textstyle\sum_{j=1}^{3} \quad \sum_{i=1}^{9} \sum_{k=1}^{z} \quad \omega_{ijk}. \qquad \textbf{[9.9]}$$

This ($\hat{\mu}$) is an estimate of the population average (μ) for each of R^2, RMSE and %predicted. The value of $\hat{\mu}$ provides the basis for the cut-off value; **[9.7]** is as compared to **[9.9].**

Set the criteria that if $\boldsymbol{\theta}^{*}_{.j} < \hat{\mu}$ or $\boldsymbol{\theta}^{*}_{.j} > \hat{\mu}$ for a given metric, then the classification of each algorithm is generated as shown in Figure 9.1. This is the simplest classification criteria used, which essentially states that if the estimates of a metric (e.g. R^2) across studies are lower or higher than the expected population average, algorithms can be classified accordingly.

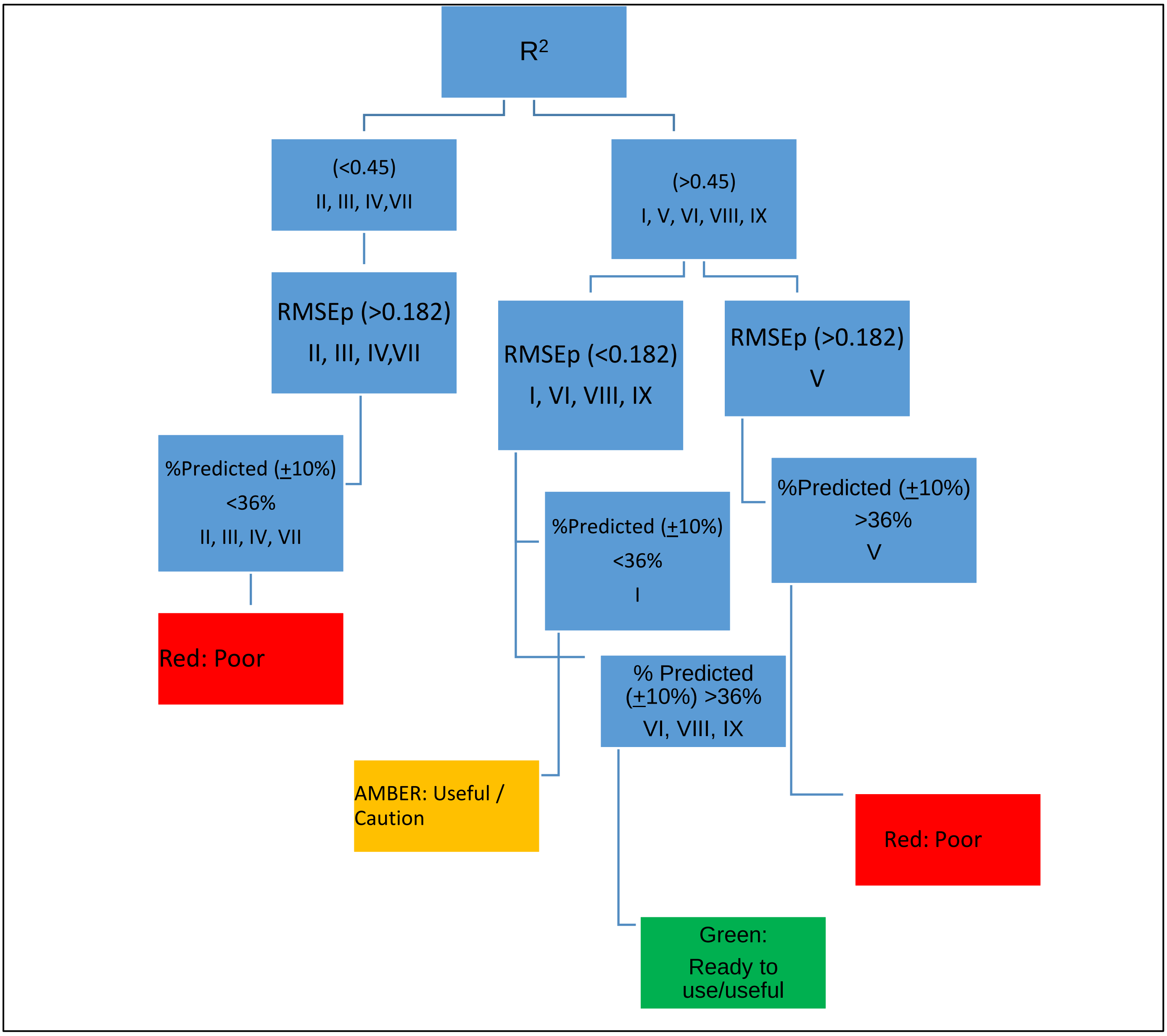


Figure 9.1: Example Classification Tree for Published Algorithms (Across All Studies)

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al., (2010); IV: Crott et al., (2012); V:Kim-k (2012); VI:Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014) using SOCCAR data; IX: Proskorovsky (2014); RMSEp: RMSE Predicted

Further justification can be shown as follows. The differences between **[9.7]** and **[9.9],** i.e. $\Omega_j = (\theta^*_{.j}$ - $\hat{\mu}$), provides an indication of how far the observed estimate is across all 3 studies compared to the expected population estimate ($\hat{\mu}$). For large positive values of Ω_j, for metrics such as R^2 and % predicted one would like this to have high probability ($\geq \Delta$) for 'useful' algorithms (Ω_j >>0) or (Ω_j << 0) for metrics such as RMSE, based on **[9.5] and [9.6]** below.

Pr [($\Omega_j \geq 0$) |Λ, A_j] $\geq \Delta$, for $\Lambda \in \{R^2$, % predicted$\}$ **[9.5]**

Pr [($\Omega_j < 0$) |Λ, A_j] $\geq \Delta$, for $\Lambda \in$ {MSE} **[9.6]**

The equations **[9.5]** and **[9.6]** are interpreted as the difference between the average R^2 across all available data sets when compared with the average R^2 values across all data sets and algorithms across all simulations. The value of Ω_j should be positive (i.e. $\Omega_j \geq 0$), for at least 50% of the time, so that algorithms can be considered useful, otherwise they are not considered useful. A similar interpretation is made for the % predicted and the inverse is used for RMSE. Hence, the chosen cut-offs for R^2, MSE and % predicted reflect at least a 50% chance a given algorithm is classified as 'Poor' or 'Useful'. A summary of the terminology and definitions are found in Table 9.5.

Parameter	**Details**
A_j	Each Aj is an algorithm (j= 1 to n); n=9
S_i	Each Si is a study (i= 1 to m); m=3
M	The number of data sets; m=3
N	The number of algorithms; n=9
Λ	The set of model fit statistics of interest.
θ_{ij}	The parameter of interest for algorithm i and study j.
$\theta^*_{.j}$	Mean of parameters for each algorithm j across the m data sets for a given statistic in Λ
Z	The matrix of m x n estimates for each of the elements in Λ
z	The number of bootstrap samples.
$\boldsymbol{\omega_{ijk}}$	Each bootstrap mean (k) for each algorithm (j) and each data set (i).
$\hat{\boldsymbol{\mu}}$	The estimate of the true mean across all algorithms and all studies for a given value of Λ
$\boldsymbol{\Omega_j}$	The difference between the mean of parameters for each algorithm and the overall mean: ($\theta^*_{.j}$ - $\hat{\mu}$).

Table 9.5: Parameter Description

9.2.5 Method Used to Simulate Data from TOPICAL and SOCCAR Trials

The data were simulated from a multivariate normal distribution (MVN) using summary statistics reported earlier [109]. The data were assumed to be multivariate normal. It is not uncommon for QLQ-C30 scores to be reported in terms of mean and standard deviations for describing treatment effects, an assumption widely noted in the literature [155]. The raw data were no longer available and therefore simulations were required. The method of simulation was as follows:

(i) Estimate the mean and SD of each the 15 QLQ-C30 domain scores and EQ-5D, for each of TOPICAL and SOCCAR. These were readily available [109]. A total of 2,038 and 1,002 observations were simulated for each data set (as these were the original number of observations in each trial).

(ii) Estimate or make assumptions about the correlations between each EQ-5D and QLQ-C30 domain score. These were assumed to be as shown in Table 4.7 and 4.8 (Chapter 5) published as supplementary tables [109]. The assumption for the correlation matrix is justified because the final estimates of the summary Mean and SDs from simulated data approximate closely with those observed in Khan and Morris [109]. The general approach to simulation is:

(iii) Using the Kaiser and Dickman (1962) approach:

$$Z_{kxN} = F_{(kxk)}\ X_{(kxN)}$$

Where Z is the resultant observation (matrix of observations).

F is the correlation matrix.

X is the input data consisting of means and SDs.

kxN represents the number of variables(k the 15 QLQ-C30 plus one EQ-5D variable).

The following SAS code was used to construct the simulated data sets:

```
Data sim (type=corr;
_type_='CORR';
Input x1-x16;
1.00  etc <correlations>;
Run;
Proc factor n=16;
run;
```

The resulting mean and SD's (Tables 9.2) were very similar to those observed in Chapter 4. Since all observations are sometimes pooled in the development and applications of mapping algorithms, this type of simulation approach is unlikely to be problematic for the testing of published algorithms.

9.3 Results

Nine algorithms were used to estimate EQ-5D-3L utilities from the observed QLQ-C30 data (Tables 9.1).

Demographic and Clinical Characteristics

Table 9.6, 9.7 and 9.8 shows a summary of the data characteristics from each algorithm. Only one previous algorithm [130] was developed using a NSCLC data set. All algorithms were OLS type models (except IX). Sample sizes used for developing algorithms ranged between 48 (algorithm II) to 893 (VI); Reported R^2 (Table 9.9) ranged between 0.49 (V) to 0.91 (IV) and RMSEs were between 0.09 (VIII) to 0.19 (II). Coefficients of each reported algorithm were shown in Table 9.1 and 9.10.

Year [Reference]	**Sample Size**	**Population**	**Model Type**	**R^{2a}**	**validation**
2016 Khan et al. (IX)	100	NSCLC	2 Part Beta OLS	75%	Cross validation & Simulation
2009 McKenzie et al. (I)	199	Oesopgeal	OLS PROBIT	61%	Independent data
2009 Kontodimopoulos N (II)	48	Gastric	OLS	91%	None
2010 Jang et al. (III)	172	NSCLC	OLS	58%	Cross validation
2010 Crott and Briggs (IV)	448	Breast	Quadratic	80%	Independent data
2012 Kim EJ et al. (V)	199	Breast Cancer	OLS	49%	Cross validation
2012 Kim SH et al. (VI)	893	Mixed	OLS	52%	Independent data
2012 Versteegh et al. (VII)	137	Mixed	OLS	82%	Independent data
2014 Proskovorsky (VIII	154	Myeloma	OLS	69%	None

Table 9.6: Selected Algorithms

[a]maximum observed was reported

	TOPICAL	**SOCCAR**	**Study 3***
Sample Size	670	130	98
Observations[#]	2038	1002	985
Age (Range) [years]	77 (42-91)	63 (36-78)	69 (39-86)
Gender			
Male	409 (61%)	79 (61%	55 (56%)
Female	261 (39%)	51 (39%)	43 (44%)
ECOG:			
0: Normal activity	12(2%)	10(17%)	12 (12%)
1: Near full activity	94(14%)	50(83%)	23 (23%)
2: In bed < 50% of time	372(56%)	0	30 (31%)
3: In bed > 50% of time	192(29%)	0	27 (28%)
4: Totally confined to bed	0		4 (4%)
# Health States for EQ-5D-3L	85	54	62
Stage			

Stage I-II	0	0	26 (27%)
Stage III	234	57 (44%)	31 (32%)
Stage IV	436(65%)	73 (56%)	37 (38%)
Histology			
Adenocarcinoma	256(38%)	35(27%)	43 (44%)
Squamous	263(39%)	83(64%)	36 (33%)
Large Cell	30(5%)	0	0
Other NSCLC	121(18%)	12(9%)	19 (23%)

Table 9.7: Baseline Characteristics for Each Study

[#] EQ-5D observations; *observational study designed for this thesis

Table 9.7: Baseline Characteristics for Each Study

Algorithm	Sample Size	# Coefficients reported for prediction	Population (Cancer)	Model
I	199	15	oesophageal	Linear (OLS)
II	48	3	Gastric	Linear (OLS)
III	172	15	NSCLC	Linear (OLS)
IV	798	12	Breast	Linear (OLS)
V	149	5	Breast	Linear (OLS)
VI	893	11	Myeloma	Linear (OLS)
VII	723	5	Colon	Linear (OLS)
VIII	800	15	NSCLC	Non-Linear (BB)
IX	154	15	Myeloma	Linear (OLS)
SOCCAR	130	15	NSCLC	Non-Linear (BB)
TOPICAL	670	15	NSCLC	Non-Linear (BB)

Table 9.8: Baseline and Data Characteristics for Each Algorithm

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al. (2010); IV: Crott et al. (2012); V: Kim-k (2012); VI: Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014); IX: Proskorovsky (2014)

Note: VIII (Khan 2014) was developed from SOCCAR data and tested on TOPICAL data and vice versa.

	Reported/ Published		**TOPICAL [N=670, #=2038]**			**SOCCAR [N=130, #=1002]**			**Study 3 [N=100, #=985]**		
	R^2_R	**$RMSE_R$**	**R^2_P**	**$RMSE_P$**	**$Mean_P$**	**R^2_P**	**$RMSE_P$**	**$Mean_P$**	**R^2_P**	**$RMSE_P$**	**$Mean_P$**
I	0.61	NR	0.59	0.19	0.5685	0.57	0.151	0.6959	0.50	0.2193	0.492
II	0.61	0.192	0.54	0.20	0.6013	0.56	0.152	0.8056	0.42	0.2360	0.597
III	0.58	NR	0.19	0.19	0.6571	0.56	0.152	0.7691	0.60	0.1963	0.638
IV	0.80	0.096	0.46	0.22	0.7250	0.22	0.203	0.7861	0.54	0.211	0.640
V	0.49	NR	0.59	0.19	0.8234	0.55	0.154	0.9093	0.60	0.196	0.808
VI	0.74	0.11	0.58	0.19	0.6229	0.58	0.150	0.7209	0.55	0.209	0.580
VII	0.52	0.087	0.59	0.19	1.0830	0.59	0.148	1.1950	0.53	0.213	1.041
VIII	0.75	0.09	0.58	0.19	0.622	0.61	0.159	0.7448	0.57	0.230	0.539
IX	0.70	0.164	0.56	0.20	0.5106	0.51	0.161	0.6783	0.58	0.201	0.522

Table 9.9: Summary of Algorithm Performance on Independent Data

Note: Observed Means were 0.61, 0.75 and 0.52 for TOPICAL, SOCCAR and Study 3 respectively

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al. (2010); IV: Crott et al. (2012); V:Kim-k (2012); VI:Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014) using SOCCAR data; IX: Proskorovsky (2014); $^{+}$as reported in published model; NR: Not Reported; $\mathbf{R^2_R}$: R-squared reported; $\mathbf{R^2_p}$: R-squared predicted; **RMSE**$_{\mathbf{R}}$: Root Mean Square Error reported (R) ; **RMSE**$_{\mathbf{P}}$ Root Mean Square predicted (P); Note: VIII (Khan 2014) was developed from TOPICAL data and tested on SOCCAR data and vice versa; $^{\#}$: number of observations

Selection of Algorithm Results

The Bootstrap estimates of cut-offs ($\hat{\mu}$) for each of R^2, RMSE and % predicted within $\underline{+}$10% were 45%, 0.182 and 36% respectively (Figure 9.1 and Tables 9.9, 9.10 & 9.11). Therefore, algorithms were classified as:

(i) **Green- Useful:** VI: Veerstegh (2012) [141]; VIII: Khan (2014) and IX: Proskorovsky (2014), because these reported R^2 >45%, better (highest) proportion of prediction of EQ-5D to within $\underline{+}$10% of the observed mean EQ-5D and relatively lower RMSE (Figure 9.2 & Table 9.6).

(ii) **Amber**- **Useful/Caution:** Algorithm I is considered ‘Amber’ since it had some useful properties as well as some potential for moderate over/under prediction: R^2 >45%, but higher RMSE and % predicted within $\underline{+}$10% were <36%.

(iii) **Red – Avoid:** Algorithms II, III, IV, V, and VII should not be the first choices. These had typically lowest R^2, highest RMSE, and poorer prediction. These algorithms may yield unreliable estimates of utilities and consequently, QALYs.

Tables 9.11 & 9.12 show % of predicted EQ-5D scores within +5% and +10% of the observed.

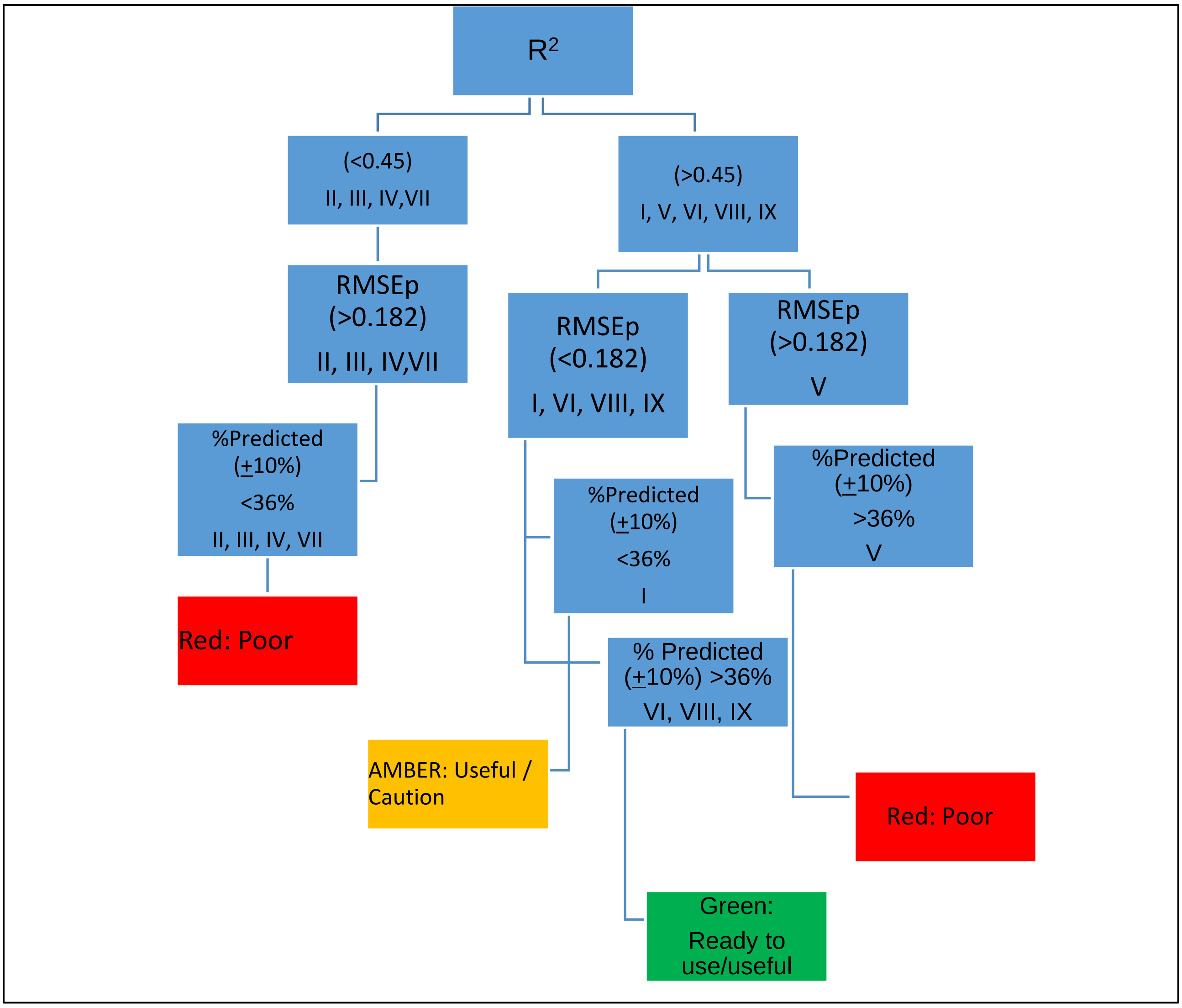


Figure 9.2: Summary of Algorithm Performance (Across All Studies)

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al., (2010); IV: Crott et al., (2012); V:Kim-k (2012); VI:Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014) using SOCCAR data; IX: Proskorovsky (2014); RMSEp: RMSE Predicted

QLQ-C30	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII***	**IX**
PF	0.0004	0.00508	0.0035	-0.006963 [0.0000355]*	0.0032	0.0032	0.002	0.260	0.00471
RF	0.0022		0.0007	NR	NR	NR	0.002	0.234	0.000731
EF	0.0028	0.00313	0.0011	-0.008734 [0.0000552]*	0.0005	0.0005	0.003	0.379	0.00104
SF	0.0002	NR	0.0007	-0.003993 [0.0000290]*	0.0005	0.0005	0.001	0.257	0.0000447
CF	0.0009	NR	0.0007	NR	NR	NR	NR	0.061	-0.000378
FA	0.0021	NR	0.0003	NR	NR	NR	0.001	-0.012	-0.0000575
NV	0.0005	NR	-0.0002	NR	NR	NR	NR	0.062	0.000573
PA	0.0024	NR	-0.0021	0.003561	-0.0013	-0.0013	-0.001	-0.235	-0.00229
DY	0.00004	NR	-0.0001		0.0008	0.0008	NR	0.0088	0.000376
SL	0.00004	NR	-0.0001	-0.0003678 [0.0000117]*	NR	NR	NR	-0.012	0.000966
AP	0.0003	NR	-0.0001	NR	NR	NR	NR	0.0017	-0.000479
CO	0.0001	NR	0.0005	0.001145	NR	NR	-0.001	0.026	-0.0005519
DI	0.0003	NR	0.0004	[-0.0000540]*	0.005	0.0005	NR	0.051	-0.0003
FI	0.0006	NR	-0.0001	NR	NR	NR	NR	0.062	0.000961
QL	0.0016	0.00546	0.0009	NR	0.0007	0.0007	0.001	0.224	0.00161
Constant	0.0004	0.00508	0.0035	-0.006963 [0.0000355]*	0.0032	0.0032	0.002	1.318	0.00471

*Using model developed from TOPICAL data and tested on SOCCAR data

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al. (2010); IV: Crott et al. (2012); V:Kim-k (2012); VI:Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014); IX: Proskorovsky (2014)

Table 9.10: Coefficients of Published Mapping Algorithms

Algorithm	**R^2_P**	**$RMSE_P$**	**% of predicted within ±10%**	**Overall**	**Interpretation**
I	0.51 (U)	0.17 (U)	31% (P)	Useful/Caution	Amber: Useful, but caution
II	0.39 (P)	0.19 (P)	23% (P)	Poor	Red: Avoid in the first instance
III	0.14 (P)	0.21 (P)	29% (P)	Poor	Red: Avoid in the first instance
IV	0.32 (P)	0.19 (P)	19% (P)	Poor	Red: Avoid in the first instance
V	0.44 (U)	0.19 (P)	23% (P)	Poor	Red: Avoid in the first instance
VI	0.58 (U)	0.17 (U)	39% (U)	Useful	Green: Ready to use/useful
VII	0.31 (P)	0.20 (U)	34% (P)	Poor	Red: Avoid in the first instance
VIII	0.57 (U)	0.17 (U)	48% (U)	Useful	Green: Ready to use/useful
IX	0.61 (U)	0.17 (U)	55% (U)	Useful	Green: Ready to use/useful
Mean	**0.45**	**0.182**	36%		

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al. (2010); IV: Crott et al. (2012); V:Kim-k (2012); VI: Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014); IX: Proskorovsky (2014); Note: VIII (Khan 2014) was developed from TOPICAL data and tested on SOCCAR data and vice versa

Table 9.11: Results from Bootstrap Simulations (Averaged across All 3 Studies)

	TOPICAL			SOCCAR			Study 3		
	% within +5%	% within +10%	Mean Predicted	% within +5%	% within +10%	Mean Predicted	% within +5%	% within +10%	Mean Predicted
I	8	25	0.5685	9	29	0.6959	6	28	0.4924
II	9	16	0.6013	9	24	0.8056	6	22	0.5972
III	12	24	0.6571	15	32	0.7691	5	29	0.6380
IV	6	19	0.7250	10	25	0.7861	4	19	0.6409
V	6	22	0.8234	7	26	0.9093	3	18	0.8089
VI	10	31	0.6229	14	40	0.7209	6	44	0.5808
VII	11	32	1.0830	12	34	1.1950	9	30	1.0410
VIII	12	46	0.6223	15	46	0.7448	14	41	0.5399
IX	10	40	0.5106	16	44	0.7503	8	39	0.5224
Observed		**0.61**			**0.75**			**0.52**	

Table 9.12: Mean Predicted EQ-5D within +5% and +10% of Observed

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al. (2010); IV: Crott et al. (2012); V:Kim-k (2012); VI:Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014) using SOCCAR data; IX: Proskorovsky (2014); U: Useful, P: Poor

Prediction at Poorer Health States

Figure 9.3 shows the predicted mean utilities for each health state for the EQ-5D-3L.

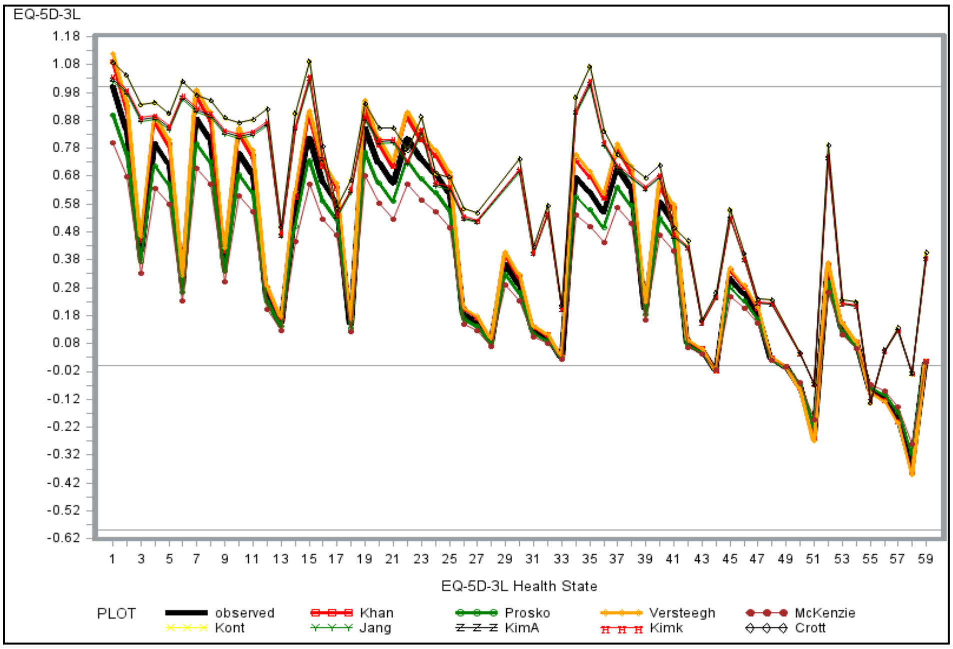


Figure 9.3: Predicted EQ-5D-3L vs. Observed EQ-5D for Each Algorithm by Health State (Study 3)

I: McKenzie (2009); II: Kontodimopoulos (2009); III: Jang et al., (2010); IV: Crott et al., (2012); V: Kim-k (2012); VI: Veerstegh (2012); VII: Kim-A (2013); VIII: Khan (2014); IX: Proskorovsky (2014); observed: bold black line

The x-axis is the ordered health state: in this case, 1 refers to the observed health state 11111 and 59 to 22322. Algorithms perform better if predictions are close to the observed (solid black line).The overprediction at poorer health states is observed for some algorithms (e.g. algorithm I). This is shown for algorithms VI, VIII, and IX. Algorithms II, III, IV, V, and VII show marked a departure from the observed health states and are broadly consistent with Figure 10.1. Some algorithms predict above 1 (outside the range).

This analysis confirms a feature of mapping algorithms noted previously [97] - the over-prediction of utility at poorer health states. Figure 9.3 shows predictions overall for each health state using Study 3 data. Most models result in over-prediction at some poorer health states. One reason for the poorer EQ-5D-3L predictions at worsening health states may be related to the deteriorating HRQoL post disease progression or some other mechanism associated with progression such as safety/toxicity (noted in earlier chapters).

9.4. Discussion

In this chapter, an approach to categorizing mapping algorithms based on several criteria have been presented. Algorithms were broadly classified as Green: 'useful'; Orange/Amber:' Useful with caution' and Red: 'Avoid in the first instance'. The overprediction at poorer health states and usefulness of a three-part (Beta Binomial) model in comparison with other models was confirmed. The estimates of predicted EQ-5D by health states also appear to agree in general, with other measures such as R^2, RMSE and predicted mean EQ-5D values. The approach presented here was based on an empirical approach, using bootstrap simulations to define the cut-off values.

Doble and Lorgelly (2016), Arnold et al., (2015) and Crott (2014) confirm that the selection and conclusion as to the 'best' model may be arbitrary and suggests further research into identifying the best model [72,163]. The approach presented here seems arbitrary as the metrics chosen for considering one algorithm more (or less) favourable over another is based on an arbitrary selection of R^2, RMSE and % predicted and an arbitrary selection of 50% as a probability threshold of a useful algorithm. However, the idea of classifying an algorithm based on the *probability* of the selected metric being close to the overall population value across all data (using simulation) is less arbitrary because each algorithm is judged against the overall 'average' performance after quantifying uncertainty in the choice.

Arnold et al., (2015) conclude that for the model by Longworth et al. (2012) [162], using a response modelling approach was 'best'. The findings here differ with those of Arnold et al. (2015) [162] in terms of their classification of the 'best' algorithm. The findings presented suggest algorithm III [130] is a less useful algorithm while algorithm I [132] is 'useful with caution'. Differences in demographic and tumour characteristics across studies may partially explain differing conclusions in algorithm performance.

There are several possible reasons, which might explain how the findings differ in their classification of the nine algorithms. Firstly, the original algorithms were generated from data that are different from the target patient population data set to which the algorithm is applied, such as, differences in baseline characteristics, prognoses, and gender. For instance, an algorithm developed from a breast cancer population applied to prostate cancer might seem inappropriate. Interestingly, the more ‘useful’ algorithms were not necessarily developed from lung cancer data, yet worked reasonably well when applied to a NSCLC data set (e.g. algorithm IX). Consequently, algorithms may not need to be disease-specific to be ‘useful’. Further research is required in order to establish whether the differences between the predicted vs. observed estimates are entirely due to the performance of the statistical/mathematical form of the algorithm or because the data used to predict the estimates are from entirely different cancer populations/tumour types, with quite different demographic and disease characteristics. A large dataset across multiple geographical locations, clinical characteristics, and tumour types is required and is a current research topic.

A second reason for the differences in algorithm performance is that about 50% of the algorithms (5/9) reported all 15 coefficients. Although coefficients may not be important (statistically), it is still possible that they have a role in predicting EQ-5D utilities in independent data sets - particularly, if the predictions are for a different patient population. Consequently, differences between algorithm performances may be due to the number of coefficients reported. Some coefficients (e.g. for treatments) will differ from trial to trial but are expected to influence the utility in different ways; these coefficients are rarely reported. Utilities are therefore predicted from available coefficients.

Thirdly, predictions are likely to depend on the differences in data ranges between the data used to develop the algorithm and the data used for testing. For instance, if an algorithm is developed from data with a smaller range (e.g. EQ-5D values between 0 to 0.95) and applied to a data set, where there are many patients with particularly poor health prognoses and health states, whose utilities lie between (-0.549 to 1.0), then prediction will involve more extrapolation.

Fourthly, differences in performance between algorithms can also be due to the design of the study from which the data were collected (e.g. assessment points for measuring HRQoL). In some trials, the presence of toxic treatments, especially after disease progression where options are limited (e.g. taking 3rd line treatments), might have a pronounced impact on HRQoL (improve or worsen).

There are several limitations in this research. Firstly, the criteria used for classifying algorithms are dependent on the data. Secondly, some algorithms may have a high R^2 and % predicted but a poor QALY prediction. Additional criteria may help improve selection, but at the same time, these can be more complex with a higher chance of contradiction between criteria (although one might expect the criteria to agree in general). When this criterion was applied to the observed TOPICAL data, for instance, algorithm IV, of Crott and Briggs (2010) [111] classified as 'Poor' but generated a QALY and ICER closer to the observed ICER compared to other algorithms (£168,810 vs. £139,019). Similar is true for the algorithm I, which was considered 'Useful/caution' but which generated a large difference compared to the observed QALY. Thirdly, the choice of metric to simulate is arbitrary (one could have chosen MAE or AIC for instance). The 3 metrics (R^2, RMSE %predicted) are assumed as being equally important (equally weighted) for simplicity. Some priority ordering could be imposed on the importance of metrics.

A further question might be a concern that the algorithms assessed as green still have R^2 and %predicted values that seem low and whether these are sufficiently high to be useful for utility estimation. The answer to this is that the magnitude of these values are relative to the overall distribution of such statistics across all algorithms and the selected algorithms perform better based on the entire distribution. It is not uncommon to find published algorithms when tested on independent data, to result in low R^2 values [161-163]. Some of these authors have made similar recommendations for choices of algorithms based on values of R^2 commensurate with what is suggested here [161-163. This may show a problem with mapping detailed earlier in Chapter 6 on the sources of over or under prediction and the role the independent data sets play in validating or testing algorithms. The important thing is that the algorithms in this analyses which were reported as good (e.g. 91% R^2 as for algorithm II not tested in an independent dataset) are found to be quite poor when rigorously tested on multiple datasets through simulation. Hence the proposed method is cautious to attribute 'Green' to an algorithm when in fact it is poor which may be a fair price to pay for what appears to be a lower R^2 (recall that for non-linear models, R^2 may not be the best metric and given more models are tending to non-linear approaches, low R^2 may not be the only concern).

Only four of the algorithms tested on independent data resulted in R^2 values (Table 9.6) that ranged between 52% to 82% with an average R^2 of 69%. No rule has been stipulated on how high an R^2 should be for use in HTA, where algorithms have been used. Hence, values of R^2 of 58%, 57% and 61% (Useful algorithms) although low in absolute terms compared to other applications (e.g. biology), in this setting might be considered adequate.

Several other approaches could be used to select algorithms such as the *proportion* of simulations above or below the average values (rather than mean values). Another approach might be to adopt a Multi-Criteria Decision Analysis Approach using an Analytic Hierarchical Process [230,231] to decide between algorithms. This form of priority setting might help to formalize a choice for future algorithm use. An alternative approach might be to use a Delphi, or consensus conference, or Analytical Hierarchical Process, around the criteria of importance to define relevant thresholds (a possible area for future research). Further research is also needed using the suggested approach here to test the generalizability with other datasets. It is clear that as more algorithms are published the focus may turn to selection criteria methods. Finally, a type of 'super' algorithm could be developed if there is collaboration and data sharing between authors of algorithms to take into account the heterogeneity in clinical and other characteristics.

9.5. Conclusion

Despite the limitations of the suggested criteria, they may still offer a useful approach to selecting from a choice of mapping algorithms. This work is, therefore, a valuable addition to limited research in this area and may increase interest towards formal methodology in algorithm selection.

Chapter 10

Chapter 10: Summary and Conclusion

10.1 Overall Summary

This thesis has investigated approaches to modelling HRQoL for economic evaluation in lung cancer using data collected from a NSCLC population in routine NHS practice. To address the aims and objectives presented earlier, this thesis introduced a number of novel issues around the subject of modelling HRQoL data.

Chapters 1 to 3 presented a detailed literature review which identified several very important research areas that required necessary research in the development, testing, and validation of approaches to modelling HRQoL data for economic evaluation of lung cancer interventions. This also included the need for understanding the behaviour of post-progression utilities and extrapolation of utilities. In addition, useful information that contextualizes the economic burden of NSCLC and the QALY estimates were summarized. Very few lung cancer treatments satisfied the current NICE thresholds (of £20,000 - £30,000 per QALY) and a major component of this, related to modelling utility data for economic evaluation. In addition, a review of the HTAs of several important NSCLC drugs revealed important critiques of the cost-effectiveness models of which one consistent theme was handling and estimating HRQoL during and also after disease progression.

To this end, a prospective observational study was designed in an NHS setting to collect simultaneously HRQoL data, including EQ-5D-3L, EQ-5D-5L, and QLQ-C30 from study registration and beyond progression for at least 12 months. This small, but very rich data source will be invaluable for additional research in understanding HRQoL in NSCLC patients beyond this thesis.

Chapter 4 developed and tested a new non-linear approach to mapping utilities. A Beta Binomial model was used. This was the first use of a three-part non-linear model as a mapping algorithm for the QLQ-C30, which was able to model the over-dispersion well. Moreover, testing mapping through simulation was introduced - also a novel approach to assessing the uncertainty of mapping algorithms.

Chapter 5 (and also Chapters 6 and 7) presents the first research of its type to compare the EQ-5D-5L and the EQ-5D-3L. The conclusion reached was that the EQ-5D-5L consistently offered a better mapping algorithm (over the EQ-5D-3L) regardless of the statistical modelling approach. Moreover, this research may also explain that current limitations in mapping may be the measurement scale or at least a combination of both the functional form of the model and the scale (the 5L seemed to perform better, perhaps due to its extended scale). The scale may also explain the enhanced sensitivity and responsiveness observed in Chapter 8.

In Chapter 6, a further nuance to mapping was introduced by modelling the toxicity and utility together. No mapping algorithm based on examining the relationship between toxicity and utility has been developed. This is shown to be a promising area of research. In addition, Chapter 6 used real world data from a NSCLC population. Previous estimates of utilities for specific clinical or toxicity features were based on survey data and not real world clinical practice. These estimates are likely to be extremely useful for researchers undertaking future economic evaluation in a NSCLC population. Further research in this area could involve using the BB model to predict EQ-5D from modelling jointly the QLQ-c30 and toxicity to avoid prediction outside the range. Using Finite Mixture models (FMM) too might be useful to capture the relationship between HRQoL and other outcomes.

In Chapter 7, the first application of a Bayesian Network to both the EQ-5D-5L and the QLQ-C30 was used. Although Bayesian methods in mapping are still relatively novel, in this application, the Bayesian did not perform as well as expected. An interesting observation was that probabilistic type mapping could under-predict utilities at poorer health states – suggesting that patients may be much worse (on average) than observed. This was found to be at odds with other algorithms where models were overly optimistic. If the over-prediction at poorer health states can be resolved through Bayesian applications, this would be a significant development in the area of mapping.

Chapter 8 presented treatment effects from three important and commonly used HRQoL instruments were compared. EQ-5D-5L was shown to be more sensitive than EQ-5D-3L and further, it was also found that treatment effects from EQ-5D-5L relative to baseline are comparable with those of QLQ-C30. EQ-5D-5L appeared also to be more sensitive than EQ-5D-3L as patients became more severe (ECOG status). Again, the novelty in this thesis is that this may be the first such study, which measures and compares the sensitivity of EQ-5D-5L and EQ-5D-3L with a condition-specific measure using a common metric. This is an important area of research, because if the hypothesis that EQ-5D-5L is at least as sensitive as the cancer specific QLQ-C30 is true, this may lead to a wider use of EQ-5D-5L as a clinical measure of benefit and minimize concerns that the QALY is not reflective of the true (unknown) HRQoL benefit. The approach introduced in Chapter 8 may also contextualize the apparently small HRQoL effects.

Chapter 9 consolidates much of the above by developing an approach to selecting a mapping algorithm. The performances of several existing algorithms were tested using NSCLC data. A 'traffic' light approach is developed to selecting mapping algorithms, given the existence and continued development of new mapping algorithms. The limitations of mapping algorithms in patients at poor health states were also confirmed. This is the first approach to developing formal criteria for selecting among several published mapping algorithms. Further work around this could involve using formal multi-criteria decision analysis and decision theoretic approaches. The contributions and implications of these findings will now be discussed.

10.2 Contribution

The central contribution offered in this thesis relates to presenting alternative ways of modelling HRQoL data for economic evaluation. These contributions are now delineated further.

10.2.1 Methodological Contributions

The main methodological contributions are as follows:

(i) Chapter 4: No application of a Beta-Binomial (BB) model has been used in this context. This presents a shift in modelling from the usual OLS type models, which have been shown to be inadequate. The investigation into the BB approach has subsequently led to several further applications that show improved ways of measuring HRQoL. These can be understood readily by clinicians and health economist and can also contextualize generic HRQoL effects when compared with disease-specific measures; and where necessary, inputs justify adjustments to health benefits through the derivation of QALYs.

(ii) Chapter 5: The contribution of EQ-5D-5L, EQ-5D-3L, and other HRQoL data within the same patients for a sustained period of time is both a methodological and empirical contribution. It shows that such data collection is possible in this patient population. The conclusions of this research show that the EQ-5D-5L offers improved mapping over EQ-5D-3L regardless of the underlying model. However, among the models used for mapping EQ-5D-5L, several models do perform better – such as those that take into account skewness and over-dispersion. This chapter too was the first mapping algorithm with the EQ-5D-5L and made a direct comparison between the BB model and the LVDM model suggested by previous authors. This provided an opportunity for further investigation of other types of models, including Bayesian models.

(iii) Chapter 6: In this Chapter, again, the first time that toxicity and HRQoL data have been modelled in the context of mapping is presented. The relationship between toxicity and EQ-5D shows much promise and may reflect the underlying ‘direct’ connection between HRQoL in the context of cancer. In this Chapter too, the utility decrements were estimated and compared to those from utility studies conducted on the general public. The patient level utility decrements will be extremely valuable for future research. Currently, recourse to published utilities is made often using data published by Nafees et al. (2008). The real world NHS setting values are likely to be much more realistic of decrements in utilities for various toxicity and clinical characteristics.

(iv) Chapter 7 introduced the first Bayesian Network (BN) applied to the QLQ-C30. There were two new levels of novelty. Firstly, the application of the BN to the QLQ-C30 and secondly it was applied to the EQ-5D-5L. The model was not as successful as reported in the literature and may show the limitations of Bayesian algorithms to the QLQ-C30, which has a relatively large number of domains.

(vii) Chapter 8 is the first research that compares the sensitivity of the EQ-5D-5L with the EQ-5D-3L and the QLQ-C30. Much has been made as to whether economic evaluations make proper use of the clinical benefits and whether generic measures are not sensitive which might lead to unrealistically high QALYs. This chapter shows that in fact not only is the EQ-5D-5L more sensitive to the QLQ-C30 (not unexpected) but importantly the EQ-5D-5L reports comparable if not larger effect sizes than the QLQ-C30. When generic and condition-specific measures show conflicting effect sizes, the approach taken here supports the use of the EQ-5D-5L for demonstrating important clinical effects as well as its obvious use and value for cost-effectiveness analyses.

(viii) Chapter 10 offers an approach to select published mapping algorithms based on their performance on independent data. A theoretical basis is developed using bootstrap estimates in order to classify mapping algorithms. Such an approach is needed because it allows one to avoid the more uncertain utilities (and therefore more uncertain QALYs). This area is again an area of development and several areas of research including multi-criteria decision-making models and decision-theoretic models can be developed.

10.2.1 Empirical Contributions

The main empirical contribution of this thesis was the prospectively designed study to collect HRQoL data in NSCLC patients (Study 3). As described above, little or no published data is available in a NSCLC population that compares HRQoL within the same patients longitudinally over at least 12 months. The details of the process of collection and design have been provided earlier in the methods Chapter (Chapter 3).

10.2 Strengths and Limitations

The main strength of this research is the availability of prospectively available data designed to investigate the pre-specified objectives above. Methods to improve mapping were identified through complex modelling approaches hypothesized to predict better utilities. The sensitivity of the EQ-5D-5L was reported in terms of both mapping and also for measuring and reporting clinical benefit compared to disease-specific measures. The feasibility to develop criteria for selecting mapping algorithms was demonstrated.

Although this thesis offers novel contributions, there are some limitations. Firstly, the utilities for the EQ-5D-5Lwere not available (and still not available) at the time of writing the thesis and hence a ‘cross walk’ had to be used. This may result in different conclusions and as and when the final tariff becomes available, the results will need to be revised for any future publication. A second limitation is that mapping is considered to be a ‘second best’ and where possible utilities from preference based condition specific measures could be considered. A comparison the impact of such condition specific preference based measures and mapping requires further research and the relative merits of each require robust quantification. Although mapping is considered to be suboptimal, it is important to note that NICE guidance does not currently consider estimates of preference based utilities as being more important than mapping and utility studies are even lower in the priority [81,82]. Thirdly, the data set of Study 3 was small. Fourth, validation in independent data sets could not always be performed. Fifth, in Study 3, the impact of administering the EQ-5D-3L and 5L at different times might contribute towards a different conclusion. Randomization of the instruments to the time points could have been a possible alternative, however, differences in observed outcomes and results may have been unlikely given that this differed by only 2 weeks; initial indications suggested this was unlikely to have impacted on the conclusions. Sixth, only data from lung cancer patients were available and therefore, some inferences may be less generalizable. Seventh, the data is restricted to a NSCLC population and whether such conclusions are generalizable across all cancers requires further work. However, given that both instruments EQ-5D and QLQ-C30 are generic (in the sense that latter is generic to cancer while the former more widely generic in a broader sense, even beyond health care), the generalizability of conclusions may, in fact, be wide. Finally, Study 3 was not randomized and therefore a direct impact on the ICER as a result of the varying methods could not be evaluated.

10.3 Implications of Findings

The main results from this thesis have shown alternative ways to predict patient level utilities for economic evaluation using various mapping models. Each of these models depart from the traditional linear prediction approach which while simple has often resulted in over or under prediction at poorer health states and in some cases over estimated QALYs. Therefore, given the non normal structure of utility data, a shift towards non-linear models is the suggested direction for mapping. A further conclusion from this thesis is the proposal to consider the EQ-5D-5L as a future measure in economic evaluation because it may be a more sensitive measure to the extent that effect sizes may be commensurate with condition specific measures. It may also offer a better mapping over the EQ-5D-3L. The implications of these main conclusions are addressed below:

(i) Implications for Economic Evaluation:

a) The approach to mapping should go beyond linear models. Although Brazier et al (2010) has suggested that more complex models did not improve estimation of utility, in this thesis the model fit from non-linear models has shown a much better model fit. Consequently, the model that shows the best 'fit' should be used (rather than performance on R^2 alone). This would offer greater confidence in economic evaluations that used more complex mapping models. For example, the BB model is shown as a very useful and powerful way to model utility data for both economic evaluation and informing about the importance of apparently small EQ-5D differences on an alternative (odds ratio) scale.

b) A corollary to the above is that improved mapping algorithms will have more reliable estimates of QALYs which will inform key stakeholders (patients, clinicians, budget holders) on the relative value of treatments with respect to cost, and HRQoL.

c) Additional data should be used such as covariate data and in particular toxicity to properly model the relationship between generic and condition specific measures. This shows great promise and seems a very plausible approach. This approach has implications on how toxicity might be collected (e.g. close to EQ-5D assessments) as well as trial design for better evaluating the relationship between toxicity and HRQoL.

(ii) Implications for health state valuation:

a) In this thesis, I examined the sensitivity of generic and CSMs and came to the conclusion that the EQ-5D-5L should be considered as an important and more useful instrument not only for mapping but also as a potential clinical measure. This is because the observed measure of clinical effect from the EQ-5D-5L are commensurate with generic cancer measures.

b) A Bayesian mapping algorithm allows for an interesting and informative estimate of utilities while predictive the *full health state* description. What this implies is that the probabilities of each profile score can be updated if these are reported. For example, the prior probability of 0.2 for each of the scores (1 to 5) for the EQ-5D-5L will generate posterior probabilities. If these posterior probabilities are reported with the mapping algorithm, users of the algorithm could then use these as prior information (rather than assuming an equal probability of 0.2 which may not be realistic). This allows for continuous feedback and updating of probabilities until a set of utility values is reached which could be recommended. This avoids user bias (i.e. selection of a particular set of prior values) to generating utilities and subsequent QALYs.

(iii) Implications for Decision Making in Cancer:

a) As a result of a more sensitive instrument (EQ-5D-5L) to determine cost-effectiveness, the implications for new treatments that offer value as well as those treatments that do not offer value (assumed) can be differentiated. This implies patients and decision makers (budget holders) can benefit from the decisions of an instrument (and a mapping process) that leads to more informative assessments of HRQoL and a decision for a more or less cost-effective treatment.

b) The above point does not only have implications for NSCLC patients (from which this data is taken) but can be extended to other cancer patients because both the EQ-5D and the QLQ-C30 are generic type HRQoL measures in their respective contexts.

c) The approach proposed for identifying a useful mapping algorithm is particularly important for decision makers. It would allow a much more objective way to select the ‘best’ mapping model based on simulation from a suite of available algorithms. This should allow a reduction in decision uncertainty for the selection from an increasing availability of mapping models.

d) I have presented a rich area of methodological innovation in mapping and modelling HRQoL in cancer which could be developed further. This ranges from further development in modelling approaches to more objective methods to select from available algorithms. The selection method I have presented is but one of alternative methods that could be developed further.

10.4 Future research

A number of areas for future research remain. Further work is required in developing and testing mapping algorithms using real-world (and not a clinical trial) data from patients, to reflect estimates of ‘real world’ utilities for economic evaluation. The functional forms of mapping functions, including the use of splines, joint modelling and other more complex functional forms, which try to interrogate the relationship with clinical data, are needed. Separately, this should be applied to EQ-5D-5L since this may be a more suitable candidate instrument for future mappings, due to its extended scale. Currently, a project is being initiated to combine data across all available data sources and countries in an attempt to develop a ‘Mega Mapping’ model (Khan, Crott, Petrou, Doble).

Additional research on the sensitivity and responsiveness of EQ-5D-5L in comparison with other cancer measures (such as the FACT-L) is also required. A post-progression mapping algorithm will also be useful. A comparison between the models presented using post-progression utilities could be developed. In particular, research is needed on how to interpret HRQoL effects between the generic and the condition-specific measures, when they reach contradictory conclusions and impact on QALY. This is an important area of research, especially in the light of several cancer drugs being rejected on the grounds of unjustified (optimistic) utility / HRQoL values. Further research also in modelling and extrapolating post-progression utilities is warranted along with the implications for the QALYs.

10.5 Recommendations and Conclusion

This thesis has investigated the approaches to modelling HRQoL in cancer patients using data from NSCLC patients. A comprehensive literature review was undertaken. The scarcity of research in some areas underlined the gaps in knowledge and need for further research. The main results of this thesis can be summarised as:

Chapter 4:

❖ A BB model has shown that it offers a much more improved fit over existing methods at the time and is an important way of modelling EQ-5D data in the future. The results are considered robust, interpretable and the mapping algorithm is relatively simple to apply.

Chapter 5:

- The EQ-5D-5L offers an improved mapping algorithm compared to the EQ-5D-3L. Moreover, the two part BB model was shown to be a better fit compared to the more complex LDVMM approach which can become uninterpretable with a CSM such as the QLQ-C30.
- Chapter 6: Models that relate toxicity and EQ-5D to map a CSM are likely to be very informative and possibly the best way for handling any over and under prediction observed in almost all mapping algorithms to date.
- The possibility to jointly model EQ-5D and toxicity with covariates and complex models (e.g. Joint Beta Models) is likely to significantly improve prediction and model fit. However, the cost of this added complexity in terms of interpretability and usability needs more investigation
- EQ-5D-5L may offer potential for better mapping algorithms in the future.

Chapter 7:

- Bayesian Networks models allow mapping the entire profile. In this application, the Bayesian approach did not offer a significant improvement in mapping using a non-informative prior.
- The BN model did, however, confirm a better mapping with the EQ-5D-5L over the EQ-5D-3L.

Chapter 8:

- The EQ-5D-5L is more sensitive to measuring treatment benefit over time than the EQ-5D-3L
- The EQ-5D-5L has measures of effect commensurate (or larger) compared with the cancer specific measure QLQ-C30 and the size of these benefits can be considered clinically relevant. The EQ-5D-5L may be very informative in assessing whether any observed HRQoL benefit is important in addition to the QLQ-C30 for future economic evaluation of treatments for cancer.

Chapter 9:

- ❖ Objective criteria for selection of mapping algorithms was developed to separate the usefulness of mapping algorithms. Such an approach was considered feasible and the results were helpful in eliminating the use of 'poor' mapping algorithms.
- ❖ The methodology could be extended to include Delphi, consensus or hierarchical process approaches for selection.

Chapter 10:

- ❖ Future areas of research were identified to add to the findings from this thesis. These include:
  a) Joint modelling using a joint Beta and toxicity model
  b) Using final EQ-5D-5L value sets
  c) Developing an algorithm across much more comprehensive data
  d) Using more sophisticated selection approaches amongst the widely available mapping algorithms
  e) e) Comparing mapped utilities from those valued using CSM (e.g. such as the QLQ-C8D)
  f) Further examination of the sensitivity of the QLQ-C30 with EQ-5D-5L using RCT data

11. References and Bibliography

References

1. Micozzi SM. Complementary and Integrative Medicine in Cancer Care and Prevention: Foundations and Evidence-Based Interventions. Springer Publishing Company; 2006.
2. Cancer.Gov. Cancer Terms [Internet]. Available from:http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=45333
3. Cancer.Gov. Cancer Terms [Internet]. Available from:https://www.cancer.gov/about-cancer/understanding/what-is-cancer
4. Cancer Research Organisation, U.K. Worldwide Cancer Statistics [Internet].Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide- cancer.
5. Ferlay J, Shin H, Bray F, Forman D, Mathers C, Parkin D. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International Journal of Cancer. 2010; 127(12): 2893-2917.
6. Cancer Research UK. Lung Cancer and smoking statistics. [Internet]. UK: UK CR. 2012. Available from: http://publications.cancerresearchuk.org/cancerstats
7. Toms JR. Cancer Stats Monograph. Cancer Research UK. 2004.
8. Jemal A, Bray F, Center M, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: *A Cancer Journal for Clinicians*. 2011; 61(2): 69-90.
9. Eco.IARC [Internet]. Available from: http://eco.iarc.fr/eucan/Country.aspx?ISOCountryCd=826.
10. http://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html
11. Sun YQ, Chen Y, Langhammer A, Skorpen F, Wu C,, Mai XM; Passive smoking in relation to lung cancer incidence and histologic types in Norwegian adults: the HUNT study; Eur Respir J. 2017 Oct 12;50(4). pii: 1700824. doi: 10.1183/13993003.00824-2017. Print 2017 Oct.
12. https://www.cancercenter.com/discussions/blog/whats-the-difference-small-cell-and-non-small-cell-lung-cancer/
13. Thunnissen E, Kerr KM, Herth FJ, Lantuejoul S, Papotti M, Rintoul RC, Rossi G, Skov BG, Weynand B, Bubendorf L, Katrien G, Johansson L, López-Ríos F, Ninane V, Olszewski W, Popper H, Jaume S, Schnabel P, Thiberville L, Laenger F. The challenge of NSCLC diagnosis and predictive analysis on small samples. Practical approach of a working group. Lung Cancer. 2012 Apr;76(1):1-18. doi: 10.1016/j.lungcan.2011.10.017. Epub 2011 Dec 3.

14. Nice.org.uk. Lung cancer (non-small-cell, squamous, metastatic) - nivolumab (after chemotherapy) [ID811] | Guidance and guidelines |. [Internet]. NICE. 2016. Available from: https://www.nice.org.uk/guidance/indevelopment/gid-tag506
15. https://www.nice.org.uk/guidance/cg121)
16. Cancer.org. The Global Economic Cost of Cancer. [Internet]. 2016. Available from: http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf
17. Cancer Research UK. Lung cancer UK price tag eclipses the cost of any other cancer. [Internet]. 2016. Available from: http://www.cancerresearchuk.org/about-us/cancer-news/press-release/2012-11-07-lung-cancer-uk-price-tag-eclipses-the-cost-of-any-other-cancer
18. England.nhs.uk. [Internet]. 2013. https://www.england.nhs.uk/wp-content/uploads/2013/04/cdf-sop.pdf
19. England.nhs.uk. [Internet]. 2015. https://www.england.nhs.uk/cancer/cdf/:
20. England.nhs.uk. NHS England ; NHS increases budget for cancer drugs fund from £280 million in 2014/15 to an expected £340 million in 2015/16. [Internet]. 2015. Available from: https://www.england.nhs.uk/2015/01/12/cancer-drug-budget/
21. England.nhs.uk. [Internet]. 2013. https://www.england.nhs.uk/wp-content/uploads/2013/04/cdf-sop.pdf
22. Ulmeanu R, Antohe I, Anisie E, Antoniu S. Nivolumab for advanced non-small cell lung cancer: an evaluation of a phase III study. *Expert Review of Anticancer Therapy.* 2015; 1-3.
23. Scottishmedicines.org.uk. Scottish Medicines Consortium afatinib (Giotrif). [Internet]. 2016. Available from: http://www.scottishmedicines.org.uk/SMC_Advice/Advice/920_13_afatinib_Giotrif/afatinib_Giotrif
24. Nice.org.uk. Afatinib for treating epidermal growth factor receptor mutation-positive locally advanced or metastatic non-small-cell lung cancer | Guidance and guidelines. [Internet]. 2014. Available from: http://www.nice.org.uk/guidance/TA310
25. bs.gov.au. *Pharmaceutical Benefits Scheme (PBS) | Afatinib, tablet, 20 mg, 30 mg, 40 mg and 50 mg, (as dimaleate), Giotrif® (first line) - July 2013. [Internet].* 2016. Available from: http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2013-07/afatinib-first-line
26. Nice.org.uk. *Crizotinib for previously treated non-small-cell lung cancer associated with an anaplastic lymphoma kinase fusion gene* | *Guidance and guidelines. [Internet].* 2013. Available from : http://www.nice.org.uk/guidance/TA296

27. Scottishmedicines.org.uk.*Scottish Medicines Consortium crizotinib (Xalkori).* [Internet]. 2016. Available from: http://www.scottishmedicines.org.uk/SMC_Advice/Advice/865_13_crizotinib_Xalkori/crizotinib_Xalkori_Resubmission

28. Cadth.ca. *Xalkori for Advanced Non-Small Cell Lung Cancer | CADTH.ca.* [Internet]. 2016. Available from: *https://www.cadth.ca/xalkori-advanced-non-small-cell-lung-cancer*

29. Nice.org.uk. *Erlotinib for the first-line treatment of locally advanced or metastatic EGFR-TK mutation-positive non-small-cell lung cancer*: [Internet]. 2012. Available from: http://www.nice.org.uk/guidance/TA258

30. Nice.org.uk. *Erlotinib for the treatment of non-small-cell lung cancer. [Internet].* 2008. http://www.nice.org.uk/guidance/TA162

31. Scottishmedicines.org.uk. *Scottish Medicines Consortium erlotinib (Tarceva).* [Internet]. 2016. Available from: http://www.scottishmedicines.org.uk/SMC_Advice/Advice/749_11_erlotinib_Tarceva/erlotinib_Tarceva

32. *Scottishmedicines.org.uk. Scottish Medicines Consortium erlotinib (Tarceva).* [Internet]. *2016. Available from:* http://www.scottishmedicines.org.uk/SMC_Advice/Advice/749_11_erlotinib_Tarceva/erlotinib_Tarceva

33. *Pbs.gov.au. Pharmaceutical Benefits Scheme (PBS)* | *Erlotinib, tablets, 25 mg, 100 mg, 150 mg (as hydrochloride), Tarceva® - July 2013*. [Internet]. *2016. Available from: http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2013-07/erlotinib*

34. *Nice.org.uk. Gefitinib for the first-line treatment of locally advanced or metastatic non-small-cell lung cancer* | Guidance and guidelines | NICE. [Internet]. *2010. Available from:* http://www.nice.org.uk/guidance/TA192

35. Scottishmedicines.org.uk. *Scottish Medicines Consortium gefitinib (Iressa) Resubmission*. [Internet]. 2016. Available from: http://www.scottishmedicines.org.uk/SMC_Advice/Advice/615_10_gefinitib_Iressa/gefitinib_Iressa_Resubmission

36. *Pbs.gov.au. Pharmaceutical Benefits Scheme (PBS)* | *Gefitinib, tablet, 250 mg, Iressa® - July 2013.* [Internet]. *2016.* Available from: http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/psd/2013-07/gefitinib

37. Nice.org.uk. *Pemetrexed for the first-line treatment of non-small-cell lung cancer | Guidance and guidelines* | *NICE.* [Internet]. 2009. Available from: http://www.nice.org.uk/guidance/TA181

38. *Nice.org.uk. Pemetrexed maintenance treatment following induction therapy with pemetrexed and cisplatin for non-squamous non-small-cell lung cancer | Guidance and guidelines | NICE.* [Internet]. *2014.* Available at: http://www.nice.org.uk/guidance/TA309

39. *Scottishmedicines.org.uk. Scottish Medicines Consortium pemetrexed (Alimta) 3.* [Internet]. *2016.* Available from :

http://www.scottishmedicines.org.uk/SMC_Advice/Advice/531_09_pemetrexed__Alimta_/pemetrexed__Alimta__3

40. *Scottishmedicines.org.uk. Scottish Medicines Consortium pemetrexed (Alimta)_restr.* [Internet]. *2016.* Available from :

41. http://www.scottishmedicines.org.uk/SMC_Advice/Advice/Pemetrexed__500mg__powder__Alimta___342_07_/pemetrexed__Alimta__restr

42. WHO. Measuring Quality of Life, WHO/MSA/MNH/PSF/97.4. [Internet].1997. Available from:

http://www.who.int/mental_health/media/68.pdf (1997

43. Cykert S, Kissling G, Hansen C. Patient Preferences Regarding Possible Outcomes of Lung Resection. Chest. 2000; 117(6):1551-1559.

44. Montazeri A, Milroy R, Hole D, McEwen J, Gillis C. Quality of life in lung cancer patients: as an important prognostic factor. Lung Cancer. 2001; 31(2-3):233-40.

45. Klein R, Muehlenbein C, Liepa A, Babineaux S, Wielage R, Schwartzberg L. Cost-Effectiveness of Pemetrexed Plus Cisplatin as First-Line Therapy for Advanced Nonsquamous Non-small Cell Lung Cancer. Journal of Thoracic Oncology. 2009; 4(11):1404-1414.

46. Temel J, Greer J, Muzikansky A, Gallagher E, Admane S, Jackson, et al. Early Palliative Care for Patients with Metastatic Non–Small-Cell Lung Cancer. New England Journal of Medicine. 2010; 363(8):733-742.

47. Goodwin P, Black J, Bordeleau L, Ganz P. Health-Related Quality-of-Life Measurement in Randomized Clinical Trials in Breast Cancer--Taking Stock. JNCI Journal of the National Cancer Institute. 2003; 95(4): 263-281.

48. Blazeby J. Health-Related Quality of Life Measurement in Randomized Clinical Trials in Surgical Oncology. Journal of Clinical Oncology. 2006; 24(19): 3178-3186.

49. Damm, K, Roeske N, Jacob C. Health-related quality of life questionnaires in lung cancer trials: a systematic literature review. Health Econ Rev. 2013; 3(1):15.

50. Slevin M, Plant H, Lynch D, Drinkwater J, Gregory W. Who should measure quality of life, the doctor or the patient?. British Journal of Cancer. 1988; 57(1):109-12

51. Montazeri A, Harirchi I, Vahdani M, Khaleghi F, Jarvandi S, Ebrahimi M, et al. The EORTC breast cancer-specific quality of life questionnaire (EORTC QLQ-BR23):

translation and validation study of the Iranian version. Quality of Life Research. 2000; 9(2): 177-84.

52. Drummond M, O'Brien B. Clinical importance, statistical significance and the assessment of economic and quality-of-life outcomes. Health Econ. 1993; 2(3): 205-212.
53. Clauser S. Use of Cancer Performance Measures in Population Health: A Macro-level Perspective. Journal of the National Cancer Institute Monographs. 2004; 2004(33): 142-154.
54. Comabella C, Gibbons E, Fitzpatrick R. A Structured Review of Patient-Reported Outcome Measures for Patients with Lung Cancer. *Patient-reported* Outcome Measurement Group. Department of Public Health: University of Oxford.
55. Lewis G, Peake M, Aultman R, Gyldmark M, Morlotti L, et al. Cost-Effectiveness of Erlotinib versus Docetaxel for Second-Line Treatment of Advanced Non-Small-Cell Lung Cancer in the United Kingdom. Journal of International Medical Research. 2010; 38(1): 9-21.
56. Hollen P, Gralla R, Cox C, et al. A dilemma in analysis: issues in the serial measurement of quality of life in patients with advanced lung cancer. Lung Cancer. 1997; 18: 119–136.
57. https://scharr.dept.shef.ac.uk/nicedsu/technical-support-documents/utilities-tsd-series/; TSD 10,
58. Calman K. Quality of life in cancer patients--an hypothesis. Journal of Medical Ethics. 1984;10(3):124-127.
59. Cella D, Wiklund I, Shumaker S, Aaronson N. Integrating health-related quality of life into cross-national clinical trials. Qual Life Res. 1993; 2(6): 433-440.
60. Maringwa J, Quinten C, King M, Ringash J, Osoba D, Coens C, et al. Minimal clinically meaningful differences for the EORTC QLQ-C30 and EORTC QLQ-BN20 scales in brain cancer patients. Annals of Oncology. 2011; 22(9): 2107-2112.
61. http://groups.eortc.be/qol/manuals
62. Ades A, Lu G, Madan J. Which Health-Related Quality-of-Life Outcome When Planning Randomized Trials: Disease-Specific or Generic, or Both? A Common Factor Model. Value in Health. 2013; 16(1): 185-194.
63. Carreon L, Berven S, Djurasovic M, Bratcher K, Glassman S. The Discriminative Properties of the SF-6D Compared With the SF-36 and ODI. Spine. 2013; 38(1): 60-64.
64. Brooks, R, Rosalind Rabin, F. de Charro; The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective: Evidence from the EuroQol BIOMED Research Programme; Springer Science & Business Media, 9 Mar 2013
65. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X.

Qual Life Res. 2011 Dec;20(10):1727-36. doi: 10.1007/s11136-011-9903-x. Epub 2011 Apr 9; Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L).

66. Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford Univ. Press.
67. Rowen D, Young T, Brazier J, Gaugris S. Comparison of Generic, Condition-Specific, and Mapped Health State Utility Values for Multiple Myeloma Cancer. Value in Health. 2012; 15(8): 1059-1068.
68. Rowen D, Brazier J, Young T, Gaugris S, Craig B, King M, et al. Deriving a Preference-Based Measure for Cancer Using the EORTC QLQ-C30. Health Economics and Decision Science Discussion Paper, University of Sheffield. 2010.
69. Donna Rowen, John Brazier, Tracey Young, Sabine Gaugris, Madeleine T King, Benjamin Craig, Galina Velikova; Deriving a Preference Based Measure for Cancer using the EORTC QLQ-C30 *Value Health*. Author manuscript; available in PMC 2013 Oct 29. Published in final edited form as: Value Health. 2011 Jul-Aug; 14(5): 10.1016/j.jval.2011.01.004. doi: 10.1016/j.jval.2011.01.004
70. Daniel SJ Costa, Neil K Aaronson, Peter M Fayers, Peter S Grimison, Monika Janda, Julie F Pallant, Donna Rowen, Galina Velikova, Rosalie Viney, Tracey A Young, Madeleine T King. Deriving a preference-based utility measure for cancer patients from the European Organisation for the Research and Treatment of Cancer's Quality of Life Questionnaire C30: a confirmatory versus exploratory approach *Patient Relat Outcome Meas*. 2014; 5: 119–129. Published online 2014 Nov 6. doi: 10.2147/PROM.S68776
71. Kularatna S, Whitty JA, Johnson NW, Jayasinghe R, Scuffham PA. Development of an EORTC-8D utility algorithm for Sri Lanka. *Medical Decision Making* 2015; 35(3): 361-70
72. Paula K. Lorgelly, Brett Doble, Donna Rowen, John Brazier, Cancer 2015 investigators Condition-specific or generic preference-based measures in oncology? A comparison of the EORTC-8D and the EQ-5D-3L *Qual Life Res*. 2017; 26(5): 1163–1176. Published online 2016 Nov 9. doi: 10.1007/s11136-016-1443-y
73. Emilio Sacco, Daniele Tienforti, Alessandro D'Addessi, Francesco Pinto, Marco Racioppi, Angelo Totaro, Daniele D'Agostino, Francesco Marangi, and Pierfrancesco Bassi Social, economic, and health utility considerations in the treatment of overactive bladderOpen Access J Urol. 2010; 2: 11–24].
74. Simes R, Coates A. Patient Preferences for Adjuvant Chemotherapy of Early Breast Cancer: How Much Benefit Is Needed?. JNCI Monographs. 2001; 2001(30):146-152.
75. Dolan P. Modelling Valuations for EuroQol Health States. Medical Care. 1997; 35(11): 1095-1108.

76. Shaw JW, Johnson JA, Coons SJ. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Med Care. 1997; 43(3): 203-220.

77. Healthutilities.com. Health Utilities Inc. "Leaders in Health-Related Quality of Life Research". [Internet]. 2016. Available from: http://www.healthutilities.com/

78. Rohde G, Moum T, Haugeberg G. Comparing 15D and SF-6D performance in fragility wrist and hip fracture patients in a two-year follow-up case-control study. Value Health. 2012; 15(8):1100-1107.

79. Kaplan R, Sieber W, Ganiats T. The quality of well-being scale: Comparison of the interviewer-administered version with a self-administered questionnaire. Psychology & Health. 1997; 12(6): 783-791.

80. Richardson J, Khan M, Iezzi A, Maxwell A. 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 Multi-attribute Utility Instruments. Medical Decision Making. 2014; 35(3): 276-291.

81. Brazier J, Rowen D. NICE DSU technical support document 11: Alternatives to EQ-5D for generating health state utility values; report by the decision support unit. School of Health and Related Research: University of Sheffield, UK. 2011.

82. Nicedsu.org.uk. NICE DSU Decision Support Unit - Utilities TSD series [Internet]. 2016. Available from: http://www.nicedsu.org.uk/Utilities-TSD-series(2391676).htm

83. Nice.org.uk. How We Work. [Internet]. 2016. Available from http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisal processguides/GuideToMethodsTA201112.jsp.

84. Van Agt H, Essink-Bot M, Krabbe P, Bonsel G. Test-retest reliability of health state valuations collected with the EuroQol questionnaire. Social Science & Medicine. 1994; 39(11): 1537-1544

85. Hurst N. Re: Quality of life measures. Rheumatology. 1997; 36(1): 147-148.

86. Brazier J, Jones N, Kind P. Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire. Qual Life Res. 1993; 2(3): 169-80.

87. Badia X, Carné X. Evaluation of quality of life in clinical trials. Med Clin (Barc). 1998; 110(14): 550-6.

88. Wailoo A, Davis S, Tosh J. The incorporation of health benefits in cost utility analysis using the EQ-5d. Report by the decision support unit. http://www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D

89. Wailoo A, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B. Mapping to estimate Health-State Utility from Non–Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value in

90. Goulart B, Ramsey S. A Trial-Based Assessment of the Cost-Utility of Bevacizumab and Chemotherapy versus Chemotherapy Alone for Advanced Non-Small Cell Lung Cancer. Value in Health. 2011; 14(6): 836-845.
91. Brown T, Boland A, Bagust A, Oyee J, Hockenhull J, Dundar Y, et al. Gefitinib for the first-line treatment of locally advanced or metastatic non-small cell lung cancer. Health Technology Assessment. 2010; 14(Suppl. 2): 71-9.
92. Berthelot J, Will B, Evans W, Coyle D, Earle C, Bordeleau L. Decision Framework for Chemotherapeutic Interventions for Metastatic Non-Small-Cell Lung Cancer. JNCI Journal of the National Cancer Institute. 2000; 92(16):1321-1329.
93. Wang S, Peng L, Li J, Zeng X, Ouyang L, Tan C, et al. A Trial-Based Cost-Effectiveness Analysis of Erlotinib Alone versus Platinum-Based Doublet Chemotherapy as First-Line Therapy for Eastern Asian Nonsquamous Non–Small-Cell Lung Cancer. PLoS ONE. 2013; 8(3):
94. Maniadakis N, Fragoulakis V, Pallis A, Simou E, Georgoulias V. Economic evaluation of docetaxel-gemcitabine versus vinorelbine-cisplatin combination as front-line treatment of patients with advanced/metastatic non-small-cell lung cancer in Greece: a cost-minimization analysis. Annals of Oncology. 2009; 21(7): 1462-1467.
95. Thongprasert S, Tinmanee S, Permsuwan U. Cost-utility and budget impact analyses of gefitinib in second-line treatment for advanced non-small cell lung cancer from Thai payer perspective. Asia-Pacific Journal of Clinical Oncology. 2012; 8(1): 53-61.
96. Asukai Y, Valladares A, Camps C, Wood E, Taipale K, Arellano J, et al. Cost-effectiveness analysis of pemetrexed versus docetaxel in the second-line treatment of non-small cell lung cancer in Spain: results for the non-squamous histology population. BMC Cancer. 2010; 10(1): 26.
97. Araújo A, Parente B, Sotto-Mayor R, Teixeira E, Almodôvar T, Barata F, et al. An economic analysis of erlotinib, docetaxel, pemetrexed and best supportive care as second or third line treatment of non-small cell lung cancer. Revista Portuguesa de Pneumologia (English Edition). 2008; 14(6): 803-827.
98. Carlson J, Reyes C, Oestreicher N, Lubeck D, Ramsey S, Veenstra D. Comparative clinical and economic outcomes of treatments for refractory non-small cell lung cancer (NSCLC). Lung Cancer. 2008; 61(3): 405-415.
99. Vergnenegre A, Corre R, Berard H, Paillotin D, Dujon C, Robinet G, et al. Cost-Effectiveness of Second-Line Chemotherapy for Non-small Cell Lung Cancer: An Economic, Randomized, Prospective, Multicenter Phase III Trial Comparing Docetaxel and Pemetrexed: The GFPC 05-06 Study. Journal of Thoracic Oncology. 2011; 6(1): 161-168.

100. Cromwell I, van der Hoek K, Melosky B, Peacock S. Erlotinib or Docetaxel for Second-Line Treatment of Non-small Cell Lung Cancer. Journal of Thoracic Oncology. 2011; 6(12): 2097-2103.

101. Greenhalgh J, McLeod C, Bagust A, Boland A, Fleeman N, Dundar Y, et al. Pemetrexed for the maintenance treatment of locally advanced or metastatic non-small cell lung cancer. Health Technology Assessment. 2010. 14 (Suppl. 2).

102. Fragoulakis V, Pallis A, Kateilidou, Maniadakis, Georgoulias. Economic evaluation of pemetrexed versus erlotinib as second-line treatment of patients with advanced/metastatic non-small cell lung cancer in Greece: a cost minimization analysis. Lung Cancer: Targets and Therapy. p: 43.

103. Zhu J, Li T, Wang X, Ye M, Cai J, Xu Y, Wu B. Gene-guided Gefitinib switch maintenance therapy for patients with advanced EGFR mutation-positive Non-small cell lung cancer: an economic analysis. BMC Cancer. 2013; 13(1): 39.

104. Gilberto de Lima Lopes G, Segel J, Tan D, Do Y, Mok T, Finkelstein E. Cost-effectiveness of epidermal growth factor receptor mutation testing and first-line treatment with gefitinib for patients with advanced adenocarcinoma of the lung. Cancer. 2011; 118(4): 1032-1039.

105. Ontario Health Technology Assessment Series. 2010; 10 (25).

106. Chouaid C, Le Caer H, Locher C, Dujon C, Thomas P, Auliac JB, Monnet I, Vergnenegre A; GFPC 0504 Team;Cost effectivenes of erlotinib versus chemotherapy for first-line treatment of non small cell lung cancer (NSCLC) in fit elderly patients participating in a prospective phase 2 study (GFPC 0504). BMC Cancer. 2012 Jul 20;12:301. doi: 10.1186/1471-2407-12-301.

107. Bradbury PA, Tu D, Seymour L, Isogai PK, Zhu L, Ng R, Mittmann N, Tsao MS, Evans WK, Shepherd FA, Leighl NB; NCIC Clinical Trials Group Working Group on Economic Analysis; Economic analysis: randomized placebo-controlled clinical trial of erlotinib in advanced non-small cell lung cancer. J Natl Cancer Inst. 2010 Mar 3;102(5):298-306. doi: 10.1093/jnci/djp518. Epub 2010 Feb 16.

108. Nafees B, Stafford M, Gavriel S, Bhalla S, Watkins J. Health state utilities for non-small cell lung cancer. Health and Quality of Life Outcomes. 2008; 6(1): 84.

109. Khan I, Morris S. 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. 2014; 12(1).

110. Dunlop W, Uhl R, Khan I, Taylor A, Barton G. Quality of life benefits and cost impact of prolonged release oxycodone/naloxone versus prolonged release oxycodone in patients with moderate-to-severe non-malignant pain and opioid-induced constipation: a UK cost-utility analysis. Journal of Medical Economics. 2012; 15(3): 564-575.

111. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. The European Journal of Health Economics. 2010. 11(4); pp: 427-434.

112. Brazier J, Yang Y, Tsuchiya A, Rowen D. 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. 2009; 11(2): 215-225.

113. Barton G, Sach T, Jenkinson C, Avery A, Doherty M, Muir K. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores?. Health and Quality of Life Outcomes. 2008.

114. Jäkel A, Plested M, Dharamshi K, Modha R, Bridge S, Johns A. A Systematic Review of Economic Evaluations in Second and Later Lines of Therapy for the Treatment of Non-Small Cell Lung Cancer. Appl Health Econ Health Policy. 2012; 11(1): 27-43

115. Longworth L, Rowen D. Mapping to Obtain EQ-5D Utility Values for Use in NICE Health Technology Assessments. Value in Health. 2013; 16(1): 202-210.

116. Schuffham PA, Whitty JA, Mitchell A, Viney R. The use of QALY weights for QALY calculations: a review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002-4; Pharmacoeconomics. 2008. 26(4); pp: 297-310.

117. CHMP. Guideline on clinical investigation of medicinal products for the treatment of multiple sclerosis. 2005.

118. Round J. Capturing information loss in estimates of uncertainty that arise from mapping algorithms. Health Economists' Study Group. 2008.

119. Chuang L, Whitehead S. Mapping for economic evaluation. British Medical Bulletin. 2012; 101(1): 1-15.

120. Tosh J, Kearns B, Brennan A, Parry G, Ricketts T, Saxon D. Innovation in health economic modelling of service improvements for longer-term depression: demonstration in a local health community. BMC Health Services Research. 2013;13(1): 150.

121. Soini E, Hallinen T, Puolakka K, Vihervaara V, Kauppi M. Cost-effectiveness of adalimumab, etanercept, and tocilizumab as first-line treatments for moderate-to-severe rheumatoid arthritis. Journal of Medical Economics. 2012; 15(2): 340-351.

122. Kielhorn A, Porter D, Diamantopoulos A, Lewis G. UK cost-utility analysis of rituximab in patients with rheumatoid arthritis that failed to respond adequately to a biologic disease-modifying antirheumatic drug. Current Medical Research and Opinion. 2008; 24(9): 2639-2650.

123. Hawton A, Green C, Telford C, Wright D, Zajicek J. The use of multiple sclerosis condition-specific measures to inform health policy decision-making: mapping from the MSWS-12 to the EQ-5D. Multiple Sclerosis Journal. 2012; 18(6): 853-861.

124. Soini E. Cost-utility and expected value of perfect information related to trabectedin in the treatment of metastatic soft-tissue sarcoma: the publicly funded comments explored. Annals of Oncology. 2011; 22(6): 1465-1466.

125. Mcgrath C, Rofail D, Gargon E, Abetz L. Using qualitative methods to inform the trade-off between content validity and consistency in utility assessment: the example of type 2 diabetes and Alzheimer's Disease. Health and Quality of Life Outcomes. 2010; 8(1): 23.

126. McDaid D, Knapp M. Black-skies planning? Prioritising mental health services in times of austerity. The British Journal of Psychiatry. 2010; 196(6): 423-424.

127. Bastani P, Kiadaliri A. Cost-utility analysis of adjuvant therapies for breast cancer in Iran. International Journal of Technology Assessment in Health Care. 2012; 28(02): 110-114.

128. Kind P. Measuring the value of quality of life in cancer: an index based on EORTC QLQ-C30 presentation, ASCO 2005 Annual Meeting. Journal of Clinical Oncology. 2005; 29 (16S, Part I and II).

129. Versteegh M, Rowen D, Brazier J, Stolk E. Mapping onto EQ-5D for patients in poor health. Health and Quality of Life Outcomes. 2010; 8(1): 141.

130. Jang R, Isogai P, Mittmann N, Bradbury P, Shepherd F, Feld R, et al. Derivation of Utility Values from European Organization for Research and Treatment of Cancer Quality of Life-Core 30 Questionnaire Values in Lung Cancer. Journal of Thoracic Oncology. 2010; 5(12): 1953-1957.

131. Kontodimopoulos N, Aletras V, Paliouras D, Niakas D. Mapping the Cancer-Specific EORTC QLQ-C30 to the Preference-Based EQ-5D, SF-6D, and 15D Instruments. Value in Health. 2009; 12(8): 1151-1157.

132. McKenzie L, Van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D Instrument: The Potential to Estimate QALYs without Generic Preference Data. Value in Health. 2009; 12(1): 167-171.

133. Gray A. Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping. Medical Decision Making. 2006; 26(1): 18-29.

134. Hernández Alava M, Wailoo A, Ara R. Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Questionnaire Health State Utility Values. Value in Health. 2012; 15(3): 550-561.

135. Crott R, Versteegh M, Uyl-de-Groot C. An assessment of the external validity of mapping QLQ-C30 to EQ-5D preferences. Qual Life Res. 2012.

136. Marriott ER, van Hazel G, Gibbs P, Hatswell AJ. Mapping EORTC-QLQ-C30 to EQ-5D-3L in Patients With Colorectal Cancer. J Med Econ. 2016;1-7.

137. Iftekhar Khan, Steve Morris, Nora Pashayan, Bashir Matata, Zahid Bashir and Joe Maguirre; Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patient (Health and Quality of Life Outcomes (2016) 14:60; DOI 10.1186/s12955-016-0455-1)

138. Young T, Mukuria C, Rowen D, Brazier J, Longworth L. Mapping Functions in Health-Related Quality of Life: Mapping from Two Cancer-Specific Health-Related Quality-of-Life Instruments to EQ-5D-3L. Medical Decision Making. 2015; 35(7): 912-926.

139. Kharroubi SA, Edlin R, Meads D, Browne C, Brown J, McCabe C. Use of Bayesian Markov chain Monte Carlo methods to estimate EQ-5D utility scores from EORTC QLQ data in myeloma for use in cost-effectiveness analysis. Med Decis Making. 2015; 35(3): 351-60.

140. Proskorovsky I, Lewis P, Williams C, Jordan K, Kyriakou C, Ishak J, et al. Mapping EORTC QLQ-C30 and QLQ-MY20 to EQ-5D in patients with multiple myeloma. Health and Quality of Life Outcomes. 2014; 12(1): 35.

141. Versteegh M, Leunis A, Luime J, Boggild M, Uyl-de Groot C, Stolk A. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. Med Decis Making. 2012; 32: 554–568.

142. Kim EJ, Ko S, Kang H. Mapping the cancer-specific EORTC QLQ-C30 and EORTC QLQ-BR23 to the generic EQ-5D in metastatic breast cancer patients. Qual Life Res. 2011; 21(7): 1193-1203.

143. Kim SH, Ahn, J. Mapping EORTC QLQ-C30 onto EQ-5D for the assessment of cancer patients. Health and Quality of Life Outcomes. 2012; 10(1): 151-155.

144. Wu EQ, Mulani P, Farrell MH, Sleep D; Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients. Value Health. 2007 Sep-Oct;10(5):408-14.

145. Basu A, Manca A. Regression Estimators for Generic Health-Related Quality of Life and Quality-Adjusted Life Years. Medical Decision Making. 2012; 32(1): 56-69.

146. Khan KA, Madan J, Petrou S, Lamb SE. Mapping between the Roland Morris Questionnaire and generic preference-based measures. Value Health. 2014; 17(6): 686-95. doi: 10.1016/j.jval.2014.07.001.

147. Le Q, Doctor J. Probabilistic Mapping of Descriptive Health Status Responses Onto Health State Utilities Using Bayesian Networks. Medical Care. 2011; 49(5): 451-460.

148. Le Q, Doctor J, Zoellner L, Feeny N. Minimal clinically important differences for the EQ-5D and QWB-SA in Post-traumatic Stress Disorder (PTSD): results from a Doubly Randomized Preference Trial (DRPT). *Health and Quality of Life Outcomes*. 2013; 11(1): 59

149. Lee C, Luo N, Ng R, Wong N, Yap Y, Lo S, et al. Comparison of the measurement properties between a short and generic instrument, the 5-level EuroQoL Group's 5-

dimension (EQ-5D-5L) questionnaire, and a longer and disease-specific instrument, the Functional Assessment of Cancer Therapy—Breast (FACT-B), in Asian breast cancer patients. Qual Life Res. 2012; 22(7): 1745-1751.

150. DeVine J, Norvell D, Ecker E, Fourney D, Vaccaro A, Wang J, et al. Evaluating the Correlation and Responsiveness of Patient-Reported Pain With Function and Quality-of-Life Outcomes After Spine Surgery. Spine. 2011; 36: S69-S74.

151. Malkin A, Goldstein J, Perlmutter M, Massof R. Responsiveness of the EQ-5D to the Effects of Low Vision Rehabilitation. Optometry and Vision Science. 2013; 90(8): 799-805.

152. Krahn M, Bremner K, Tomlinson G, Ritvo P, Irvine J, Naglie G. Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. Qual Life Res. 2006; 16(3): 509-522.

153. Buchholz I, Thielker K, Feng Y, Kupatz P, Kohlmann T. Measuring changes in health over time using the EQ-5D 3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample. Qual Life Res. 2014; 24(4): 829-835.

154. Richardson J, Khan M, Iezzi A, Maxwell A. 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 Multi-attribute Utility Instruments. Medical Decision Making. 2014; 35(3): 276-291.

155. Khan I, Bashir Z, Forster M. Interpreting small treatment differences from quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30. Health and Quality of Life Outcomes. 2015; 13(1).

156. Bongers M, Coupe V, Jansma E, Smit E, Uyl-de Groot C. PCN93 Cost-Effectiveness of Treatment with New Agents in Advanced Non-Small-Cell Lung Cancer: A Systematic Review. Value in Health. 2011; 14(7): A451.

157. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ref:Ann Med. 2001;33(5):344–9)

158. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D.

159. Lloyd A, Nafees B, Narewska J, Dewilde S, Watkins J. Health state utilities for metastatic breast cancer. Br J Cancer. 2006; 95(6): 683-690.

160. Stein D, Joulain F, Naoshy S, Iqbal U, Muszbek N, Payne KA, et al. Assessing health-state utility values in patients with metastatic colorectal cancer: a utility study in the United Kingdom and the Netherlands. Int J Colorectal Dis. 2014; (10): 1203-10.

161. Crott R. Mapping algorithms from QLQ-C30 to EQ-5D utilities: no firm ground to stand on yet. Expert Review of Pharmacoeconomics & Outcomes Research. 2014; 14(4): 569-576.

162. Arnold D, Rowen D, Versteegh M, Morley A, Hooper C, Maskell. N. Testing mapping algorithms of the cancer-specific EORTC QLQ-C30 onto EQ-5D in malignant mesothelioma. Health and Quality of Life Outcomes. 2015.

163. Doble B, Lorgelly P. Mapping the EORTC QLQ-C30 onto the EQ-5D-3L: assessing the external validity of existing mapping algorithms. Qual Life Res. 2015.

164. Latimer N. NICE DSU technical support document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data report by the decision support unit June 2011 (last updated March 2013). [Internet]. Available from: http://www.nicedsu.org.uk/NICE%20DSU%20TSD%20Survival%20analysis.updated%20March%202013.v2.pdf

165. Lee S, Khan I, Upadhyay S, Lewanski C, Falk S, Skailes G, et al. First-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy (TOPICAL): a double-blind, placebo-controlled, phase 3 trial. The Lancet Oncology. 2012; 13(11): 1161-1170.

166. Maguire J, Khan I, McMenemin R, O'Rourke N, McNee S, Kelly V, et al. SOCCAR: A randomised phase II trial comparing sequential versus concurrent chemotherapy and radical hypofractionated radiotherapy in patients with inoperable stage III Non-Small Cell Lung Cancer and good performance status. European Journal of Cancer. 2014; 50(17): 2939-2949.

167. Fleishman A. A method for simulating non-normal distributions. *Psychometrika*. 1978; 43(4): 521-532.

168. Pourahmadi M, Daniels M, Park T. Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*. 2007; 98(3): 568-587.

169. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99.

170. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004

171. Graham et al, 2007; How many imputations are really needed? Some practical clarifications of multiple imputation theory; Prev Sci (2007) 8:206–213).

172. Koenker R: Quantile Regression (Econometric Society Monographs); 2005.

173. Honore B, Khan S, Powell JL: Quantile regression under random censoring. J Econometrics 2002, 109(1):67–105.

174. Khan S, Powell JL: Two-step estimation of semiparametric censored regression models. J Econometrics 2001, 103(1–2):73–110.

175. Powell JL: Least absolute deviations estimation for the censored regression-model. J Econometrics 1984, 25(3):303–325.

176. Briggs AH, Claxton K, Sculpher MJ: Decision Modelling for Health Economic Evaluation. Oxford: Oxford University Press; 2006. Ff

177. Paolino P. Maximum Likelihood Estimation of Models with Beta-Distributed Dependent Variables. Political Analysis. 2001; 9(4): 325-346.

178. Ospina R, Ferrari SLP, Cribari-Neto F: A general class of zero-or-one inflated beta regression models. Comp Stat Data Analysis 2004, 31(7):799–815.

179. Ospina R, Ferrari SLP: A general class of zero-or-one inflated beta regression models. Comput Stat Data Anal 2012, 56:1609–1623.

180. Ferrari SLP, Cribari-Neto F: Beta regression for modelling rates and proportions. J Appl Stat 2004, 31(7):799–815.

181. Swearingen CJ, Castro MSM, Bursac Z: Inflated Beta Regression: Zero, one and Everything in Between, SAS Global Forum; 2012.

182. Kieschnick R, McCullough BD: Regression analysis of variates observed on (0,1): percentages, proportions and fractions. Stat Model 2003, 3(3):193–213.

183. Draper N, Smith H; Applied Regression Analysis 3rd edition; John Wiley Publication, 2014; ISBN: 9780471170822

184. Khan I, Morris S, Hackshaw A, *et al*. Cost-effectiveness of first-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy. *BMJ Open* 2015;**5**:e006733. doi: 10.1136/bmjopen-2014-006733

185. McCabe, C, Edlin, R, Meads, D, Browne, C, Kharroubi, S. Constructing indirect utility models: some observations on the principles and practice of mapping to obtain health state utilities. Pharmacoeconomics. 2013;31(8):635–41

186. Round J. Is a QALY still a QALY at the end of life?. *Journal of Health Economics*. 2012; 31(3): 521-527.

187. Van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS; Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health. 2012; 15(5): 708-15.

188. Oppe M, Devlin N, van Hout B, Krabbe P, de Charro F. A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol. *Value in Health*. 2014; 17(4): 445-453

189. Schlattmann P. Medical Applications of Finite Mixture Models (Statistics for Biology and Health) Hardcover: Springer. 2009.

190. Kessler D, McDowell A. Introducing the FMM Procedure for Finite Mixture Models Paper 328-2012; SAS Institute Inc., Cary, NC; SAS Global Forum. 2012.

191. Chai T, Draxler R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature; Geosci. 2014; Model Dev., 7: 1247-1250.

192. Sabourin C, Crott R, Aballea S, Toumi M. Alternative regression methods for mapping utilities in oncology; ISPOR 18th Annual European Congress; Milan, Italy; November, 2015.[Internet]. Available from: http://www.ispor.org/ScientificPresentationsDatabase/Presentation/60984.

193. Crott R. Direct Mapping of the QLQ-C30 to EQ-5D Preferences: a comparison of regression methods; Pharmacoeconomics. [in press]. 2016.

194. Pattanaphesaj J, Thavorncharoensap M. Measurement properties of the EQ-5D-5L compared to EQ-5D-3L in the Thai diabetes patients. Health and Quality of Life Outcomes. 2015; 3(1): 14.

195. Cheung Y, Luo N, Ng R, Lee C. 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.

196. NCI CTC Verions 2.0 https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcv20_4-30-992.pdf

197. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. *Biostatistics;* 1(4): 465-480.

198. Joseph C. Gardiner. Joint Modelling of Mixed Outcomes in Health Services Research; Paper 435-2013 SAS Global Forum 2013. [Internet]. 2013. Available from: http://support.sas.com/resources/papers/proceedings13/435-2013.pdf

199. Sturza J. A Review and Meta-Analysis of Utility Values for Lung Cancer. *Medical Decision Making*. 2010; 30(6): 685-693.

200. Launois R, Reboul-Marty J, Henry B, Bonneterre J. A cost-utility analysis of second-line chemotherapy in metastatic breast cancer. Docetaxel versus paclitaxel versus vinorelbine; Pharmacoeconomics. 1996;10(5):504-21.

201. Khan I. 2015 Probabilistic Sensitivity Analyses For Clinical Trials of Cost-effectiveness Using the Method of Copulas: A Comparison of Simulation Methods; SCT Meeting 2014; Oral presentation. [Internet]. Available from:http://www.sctweb.org/public/search/detail.cfm?ID=A7C6951B-EAD9-AA85-9BD63CC8386DE3E1

202. McLachlan G, Basford K. Mixture models : inference and applications to clustering. Marcel Dekker. 1988

203. Heckerman D. A tutorial on learning with Bayesian networks. In: M. Jordan, ed., Learning in Graphical Models, Cambridge, MA: MIT Press. 1999:301-354. Ispor.org. Pharmaceutical HTA and Reimbursement Processes - Germany. [Internet]. 2009. Available from: http://www.ispor.org/htaroadmaps/germany.asp#5

204. Borchani H, Bielza C, Martinez-Martin P, Larranage P. Markov blanket-based approach for learning multidimensional Bayesian network classifiers: An application to predict the European Quality of Life-5 dimensions (EQ-5D) from the 39-item Parkinson's Disease Questionnaire (PDQ-39). Journal of Biomedical Informatics. 2012; 45: 1175–118.

205. Bayesserver.com. BayesServer - Bayesian network software for Data Science and Decision Science - Table of Content. [Internet]. 2016. Available from: http://www.bayesserver.com/Help/v7.8/help/

206. Lemonnier I, Guillemin F, Arveux P, Clément-Duchêne C, Velten M, Woronoff-Lemsi M, et al. Quality of life after the initial treatments of non-small cell lung cancer: a persistent predictor for patients' survival. *Health and Quality of Life Outcomes*. 2014; 12(1): 73.

207. Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. *Journal of Clinical Epidemiology*. 2003; 56(1): 52-60.

208. Krabbe P, Peerenboom L, Langenhoff B, Ruers T. Responsiveness of the generic EQ-5D summary measure compared to the disease-specific EORTC QLQ C-30. *Qual Life Res.* 2004; 13(7): 1247-1253.

209. Cancer Research UK. Lung cancer statistics. [Internet]. 2015. Available from: http://www.cancerresearchuk.org/cancer-info/cancerstats/keyfacts/lung-cancer/cancerstats-key-facts-on-lung-cancer-and-smoking

210. Aiken L. Note on Sensitivity: A Neglected Psychometric Concept. Perceptual and Motor Skills. 1977; 45(3f): 1330-1330.

211. Norman G, Sloan J, Wyrwich K. Interpretation of Changes in Health-related Quality of Life. *Medical Care*. 2003; 41(5): 582-592.

212. Crosby R, Kolotkin R, Williams G. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol*. 2003; 56(5): 395-407.

213. Revicki D, Cella D, Hays R, Sloan J, Lenderking W, Aaronson N. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes.* 2006; 27: 4-70.

214. Revicki D, Hays R, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *The Journal of Clinical Epidemiology*. 2008; 61(2): 102-9.

215. Khan I. *Design & Analysis of Clinical Trials for Economic Evaluation & Reimbursement: An Applied Approach Using SAS & STATA*. Chapman and Hall; 2015:320.

216. Lindeboom R, Sprangers MA, Zwinderman AH. Responsiveness: a reinvention of the wheel?. Health and Quality of Life Outcomes. 2005; 3:8.

217. Mulhern B, Bansback N, Brazier J, Buckingham K, Cairns J, Devlin N, et al. Preparatory study for the revaluation of the EQ-5D tariff: methodology report. *Health Technology Assessment*; 2014: 18(12).

218. Luo N, Cheung Y, Ng R, Lee C. Mapping and direct valuation: do they give equivalent EQ-5D-5L index scores?. *Health and Quality of Life Outcomes*; 2015:13(1).

219. Osoba D. Translating the Science of Patient-Reported Outcomes Assessment Into Clinical Practice. *JNCI Monographs*. 2007; 2007(37): 5-11.

220. King M. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res*. 1996; 5(6): 555-567.

221. Pickard A, Neary M, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health and Quality of Life Outcomes*. 2007; 5(1): 70.

222. Feddern M, Jensen T, Laurberg S. Chronic pain in the pelvic area or lower extremities after rectal cancer treatment and its impact on quality of life. *PAIN*. 2015; 156(9): 1765-1771.

223. Chie W, Yu F, Li M, Baccaglini L, Blazeby J, Hsiao C, et al. Quality of life changes in patients undergoing treatment for hepatocellular carcinoma. *Qual Life Res.* 2015; 24(10): 2499-2506.

224. Kurita G, Lundström S, Sjøgren P, Ekholm O, Christrup L, Davies A, et al. Renal function and symptoms/adverse effects in opioid-treated patients with cancer. *Acta Anaesthesiologica Scandinavica*. 2015; 59(8):1049-1059.

225. Marra CA, Woolcott JC, Kopec JA, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med. 2005;60(7):1571–82;

226. Gerhards SA, Huibers MJ, Theunissen KA, de Graaf LE, Widdershoven GA, Evers SM. The responsiveness of quality of life utilities to change in depression: a comparison of instruments (SF-6D, EQ-5D, and DFD). Value Health. 2011 Jul-Aug;14(5):732-9;

227. Kontodimopoulos N, Pappa E, Chadjiapostolou Z, Arvanitaki E, Papadopoulos A, Niakas D. Comparing the sensitivity of EQ-5D, SF-6D and 15D utilities to the specific effect of diabetic complications. *The European Journal of Health Economics*. 2010; 13(1): 111-120.

228. Efron R., Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. 1986; 1(1): 54-75.

229. Efron B, Tibshirani R. An introduction to the bootstrap. London: Chapman & Hall. 456.

230. Saaty T. Journal of Multi-Criteria Decision Analysis. 1999; 8 (1): 23–24.

231. Saaty T. Decision-making with the AHP: Why is the principal eigenvector necessary?. *European Journal of Operational Research*. 2003; 145(1): 85-91.

Bibliography

1. Koenker R. Quantile Regression (Econometric Society Monographs). 2005.
2. Kind P, Macran S. Eliciting Social Preference Weights for Functional Assessment of Cancer Therapy-Lung Health States. *PharmacoEconomics*. 2005; 23(11):1143-1153.
3. National Cancer Institute. *Comprehensive Cancer Information*. [Internet]. 2016. Available from: http://www.cancer.gov/
4. Billingham L, Bathers S, Burton A, Bryan S, Cullen M. Patterns. Costs and cost-effectiveness of care in a trial of chemotherapy for advanced non-small cell lung cancer. *Lung Cancer.* 2002; 37(2): 219-225.
5. Lees M, Aristides M, Maniadakis N, McKendrick J, Botwood N, et al. Economic evaluation of gemcitabine alone and in combination with cisplatin in the treatment of nonsmall cell lung cancer. *Pharmacoeconomics*. 2002; 20(5): 325-337.
6. Le Lay K, Myon E, Hill S, Riou-Franca L, Scott D, et al. Comparative cost-minimisation of oral and intravenous chemotherapy for first-line treatment of non-small cell lung cancer in the UK NHS system. *The European Journal of Health Economics*. 2007; 8(2):145-151.
7. Schiller J, Harrington D, Belani C, Langer C, Sandler A, et al. Comparison of Four Chemotherapy Regimens for Advanced Non–Small-Cell Lung Cancer. *New England Journal of Medicine.* 2002; 346(2): 92-98.
8. Chouaid C, Mitchell P, Agulnik J, Herder G, Lester J, Vansteenkiste J, et al. PCN110 Health-Related Quality of Life in Advanced Non-Small Cell Lung Cancer (NSCLC) Patients. *Value in Health.* 2012; 15(4): A227.
9. Doyle S, Lloyd A, Walker M. Health state utility scores in advanced non-small cell lung cancer. *Lung Cancer*. 2008; 62(3): 374-380.
10. Darlison L, Beckett P, Calman L, Mulatero C, O'Byrne K, Peake M, et al. 91 Follow-up of patients with advanced NSCLC following 1st line chemotherapy. A British Thoracic Oncology Group national survey. *Lung Cancer*. 2012; 75:S30-S31.
11. Aaronson N, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez N, 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. *JNCI Journal of the National Cancer Institute.* 1993; 85(5): 365-376.
12. Cocks K, King M, Velikova G, Fayers P, Brown J. Quality, interpretation and presentation of European Organisation for Research and Treatment of Cancer quality of life questionnaire core 30 data in randomised controlled trials. *European Journal of Cancer.* 2008; 44(13): 1793-1798.

13. Ignacio J, et al. The EORTC quality of life questionnaire QLO-C30 (version 3.0): Validation study for spanish prostate cancer patients. 2008.
14. Ara R, Brazier J. Predicting the Short Form-6D Preference-Based Index Using the Eight Mean Short Form-36 Health Dimension Scores: Estimating Preference-Based Health-Related Utilities When Patient Level Data Are not Available. *Value in Health.* 2009; 12(2).
15. Chuang L, Kind P, Converting the SF-12 into the EQ-5D. *PharmacoEconomics*. 2009; 27(6): 491-505.
16. Mortimer D, Segal L, Sturm J. Can we derive an 'exchange rate' between descriptive and preference-based outcome measures for stroke? Results from the transfer to utility (TTU) technique. *Health and Quality of Life Outcomes.* 2009; 7(1): 33.
17. Payakachat N, Summers K, Pleil A, Murawski M, Thomas J, Jennings K, et al. Predicting EQ-5D utility scores from the 25-item National Eye Institute Vision Function Questionnaire (NEI-VFQ 25) in patients with age-related macular degeneration. *Qual Life Res.* 2009; 18(7): 801-813.
18. Young T, Yang Y, Brazier J, et al. Using Rasch analysis to aid the construction of preference based measures from existing quality of life instruments. Health Economics and Decision Science Discussion Paper 2008; No. 08/0.
19. Brazier J, Czoski-Murray C, Roberts J, Brown, M, Symonds T, Kelleher C. Estimation of a Preference-Based Index from a Condition-Specific Measure: The King's Health Questionnaire. *Medical Decision Making.* 2007; 28(1):113-126.
20. Ferreira P, Ferreira L, Pereira L. How consistent are health utility values?. *Qual Life Res.* 2008; 17(9):1205-1205.
21. Lamers L, Uyl-de Groot C, Buijt I. The Use of Disease-Specific Outcome Measures in Cost-Utility Analysis. *PharmacoEconomics*. 2007; 25(7): 591-603.
22. Brazier J, Tsuchiya A. Preference-based condition-specific measures of health: what happens to cross programme comparability?. *Health Econ*. 2010; 19(2):125-129.
23. Versteegh M, Leunis A, Uyl-de Groot C, Stolk E. Condition-Specific Preference-Based Measures: Benefit or Burden?. *Value in Health*. 2012; 15(3): 504-513.
24. Davidoff A, Tang M, Seal B, Edelman M. Chemotherapy and Survival Benefit in Elderly Patients With Advanced Non-Small-Cell Lung Cancer. *Journal of Clinical Oncology*. 2010; 28: 2191-2197.
25. Cagney H. UK Cancer Drugs Fund to reassess 42 agents. *The Lancet Oncology*. 2015; 16(1): e8.
26. Groups.eortc.be. *Questionnaires | EORTC*. [Internet]. 2016. Available from: http://groups.eortc.be/qol/eortc-qlq-c30

27. Whitehead S, Ali S. Health outcomes in economic evaluation: the QALY and utilities. *British Medical Bulletin*. 2010; 96(1): 5-21.
28. Round J, Jones L, Morris S. Estimating the cost of caring for people with cancer at the end of life: A modelling study. *Palliative Medicine*. 2015; 29(10): 899-907.
29. Montazeri A, Milroy R, Hole D, McEwen J, Gillis CR. Quality of life in lung cancer patients: as an important prognostic factor; Lung Cancer. 2001;31(2-3):233-40.
30. Agborsangaya C, Lahtinen M, Cooke T, Johnson J. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. *Health and Quality of Life Outcomes*. 2014; 12(1): 74.
31. Bilmes J. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. 1998.
32. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. *The EORTC QLQ-C30 Scoring Manual. 3rd ed.* Brussels: European Organisation for Research and Treatment of Cancer; 2001.
33. Gu N, Bell C, Ji X, Carter J, Botteman M, Van Hout B. PSY67 Predicting eq-5d utilities from neuropathic pain scores: comparing indirect mapping of predicted item responses with direct mapping of scores. *Value in Health.* 2010; 13(7): A472-A473.
34. Janssen M, Pickard A, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Qual Life Res.* 2012; 22(7): 1717-1727.
35. Teckle P, Peacock S, van der Hoek K, Chia S, Melosky B, Gelmon K. QA2 cross-walking cancer-specific instruments to the EQ-5D and SF-6D. *Value in Health*. 2011; 14(3): A11.
36. Madsen A, Lang M, Kjærulff U, Jensen F. The Hugin Tool for Learning Bayesian Networks. In: T. Nielsen and N. Zhang, ed., S*ymbolic and quantitative approaches to reasoning with uncertainty*, Berlin: Springer; 2003: 594-605.
37. Mohan A, Guleria R, Pathak A, Bhutani M, Pal H, Mohan C, et al. Quality of Life Measures in Lung Cancer. *Indian Journal of Cancer*. 2005; 42(3):125-132.
38. Pacifico D. Fitting nonparametric mixed logit models via expectation-maximization algorithm. *The Stata journal*. 2012; 12(2): 284–298.
39. Shen Y. *Probability, Bayes Nets, Naive Bayes, Model Selection*. 1st ed. [ebook] MIT OpenCourseWare. 2010. Available from: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/tutorials/MIT6_034F10_tutor06.pdf
40. Stata.com. *Data Analysis and Statistical Software | Stata*. [Internet]. 2016. Available from: http://www.stata.com

41. Sullivan P, Ghushchyan V. Mapping the EQ-5D Index from the SF-12: US General Population Preferences in a Nationally Representative Sample. *Medical Decision Making*. 2006; 26(4): 401-409.
42. Crowley J, Ankerst DP. Handbook of statistics in Oncology, 2 nd ed. Chapman & Hall; 2006.
43. Thunnissen E, Kerr K, Herth F, Lantuejoul S, Papotti M, Rintoul R, et al. The challenge of NSCLC diagnosis and predictive analysis on small samples. Practical approach of a working group. *Lung Cancer.* 2012; 76(1):1-18.
44. Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, et al. Screening for Epidermal Growth Factor Receptor Mutations in Lung Cancer. *New England Journal of Medicine.* 2009; 361(10): 958-967.
45. Tukumo M, Toyooka S, Kiura K, Shigematsu H, Asano H, Aoe M, et al. PD-160 The relationship between epidermal growth factor receptor (EGFR) mutations and clinicopathologic features in non-small cell lung cancers. *Lung Cancer*. 2005; 49: S113.
46. Nccn.org. *NCCN Clinical Practice Guidelines in Oncology*. [Internet]. 2016. Available from: http://www.nccn.org/professionals/physician_gls/f_guidelines.asp
47. Reck M, Popat S, Reinmuth N, De Ruysscher D, Kerr K, Peters S. Metastatic non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology.* 2014; 25(suppl 3): iii27-iii39.
48. Maio M, Perrone F. Quality of Life in elderly patients with cancer ; Health and Quality of Life Outcomes. 2003, 1:44.
49. Calman K. Quality of life in cancer patients--an hypothesis. *Journal of Medical Ethics*. 1984; 10(3):124-127.
50. Cipriano L, Romanus D, Earle C, Neville B, Halpern E, Gazelle G, et al. Lung cancer treatment costs, including patient responsibility, by disease stage and treatment modality, 1992 to 2003. *Value in Health*. 2011; 14(1): 41-52.
51. Donnelly L. *NHS accused of 'shambles' as dying cancer sufferers denied drugs.* Telegraph.co.uk. [Internet]. 2015 Available from: http://www.telegraph.co.uk/news/11788507/NHS-accused-of-shambles-as-dying-cancer-sufferers-denied-drugs.html
52. *National Cancer Institute*, 102(5): 298-306.
53. NICE 2011
54. NICE ERG / HTA 192 Gefitinib (2009)
55. NICE ERG Erlotinib (2006)
56. NICE HTA 192 Gefitinib (2009)
57. NICE Cost Impact Report (2011)

58. EGFR Draft Consultation (2013)

59. www.ox.ac.uk. [Internet]. 2012. Available from:http://www.ox.ac.uk/news/2012-11-07-cancer-costs-uk-economy-%C2%A3158bn-year

60. www. gov. uk. [Internet]. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213788/dh_123576.pdf

61. Pickard A, Ray S, Ganguli A, Cella D. Comparison of FACT- and EQ-5D–Based Utility Scores in Cancer. *Value in Health*. 2012:15(2): 305-311.

62. Khan I, Morris S, Hackshaw A, Lee S. Cost-effectiveness of first-line erlotinib in patients with advanced non-small-cell lung cancer unsuitable for chemotherapy. *BMJ Open.* 2015; 5(7): p.e006733.

63. Tester W, Jin P, Reardon D, Cohn J, Cohen M. Phase II study of patients with metastatic nonsmall cell carcinoma of the lung treated with paclitaxel by 3-hour infusion. *Cancer.* 1997; 79(4): 724-729.

64. Langendijk J, ten Velde G, Aaronson N, de Jong J, Muller M, Wouters E. Quality of life after palliative radiotherapy in non-small cell lung cancer: a prospective study. *International Journal of Radiation Oncology*Biology*Physics*. 2000; 47(1):149-155.

65. Harper R, Brazier J, Waterhouse J, Walters S, Jones N, Howard P. Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. *Thorax.* 1997; 52(10): 879-887.

66. Kobelt et al. Quality-of-life aspects of the overactive bladder and the effect of treatment with tolterodine. *BJU International*. 1999; 83(6): 583-590.

67. Barton G, Bankart J, Davis A, Summerfield Q. Comparing Utility Scores Before and After Hearing-Aid Provision. *Applied Health Economics and Health Policy.* 2004; 3(2):103-105.

68. Wu EQ, Mulani P, Farrell MH, Sleep D. Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients; Value Health. 2007;10(5):408-14.

69. Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, et al. Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement. Health and Quality of Life Outcomes; 2015.

70. Marra C, Esdaile J, Guh D, Kopec J, Brazier J, Koehler B. A Comparison of Four Indirect Methods of Assessing Utility Values in Rheumatoid Arthritis. Medical Care. 2004; 42(11): 1125-1131.

71. Feeny D, Wu L, Eng K. Comparing Short Form 6D, Standard Gamble, and Health Utilities Index Mark 2 and Mark 3 utility scores: Results from total hip arthroplasty patients. *Qual Life Res.* 2004; 3(10):1659-1670.

12. Appendices

Appendix A1: Examples of HRQoL Instruments

A1.1: EQ-5D-3L
A1.2: EQ-5D-5L
A1.3: QLQ-C30

EQ-5D-3L

Health Questionnaire

English version for the UK

(Validated for Ireland)

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems in walking about □

I have some problems in walking about □

I am confined to bed □

Self-Care

I have no problems with self-care □

I have some problems washing or dressing myself □

I am unable to wash or dress myself □

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

I have no problems with performing my usual activities □

I have some problems with performing my usual activities □

I am unable to perform my usual activities □

Pain / Discomfort

I have no pain or discomfort □

I have moderate pain or discomfort □

I have extreme pain or discomfort □

Anxiety / Depression

I am not anxious or depressed □

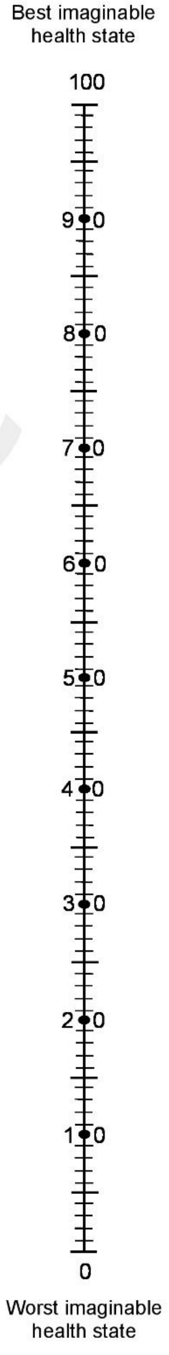
I am moderately anxious or depressed □

I am extremely anxious or depressed □

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your health state is today.

Your own health state today



EQ-5D-5L

Health Questionnaire

English version for the UK

Under each heading, please tick the ONE box that best describes your health TODAY.

MOBILITY

I have no problems 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 □

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 □

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 □

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 □

ANXIETY / DEPRESSION

I am not anxious or depressed □

I am slightly anxious or depressed □

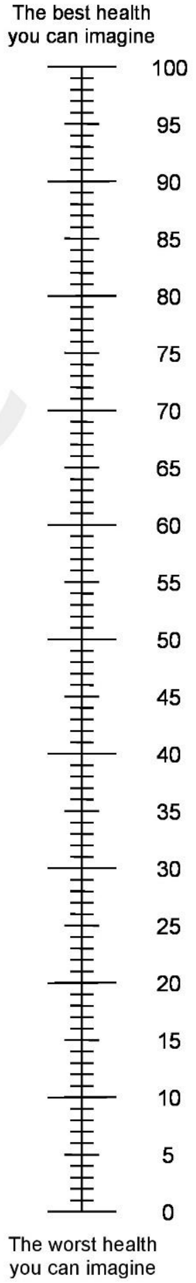
I am moderately anxious or depressed □

I am severely anxious or depressed □

I am extremely anxious or depressed □

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
  0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY = □



EORTC QLQ-C30 (version 3)

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

Please fill in your initials:

Your birthdate (Day, Month, Year):

Today's date (Day, Month, Year): 31

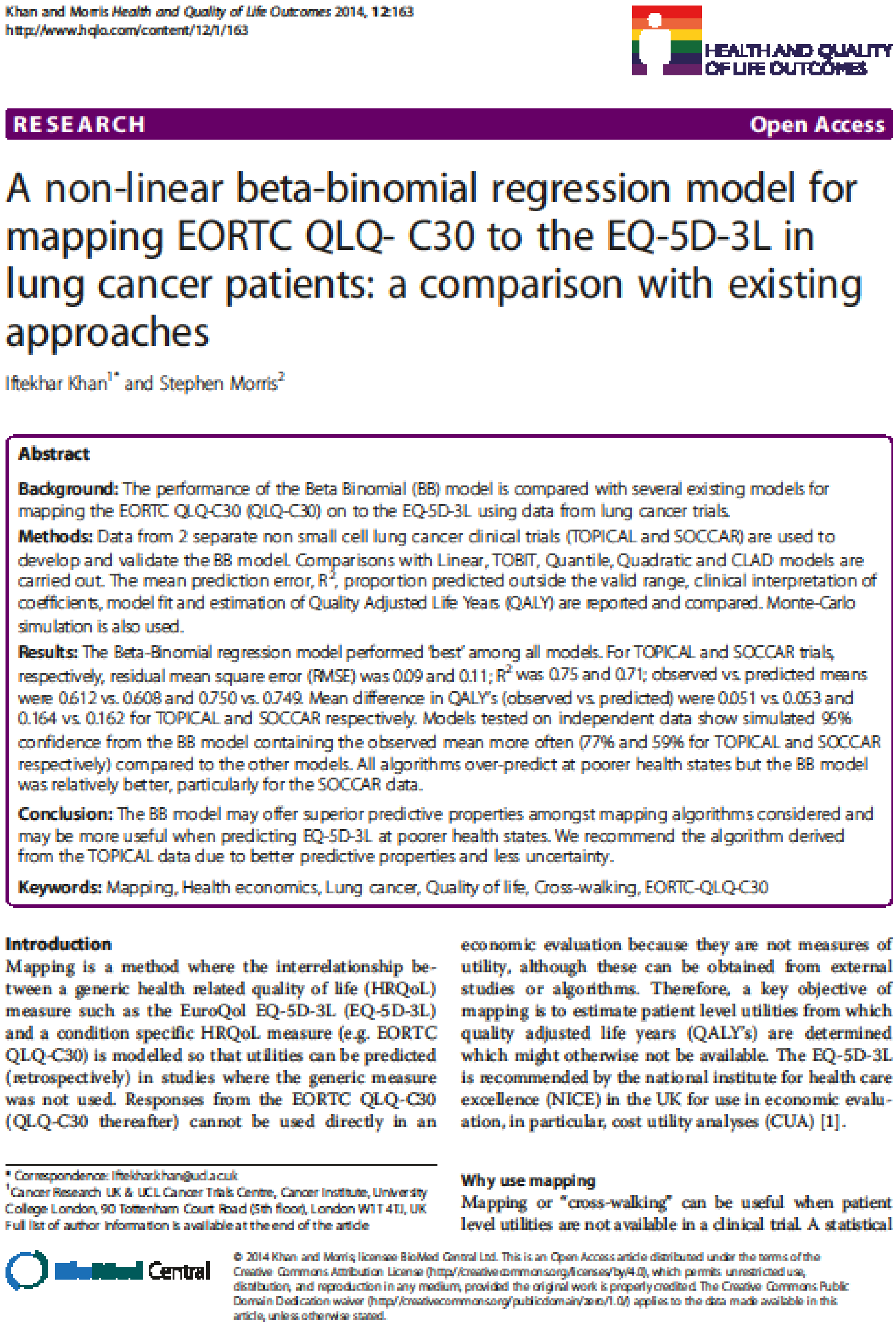
	Not at All	A Little	Quite a Bit	Very Much
1. Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2	3	4
2. Do you have any trouble taking a long walk?	1	2	3	4
3. Do you have any trouble taking a short walk outside of the house?	1	2	3	4
4. Do you need to stay in bed or a chair during the day?	1	2	3	4
5. Do you need help with eating, dressing, washing yourself or using the toilet?	1	2	3	4

During the past week:	Not at All	A Little	Quite a Bit	Very Much
6. Were you limited in doing either your work or other daily activities?	1	2	3	4
7. Were you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
8. Were you short of breath?	1	2	3	4
9. Have you had pain?	1	2	3	4
10. Did you need to rest?	1	2	3	4
11. Have you had trouble sleeping?	1	2	3	4
12. Have you felt weak?	1	2	3	4
13. Have you lacked appetite?	1	2	3	4
14. Have you felt nauseated?	1	2	3	4
15. Have you vomited?	1	2	3	4
16. Have you been constipated?	1	2	3	4

Please go on to the next page

Appendix A2: Published Articles Related to this thesis

A2.1: Chapter 4: *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* (Khan. I and Morris. S, 2014, Health and Quality of Life Outcomes 2014 12:163.).



Khan and Morris *Health and Quality of Life Outcomes* 2014, **12**:163
http://www.hqlo.com/content/12/1/163

HEALTH AND QUALITY OF LIFE OUTCOMES

RESEARCH **Open Access**

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

Iftekhar Khan[1*] and Stephen Morris[2]

Abstract

Background: The performance of the Beta Binomial (BB) model is compared with several existing models for mapping the EORTC QLQ-C30 (QLQ-C30) on to the EQ-5D-3L using data from lung cancer trials.

Methods: Data from 2 separate non small cell lung cancer clinical trials (TOPICAL and SOCCAR) are used to develop and validate the BB model. Comparisons with Linear, TOBIT, Quantile, Quadratic and CLAD models are carried out. The mean prediction error, R^2, proportion predicted outside the valid range, clinical interpretation of coefficients, model fit and estimation of Quality Adjusted Life Years (QALY) are reported and compared. Monte-Carlo simulation is also used.

Results: The Beta-Binomial regression model performed 'best' among all models. For TOPICAL and SOCCAR trials, respectively, residual mean square error (RMSE) was 0.09 and 0.11; R^2 was 0.75 and 0.71; observed vs. predicted means were 0.612 vs. 0.608 and 0.750 vs. 0.749. Mean difference in QALY's (observed vs. predicted) were 0.051 vs. 0.053 and 0.164 vs. 0.162 for TOPICAL and SOCCAR respectively. Models tested on independent data show simulated 95% confidence from the BB model containing the observed mean more often (77% and 59% for TOPICAL and SOCCAR respectively) compared to the other models. All algorithms over-predict at poorer health states but the BB model was relatively better, particularly for the SOCCAR data.

Conclusion: The BB model may offer superior predictive properties amongst mapping algorithms considered and may be more useful when predicting EQ-5D-3L at poorer health states. We recommend the algorithm derived from the TOPICAL data due to better predictive properties and less uncertainty.

Keywords: Mapping, Health economics, Lung cancer, Quality of life, Cross-walking, EORTC-QLQ-C30

Introduction

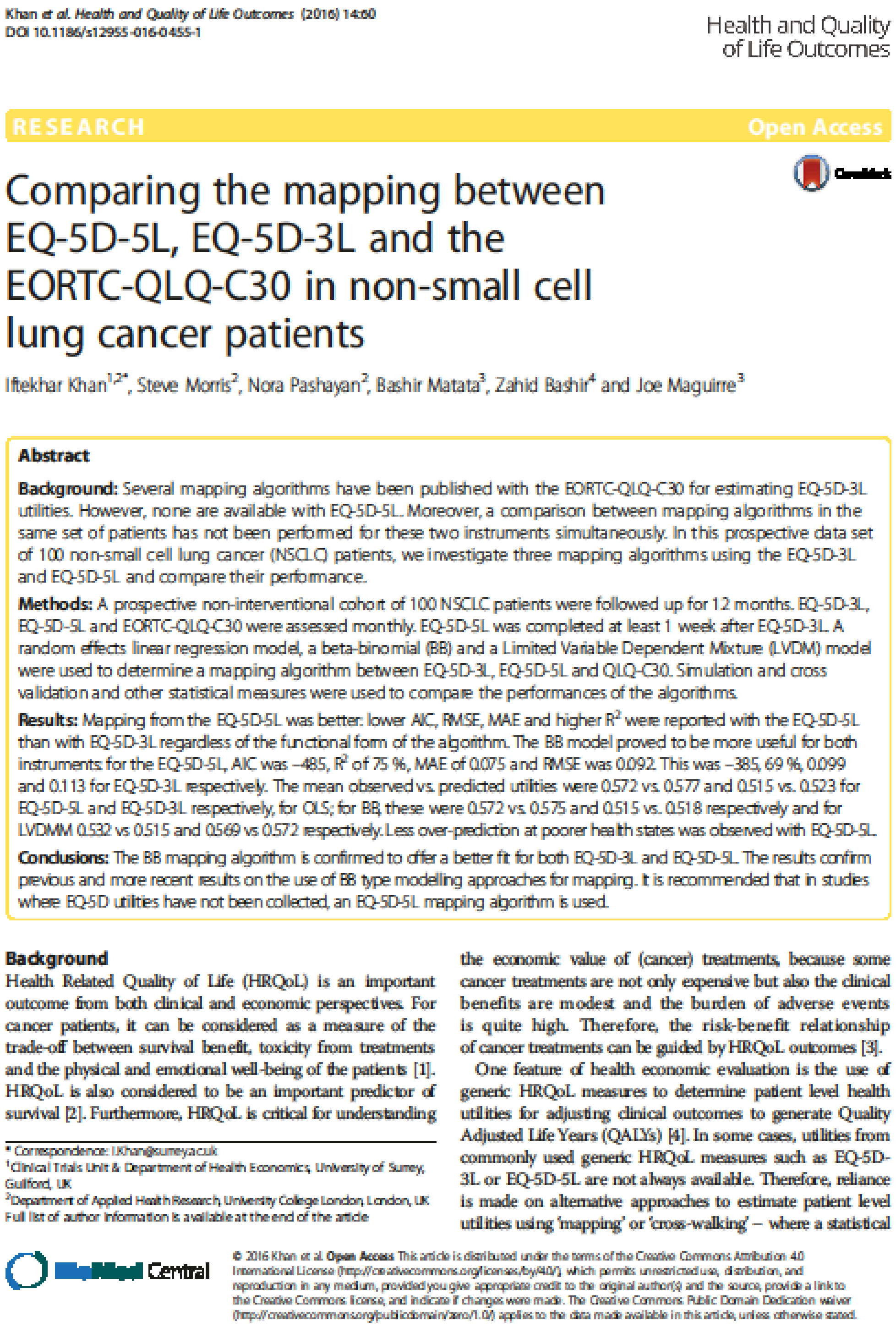
Mapping is a method where the interrelationship between a generic health related quality of life (HRQoL) measure such as the EuroQol EQ-5D-3L (EQ-5D-3L) and a condition specific HRQoL measure (e.g. EORTC QLQ-C30) is modelled so that utilities can be predicted (retrospectively) in studies where the generic measure was not used. Responses from the EORTC QLQ-C30 (QLQ-C30 thereafter) cannot be used directly in an economic evaluation because they are not measures of utility, although these can be obtained from external studies or algorithms. Therefore, a key objective of mapping is to estimate patient level utilities from which quality adjusted life years (QALY's) are determined which might otherwise not be available. The EQ-5D-3L is recommended by the national institute for health care excellence (NICE) in the UK for use in economic evaluation, in particular, cost utility analyses (CUA) [1].

Why use mapping

Mapping or "cross-walking" can be useful when patient level utilities are not available in a clinical trial. A statistical

* Correspondence: Iftekhar.khan@ucl.ac.uk
[1]Cancer Research UK & UCL Cancer Trials Centre, Cancer Institute, University College London, 90 Tottenham Court Road (5th floor), London W1T 4TJ, UK
Full list of author information is available at the end of the article

A2.2: Chapter 5: *Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients* (Khan. I, Morris. S, Pashayan. N, Matata. B, Bashir. Z and Maguirre. J; Health and Quality of Life Outcomes 201614:60).



Khan *et al. Health and Quality of Life Outcomes* (2016) 14:60
DOI 10.1186/s12955-016-0455-1

Health and Quality
of Life Outcomes

RESEARCH Open Access

Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients

Iftekhar Khan[1,2*], Steve Morris[2], Nora Pashayan[2], Bashir Matata[3], Zahid Bashir[4] and Joe Maguirre[3]

Abstract

Background: Several mapping algorithms have been published with the EORTC-QLQ-C30 for estimating EQ-5D-3L utilities. However, none are available with EQ-5D-5L. Moreover, a comparison between mapping algorithms in the same set of patients has not been performed for these two instruments simultaneously. In this prospective data set of 100 non-small cell lung cancer (NSCLC) patients, we investigate three mapping algorithms using the EQ-5D-3L and EQ-5D-5L and compare their performance.

Methods: A prospective non-interventional cohort of 100 NSCLC patients were followed up for 12 months. EQ-5D-3L, EQ-5D-5L and EORTC-QLQ-C30 were assessed monthly. EQ-5D-5L was completed at least 1 week after EQ-5D-3L. A random effects linear regression model, a beta-binomial (BB) and a Limited Variable Dependent Mixture (LVDM) model were used to determine a mapping algorithm between EQ-5D-3L, EQ-5D-5L and QLQ-C30. Simulation and cross validation and other statistical measures were used to compare the performances of the algorithms.

Results: Mapping from the EQ-5D-5L was better: lower AIC, RMSE, MAE and higher R^2 were reported with the EQ-5D-5L than with EQ-5D-3L regardless of the functional form of the algorithm. The BB model proved to be more useful for both instruments: for the EQ-5D-5L, AIC was −485, R^2 of 75 %, MAE of 0.075 and RMSE was 0.092. This was −385, 69 %, 0.099 and 0.113 for EQ-5D-3L respectively. The mean observed vs. predicted utilities were 0.572 vs. 0.577 and 0.515 vs. 0.523 for EQ-5D-5L and EQ-5D-3L respectively, for OLS; for BB, these were 0.572 vs. 0.575 and 0.515 vs. 0.518 respectively and for LVDMM 0.532 vs 0.515 and 0.569 vs 0.572 respectively. Less over-prediction at poorer health states was observed with EQ-5D-5L.

Conclusions: The BB mapping algorithm is confirmed to offer a better fit for both EQ-5D-3L and EQ-5D-5L. The results confirm previous and more recent results on the use of BB type modelling approaches for mapping. It is recommended that in studies where EQ-5D utilities have not been collected, an EQ-5D-5L mapping algorithm is used.

Background

Health Related Quality of Life (HRQoL) is an important outcome from both clinical and economic perspectives. For cancer patients, it can be considered as a measure of the trade-off between survival benefit, toxicity from treatments and the physical and emotional well-being of the patients [1]. HRQoL is also considered to be an important predictor of survival [2]. Furthermore, HRQoL is critical for understanding the economic value of (cancer) treatments, because some cancer treatments are not only expensive but also the clinical benefits are modest and the burden of adverse events is quite high. Therefore, the risk-benefit relationship of cancer treatments can be guided by HRQoL outcomes [3].

One feature of health economic evaluation is the use of generic HRQoL measures to determine patient level health utilities for adjusting clinical outcomes to generate Quality Adjusted Life Years (QALYs) [4]. In some cases, utilities from commonly used generic HRQoL measures such as EQ-5D-3L or EQ-5D-5L are not always available. Therefore, reliance is made on alternative approaches to estimate patient level utilities using 'mapping' or 'cross-walking' – where a statistical

* Correspondence: I.Khan@surrey.ac.uk
[1]Clinical Trials Unit & Department of Health Economics, University of Surrey, Guilford, UK
[2]Department of Applied Health Research, University College London, London, UK
Full list of author information is available at the end of the article

A2.3: Chapters 4 & 8: *Interpreting small treatment differences from the quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30* (Khan. I**,** Bashir. Z and Forster. M; Health and Quality of Life Outcomes 201513:180.) [Concepts developed in Chapters 4 and 8].

Khan *et al. Health and Quality of Life Outcomes* (2015) 13:180
DOI 10.1186/s12955-015-0374-6

HEALTH AND QUALITY OF LIFE OUTCOMES

RESEARCH Open Access

Interpreting small treatment differences from quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30

Iftekhar Khan[1]*, Zahid Bashir[2] and Martin Forster[3]

Abstract

Background: The EORTC-QLQ-C30 is a widely used health related quality of life (HRQoL) questionnaire in lung cancer patients. Small HRQoL treatment effects are often reported as mean differences (MDs) between treatments, which are rarely justified or understood by patients and clinicians. An alternative approach using odds ratios (OR) for reporting effects is proposed. This may offer advantages including facilitating alignment between patient and clinician understanding of HRQoL effects.

Methods: Data from six CRUK sponsored randomized controlled lung cancer trials (2 small cell and 4 in non-small cell, in 2909 patients) were used to HRQoL effects. Results from Beta-Binomial (BB) standard mixed effects were compared. Preferences for ORs vs MDs were determined and Time to Deterioration (TD) was also compared.

Results: HRQoL effects using ORs offered coherent interpretations: MDs >0 resulted in ORs >1 and vice versa; effect sizes were classified as 'Trivial' if the OR was between 1 ± 0.05 (i.e. 0.95 to 1.05); 'Small': for 1 ± 0.1; 'Medium': 1 ± 0.2 and 'Large': OR <0.8 or >1.20. Small HRQoL effects on the MD scale may translate to important treatment differences on the OR scale: for example, a worsening in symptoms (MD) by 2.6 points ($p = 0.1314$) would be a 17 % deterioration ($p < 0.0001$) with an OR. Hence important differences may be missed with MD; conversely, small ORs are unlikely to yield large MDs because methods based on OR model skewed data well. Initial evidence also suggests oncologists prefer ORs over MDs since interpretation is similar to hazard ratios.

Conclusion: Reporting HRQoL benefits as MDs can be misleading. Estimates of HRQoL treatment effects in terms of ORs are preferred over MDs. Future analysis of QLQ-C30 and other HRQoL measures should consider reporting HRQoL treatment effects as ORs.

Keywords: EORTC-QLQ-C30, Lung cancer, Quality of life, Beta binomial, Treatment effect size, MD: Mean Differences, ORs: Odds Ratios

Background

Health related quality of life (HRQoL) is an important endpoint in cancer trials for several reasons. First, where effect sizes are small, HRQoL can 'add value' to expensive cancer treatments. Secondly, considerable time is spent completing instruments for the purpose of estimating the impact of treatments on HRQoL. Therefore, such efforts should result in HRQoL effects that are meaningful and interpretable, especially where HRQoL is a primary or co-primary endpoint [1]. Thirdly, some anti-cancer treatments exhibit serious side-effects, despite improvements in overall survival (OS); HRQoL is also reported to be a predictor of survival in lung cancer patients [2], the leading cause of death among cancers [3]. It would be important to understand for example, how survival differs between patients with 'poor' baseline HRQoL, compared to those with 'Good' HRQoL. Finally, HRQoL outcomes are often required for cost-effectiveness

* Correspondence: iftekhar.khan@ucl.ac.uk
[1]Department of Applied Health Research, University College London, 1-19 Torrington Place, London WC1E 7HB, UK
Full list of author information is available at the end of the article