Methodological Issues in the Analysis of Health-Related Quality of Life Data for Cost-Effectiveness Analysis

Thomas Patton

Doctor of Philosophy

University of York

Economics

September 2015

Abstract

Health economic evaluations, developed for the purposes of informing technology adoption decisions in publicly-funded health care systems, should strive to make use of all relevant evidence. Any failure to meet this objective will result in a partial representation of the evidence base and, consequently, there is a potential risk of obtaining misleading results. Unfortunately, a lack of comparability amongst the alternative measures of health-related quality of life (HRQoL) complicates the synthesis of this type of evidence. One solution to this problem is to specify a reference case measurement to promote comparability, although this may provide an incomplete representation of HRQoL effects or, in some cases, no evidence at all.

The application of mapping functions – statistical algorithms that link HRQoL measures – might provide a means to incorporate a broader range of heterogeneous outcome measures for evidence synthesis. One method in particular, known as the common factor model (CFM), has been proposed in this regard due to its *coherent* mapping properties. Research involving the CFM has been conceptual to date and only a handful of case studies have ever been conducted. However, this method can be formulated as a structural equation model (SEM), an approach that has benefited from extensive application in other areas of research.

The primary aim of this thesis is to investigate the plausibility of SEM methods serving as a generalised framework for the handling of HRQoL evidence. SEM methods are tested across scenarios involving aggregate data, individual patient data and a combination of both; in each case, a comprehensive synthesis of heterogeneous HRQoL outcomes using the SEM approach is compared against a restrictive synthesis involving a reference case measurement. In addition, the implications of these alternative approaches are explored from a decision-making viewpoint.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to start by thanking my supervisor, Professor Andrea Manca, for his continued encouragement and support. Thank you for every discussion, the time you have spent editing drafts and all of the opportunities you have so generously allowed me.

Thank you is also extended to Professor Mark Sculpher and Professor Stephen Palmer for their expert advice. An important thank you goes to Dr Jan Boehnke, a valuable late addition to the thesis advisory group, for all of our discussions about structural equation models!

It has been a privilege to undertake my studies at the Centre for Health Economics (CHE), which has provided a wonderful environment for me to develop as a researcher. I would like to thank the staff and students at CHE, who have been a great source of support.

I would like to thank my family – Mum, Dad and Lydi – for their unyielding love and support.

Finally, I would like to dedicate this thesis to my wife, Erica. You are my inspiration! I am so thankful for all of the advice and encouragement that you have provided throughout my time writing this thesis.

Author's Declaration

I declare that this doctoral thesis is the result of my original work. I also affirm that this thesis has not previously been presented to any other university or educational institution for examination. In addition, any views expressed in this document are exclusive responsibility of the author. I hereby give my authorization for my thesis, if accepted, to be used for photocopying and inter-library loan. Likewise, I give my consent for the title and the abstract to be made available in all academic dissemination sources, making reference to authorship and copyright. All sources are acknowledged as references.

The material developed for this thesis has been communicated at one meeting of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR)

ISPOR Amsterdam, 2014: Patton, T. & Manca, A. "Integrating health psychometrics with health economics: can the 'mapping' toolbox be extended using ordinal structural equation models?" Value in Health, 17(7), A576

Chapter 1

Introduction

This thesis looks at some of the methodological issues encountered in the analysis of health-related quality of life (HRQoL) data for the purposes of health economic evaluation (HEE). In recent years, this has become an important area of research for two reasons: first, because policy-makers around the world are increasingly looking at HEE as a tool to inform the reimbursement decisions in health care (Drummond, 2013a,b); and second, because researchers and policy-makers alike have recognised the importance of valuing changes in HRQoL when conducting these evaluations.

The purpose of this chapter is to provide some necessary context relating to the topics covered in this thesis. A description of HEE is provided along with an outline of its role in health policy. An outline of the methods used for measuring and valuing HRQoL as a part of the evaluation process is then provided. This is followed by a description of the role that modelling techniques play in the consolidation of evidence for HEE. The final section of the chapter provides an overview of the content of each of the chapters covered in the thesis.

1.1 Health Economic Evaluation

Health policy-makers around the world are faced with the fundamental economic problem - how to allocate finite resources when there are (potentially) infinite, competing demands for those resources. Health economic research strives to address this problem by producing evidence which can be used to guide the efficient and equitable allocation of healthcare resources. HEE is a branch of health economics that examines the costs and benefits of alternative healthcare interventions (e.g. services, drugs, devices) in order to inform decisions about the reimbursement of those services.

Economic evaluation is an umbrella term covering a range of methodologies, the alternative forms of which differ chiefly in two regards: firstly, the way in which the benefits from health care services are measured and valued; and second, the assumptions made with regards to the changes in welfare occurring as a result of a policy change. Outside the field of health care, cost-benefit analysis (CBA) has been the most prominent method of economic evaluation, with notable applications having been conducted in the evaluation of public investments in infrastructure and transportation (Fuguitt and Wilcox, 1999). This branch of evaluation involves placing a monetary value upon the costs and benefits attributed to the alternative courses of action available. Sometimes, these values can be determined using data from real world markets, however, such values will not always be available. In these cases, analysts typically go about deriving non-market values by other means, principally using stated preference or revealed preference techniques.

Thus far, application of the CBA approach for the evaluation of health care services has been rare in practice. A body of methodological research exists in this area, particularly willingness to pay studies (Baker et al., 2014), although there have been reservations expressed regarding the appropriateness of measuring the benefits attributable to health care services in monetary terms. The willingness to pay approach has attracted criticism over concerns that individuals' responses are influenced by their ability to pay which could skew valuations disproportionally in favour of health care services that improve the health of wealthier respondents. Above all, this type of approach is unlikely to serve a publicly-funded health care system seeking to deliver health care services for all, irrespective of ability to pay (e.g. the National Health Service in the United Kingdom).

Another approach, more commonly used, is broadly referred to as cost-effectiveness analysis (CEA). Historically, the application of CEA has, for the most part, focused upon the identification of health care services expected to deliver maximum health improvements for a given budget (Dolan and Tsuchiya, 2006). Note that the exact measurement of health has yet to be specified; in practice, there are a number of ways in which one might go about quantifying the health effects of a given intervention including patient survival, clinical endpoints or through patients' responses to self-assessed questionnaires.

A critical aspect of the CEA methodology is the selection of an outcome measure capable of capturing all of the aspects of health considered to be relevant to the decision problem. In publicly-funded health care systems, reimbursement decisions need to be made across a broad spectrum of disease areas and consequently the outcome measure selected should be sufficiently broad to account for the health domains encompassed across this spectrum. To meet this requirement, the health economics research community developed the quality-

adjusted life year (QALY), a measure of health that combines survival and HRQoL into one numeraire. QALYs are estimated by adjusting the length of time alive by valuations of HRQoL typically ranging between 0 and 1, where 1 reflects perfect health and 0 reflects death (the methods for valuing HRQoL are discussed in the next section). For example, four years spent in a health state with an associated HRQoL value of 0.5 would be equivalent to two QALYs. The generic nature of the QALY enables policy-makers to compare the estimated health effects stemming from policy decisions across-the-board of disease areas.

The investigations carried out in this thesis are focused exclusively upon the application of economic evaluation where the benefits of health care are captured by the QALY. This approach is consistent with the methods employed in practice. Economic evaluation with the QALY has been adopted as a key tool in health technology assessment (HTA), a multi-disciplinary process aiming to inform best practice and, in some cases, reimbursement decisions in health care. Following the establishment of the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia over 20 years ago, many countries have set up HTA agencies applying the QALY-based approach to economic evaluation to inform health policy decisions.

The generic nature of the QALY facilitates informed decision-making in the reimbursement of services across the entire spectrum of health care. However, HEE involving QALYs require decision rules in order to determine whether or not services are considered to be welfare improving. For instance, the simple rule of dominance states that any policy that generates more health benefits and lower costs than the alternative options dominates those options and represents an improvement in welfare. In reality, many decisions need to be made in relation to the adoption of expensive new technologies that result in more QALYs as well as increased costs relative to standard practice. With this in mind, the optimal bundle of services could theoretically be determined by ranking all of the services according to their incremental cost-per-QALY and then select those that make the best use of the budget available.[1] Given that this is not a feasible task in practice, many decision makers have adopted an alternative solution which involves setting a (arbitrary) cost per QALY threshold below which new technologies could be considered to be cost-effective (for examples see Cleemput et al. (2008)). This threshold should represent the cost-per-QALY of the marginal technology currently being reimbursed, thus representing the opportunity cost in any decision involving the reimbursement of a new technology.

[1] That is, maximizing population net (of costs) health benefits (as measured by QALYs), given existing constraints

Identification of the true threshold is a challenging task and work in this area has only recently been undertaken (Claxton et al., 2013).

Despite having a central role in HEE, it is important to recognise that there are on-going methodological issues that have yet to be resolved regarding the use of the QALY, as well as normative issues to consider that are often overlooked. In particular, the often-quoted mantra 'a QALY is a QALY is a QALY' – used in reference to the assumption that all health gains captured by the QALY should be valued equally – has been subject to considerable scrutiny (Dolan et al., 2005; Rawlins and Culyer, 2004). Amongst other things, debates have revolved around claims that the QALY is ageist (Harris, 2005; Paulden et al., 2010) and questioned whether society might value health improvements differently depending upon severity of illness (Shah, 2009) and remaining life expectancy (Shah et al., 2014). Whilst work in this area is still ongoing, these questions fall outside the scope of this thesis.

Moving forward, it is important to recognize the normative underpinnings associated with the cost-per-QALY approach to evaluation and to understand where these fit in relation to the foundations of welfare theory in economics. The conventional approach to welfare economics, often referred to as the welfarist approach, assumes that social welfare is optimised according to policies that maximize the sum of individual utilities, where utility represents an individual's preferences for different states of the world (Brouwer et al., 2008; Hurley, 2014). The term extra-welfarism is used to describe any approach that deviates from the assumptions of the welfarist approach (Brouwer et al., 2008; Hurley, 2014). As we shall see in the next section, CEA with QALYs might be seen as being an extra-welfarist approach owing to the fact that the preferences beyond those of the individual are seen as being important when valuing health. Although these concepts might be beyond the scope of this thesis, a pragmatic perspective is assumed to be the most appropriate standpoint for the purposes of decision-making in health policy: that the assumptions required for the welfarist approach to be valid are unnecessarily restrictive for this to be a requirement (i.e. that the extra-welfarist approach is acceptable).

1.2 The Measurement and Valuation of HRQoL

Having established the rationale for using QALYs to capture health benefits when evaluating health care interventions, this section considers some of the competing approaches for the measurement and valuation of HRQoL. The development of techniques for the valuation of HRQoL has been one of the critical areas of research in health economic evaluation.

In particular, efforts have focused upon techniques that enable the quantification of peoples' preferences for specific health states. Valuation exercises that have predominated – the Standard Gamble (SG) and Time Trade-Off (TTO) – obtain values by exploring the willingness of respondents to accept increases in the risk of death or decreases in life expectancy (Gafni, 1994; Gudex et al., 1994). The origins of these valuation exercises can be traced back to attempts to derive health-related utility scores with methods similar to those in the expected utility theory framework. It is for this reason that many people still refer to SG and TTO values as being utilities despite both measures having been shown to produced biased estimates that violate expected utility theory (Bleichrodt, 2002). However, the question over whether or not HRQoL values should constitute utilities in the true sense is not a priority for research in this thesis. Instead, a pragmatic stance is assumed, one that sees SG and TTO values as being acceptable for the purposes of the application at hand.

Another source of debate in the methodological literature has been the question of whose values should be used. Arguments in favour of the use of patient values typically point to the fact that patients have first-hand experience of the condition and so have the best understanding of the impacts upon HRQoL (Menzel, 2014). An opposing view to this is that, for a government-provided health care system, values from the general population should be used on the basis that provision of health care should reflect the preferences of the tax-payer (Gold and Siegel, 1996). What makes this debate more complicated is the fact that the general population tends to value the HRQoL of a given health state lower than patients experiencing that health state (Menzel, 2014). This phenomenon is sometimes attributed to the idea that patients adapt to their condition over time and, accordingly, don't see the negative aspects.

The debate over whose values should be used in economic evaluation has largely been put to one side since the development and adoption of preference-based measures (PBMs) of HRQoL. This might be partly due to the fact that these measures involve both patients and the general population in the valuation exercise. There are two components to these methods, the first being a multilevel descriptive system designed to cover the core dimensions of health and the second being a pre-specified set of values associated with each of the states in the descriptive system. The value set is usually derived from a random sample of the general population using valuation techniques such as the standard gamble, time trade off or visual analogue scale (Brazier et al., 2002; Dolan, 1997; Feeny et al., 2002).

PBMs are widely used thanks to their relative ease of use, the relatively low burden for respondents and their claimed generic properties (Brazier, 2007). The collection of HRQoL values using PBMs is straightforward because it simply involves asking patients to rate their own health using the descriptive system and then assigning the pre-specified value set. Despite the advantages of these instruments, research has shown that there are important differences in the values derived using the different instruments available (Conner-Spady and Suarez-Almazor, 2003). The lack of comparability amongst the instruments available has led some policy-makers to recommend that all studies should use the same generic measure - a reference case measurement - in order to promote comparability across studies (NICE, 2013). However, the downside to this approach is that it may result in an incomplete representation of HRQoL effects or, in some cases, no evidence at all. This thesis will focus attention upon the notion of setting a reference case instrument and the challenges that the challenges this poses around the use of available evidence.

1.3 Decision Analytic Modelling

As previously mentioned, CEA is used to inform reimbursement decisions in health care. Informing such a decision requires appropriate consideration of all relevant treatment options, inclusion of all relevant evidence and an appropriate time horizon (Drummond, 2005). It is unlikely that one source of evidence, such as a clinical trial, will provide all of the evidence necessary to derive a relevant estimate of cost-effectiveness (Sculpher et al., 2006). Instead, Sculpher and colleagues recommended an evidence synthesis and decision modelling approach to CEA for decision-making (Sculpher et al., 2006). The decision modelling approach provides a framework to incorporate evidence form a range of sources in order to address a specific policy decision. More specifically, decision models estimate outputs, in the form of expected costs and QALYs for different treatment strategies, using model inputs that relate to patient survival, treatment effects, HRQoL, resource use and costs. These models are typically structured to reflect some underlying biological or clinical process that characterises patients' movements between defined health states using clinical evidence (Briggs et al., 2006).

In principle, all of the parameter types employed in decision analytic models for CEA should make comprehensive use of relevant evidence (Sculpher et al., 2006). Any failure to meet this objective will result in a partial representation of the evidence base bringing with it the potential risk of obtaining misleading results. One of the main challenges for researchers undertaking this task is the matter of how to utilize the available evidence

when there are multiple studies addressing the same research question. A large body of methodological research can be found concerning the synthesis of clinical evidence from published studies using statistical methods, a practice broadly known as meta-analysis (Sutton et al., 2012). Meta-analysis can be defined as a statistical analysis that combines or integrates the results from several independent studies (Egger and Smith, 1997). One of the main reasons for implementing this type of procedure is the possibility of attaining increased statistical power for the estimation of parameters (Higgins et al., 2009). This thesis will explore how these methods can be used to incorporate HRQoL evidence within the decision modelling framework.

1.4 Economic Evaluation in Policy

Health technology assessment (HTA) is a multi-disciplinary process that involves the evaluation of evidence of health care services in order to inform the reimbursement of those services (Hutton et al., 2006). Economic evaluation has come to play an important role in HTA, along with evidence on clinical effects and safety. The routine application of economic evaluation in some countries has led to the production of guidance outlining recommendations for best practice in the methods employed (CADTH, 2006; NICE, 2013; PBAC, 2008). This followed advice from Gold and colleagues proposing the implementation of a methodological reference case to ensure consistency and comparability across cost-effectiveness studies (Weinstein et al., 1996). Many agencies have specified a preference for the submission of economic evidence with QALYs, although there is variation in the degree to which this is a requirement. One agency – the National Institute for Health Care Excellence (NICE) in England and Wales – has gone as far as recommending a reference case measure for the valuation of HRQoL, the EQ-5D, due to the inconsistences that can arise between the different measurement options (NICE, 2013).

1.5 Thesis Overview

The broad aims of this thesis are to explore the methodological issues in the analysis of HRQoL for cost-effectiveness analysis. More specifically, this thesis is concerned with statistical methods for incorporating HRQoL evidence within the decision modelling framework. The objectives of this thesis can be set out into four research questions:

- **Research question 1:** Evaluate the current state of practice with respect to the statistical methods used to analyse HRQoL data in health economic modelling (chapter 2).

- **Research question 2:** Evaluate the methods and guidance currently available, offer further guidance where possible and identify areas where future research would be worthwhile (chapters 3 and 4).

- **Research question 3:** Undertake a series of empirical case studies to test the methodological issues identified in research question 2 (chapters 5 – 7).

- **Research question 4:** What recommendations can be made on the basis of the findings in this thesis? (chapter 8).

In Chapter 2, a review of submissions to NICE is carried out to characterise the statistical methods used to incorporate HRQoL evidence in HEEs. The findings show that statistical methods are used on an irregular basis and, as a result, there are fundamental inconsistencies with respect to the use of HRQoL evidence. Chapter 3 reviews the methodological literature and policy guidance relating to the methods for synthesising HRQoL evidence. The review finds the policy guidance in this area to be vague and the statistical methods available to be inadequate for researchers seeking to make comprehensive use of HRQoL evidence. One of the key challenges in this regard is the issue of between-instrument heterogeneity. A recent study has suggested that this problem may be circumvented through the use of mapping techniques. Chapter 4 reviews some of the statistical issues associated with the development of mapping algorithms. A handful of studies are found to show that structural equation modelling (SEM) techniques exhibit a great deal of promise when it comes to the synthesis of HRQoL evidence involving heterogeneous outcome measures. In light of these findings, investigations throughout the remainder of the thesis focus upon the application of these methods.

Chapters 5 to 7 present a series of case studies applying SEM methods across the range of plausible scenarios that researchers might expect to encounter in practice :

- **Chapter 5** looks at the synthesis of multiple sources of evidence in the aggregate format;

- **Chapter 6** looks at the synthesis of multiple sources of patient-level evidence;

- **Chapter 7** looks at synthesising a combination of aggregate and patient-level studies

Chapter 8 provides a discussion that relates the findings of the thesis to the original aims and objectives. Furthermore, the work covered in the thesis is considered in relation to the broader implications for cost-effectiveness research in practice for policy-makers.

Chapter 2

Taxonomy and Review of Current Practice

The previous chapter argued that HRQoL evidence, along with other parameters used in CEA, should be identified and synthesised from all available studies in order to avoid the bias in the selection of evidence and to ensure that uncertainty surrounding the HRQoL parameter estimates is fully characterised (Sutton et al., 2000). Yet this would rarely appear to be the case in practice when compared to the synthesis of clinical evidence (Cooper et al., 2005). The objective of this chapter is to review the current state of practice with respect to the use of HRQoL data in health economic modeling and consider this in relation to the evidence space that a researcher might plausibly encounter (Saramago et al., 2012). A review of NICE technology appraisals is carried out in order to explore the methodological landscape in this area. Existing research has found considerable variety in the methods used to select and incorporate evidence into cost-effectiveness models for NICE appraisals (Tosh et al., 2011). This chapter looks more closely at how the format and diversity of the available HRQoL evidence influences the statistical methods employed to derive model inputs. The methods are categorised using an existing taxonomy outlining the possible scenarios faced by the analyst when dealing with available evidence (Saramago et al., 2012). The original study focused mainly upon available techniques for the statistical synthesis of clinical evidence given that the majority of the methodological developments have taken place in this area. The review in this chapter looks at two of the key issues from the study by Saramago and colleagues – the format of the evidence and the number of data sources – with the objective to explore how the choice of methods for analysing and synthesising the available HRQoL evidence depend upon these issues. An additional data feature considered in this chapter is the instrument(s) used to measure HRQoL and

how that might affect the choice of analytical method.

2.1 Methods

This chapter looks to characterize the methodological landscape with regards to the techniques for analysing and synthesising HRQoL evidence for the purposes of health economic evaluation. This section provides a description of the following: the process of selecting NICE technology appraisals for review; the taxonomy employed to categorise methodological approaches for the analysis of HRQoL data; and the process of identifying, selecting and critically appraising information from the NICE technology appraisals.

Selection of NICE Technology Appraisals

The decision to look specifically at evidence submitted to NICE was predicated by two key factors. Firstly, the methodological issues under investigation are only relevant insofar as they have important implications for health policy decision-making. Given that NICE submissions are used to inform real-world policy decisions, these were considered on the grounds that they would provide an understanding of current practice in this context. The second reason for looking at submissions to NICE was more practical; whilst there might be other organisations with a similar remit to NICE, none are as transparent or comprehensive in the reporting of evidence used to inform decisions. Despite focusing upon NICE in this review, it is important to note that the thesis is not solely concerned with NICE but rather deals with issues that are relevant to any organization interested in the evaluation of technology adoption decisions using QALYs.

The sample of NICE technology appraisals selected for review in this chapter includes all Single Technology Appraisals (STA) with recommendations issued in 2012. STAs consider technology adoption decisions for a single health technology, typically a branded pharmaceutical scheduled for release, and differ from Multiple Technology Appraisals (MTA) that evaluates technology adoption decisions for multiple competing health technologies developed for the treatment of patients in the same clinical indication. The STA process was established in 2006 and involves the submission of an evidence dossier, including a health economic evaluation, by the manufacturer of the technology. This differs from the MTA process where there are multiple submissions from each of the relevant manufacturers as well as the submission of a health economic evaluation from an Evidence Review Group (ERG). The decision to focus solely upon STAs was pragmatic and taken in light of the differences that exist in the public availability of reports (www.nice.org.uk) for the two

processes, specifically: (i) the reporting of manufacturers' submissions in the MTA process is extremely limited and (ii) the format selected for the reporting of AGs' submissions differed from that in manufacturers' submissions in the STA process.

Taxonomy

Given that this task is largely a qualitative exercise, a taxonomy has been employed in order to categorise the methods used to analyse and synthesise HRQoL data, as well as identifying areas requiring further research. The taxonomy has been adapted from an existing study that categorised methods according to the possible scenarios that an analyst might encounter when seeking to populate a cost-effectiveness model (Saramago et al., 2012). The original study put forth three factors for consideration when choosing a method for estimating model parameters describing clinical effects.

The first factor to consider is the number of sources of evidence available. Often, the analyst might encounter a scenario where there are multiple studies available addressing the same research question. Where this is the case, Saramago and colleagues emphasized the importance of combining all relevant evidence in order to avoid a biased selection of evidence (also known as evidence synthesis). Meta-analytic techniques are typically used to pool data from multiple (reasonably homogeneous) studies and obtain more precise treatment effects, which is particularly important when these effects are small (Egger and Smith, 1997). Meta-analytic techniques have proved to be extremely popular with regards to the synthesis of clinical evidence for health economic evaluation (Cooper et al., 2005).

A second factor considered as being an important determinant of the statistical methods employed to analyse evidence for the purposes of HEE and included in Saramago and colleagues' taxonomy is the format of the available evidence. The evidence may be available either 1) in aggregate form only (also known as summary data), 2) at the individual-patient level only, or 3) in a combination of aggregate- and individual-level data. Whichever scenario applies, the researcher will typically require a mean estimate of the parameter(s) of interest in order to calculate expected cost-effectiveness, and an associated measure of uncertainty with a view to understanding the uncertainty associated with the decision at hand (Briggs et al., 2006). Individual patient data (IPD) holds a number of advantages over aggregate data (AD). With AD, the validity of the results is constrained by the statistical methods undertaken in the original study; in contrast, access to IPD allows researchers to use a wider range of methods for the analysis and synthesis of the data. Furthermore, there may be the option to explore statistical heterogeneity in the cost-effectiveness results using IPD by controlling for variables describing patient char-

acteristics. This capability means that cost-effectiveness research with IPD can result in more appropriate decision-making when compared to research with AD, particularly when there are substantial differences in the estimated cost-effectiveness for a given intervention across different patient subgroups.

The third and final factor in the taxonomy developed by Saramago and colleagues relates to the number of parameters scheduled for estimation (single or multiple). This feature bears particular relevance to the application of MCMC methods, conducted to explore the uncertainty surrounding the cost-effectiveness results. Where possible, researchers should seek to account for correlation between the various model input parameters when conducting this type of exercise (Ades and Sutton, 2006); otherwise, they risk simulating implausible combinations of input parameters that lead to a misrepresentation of the uncertainty surrounding the cost-effectiveness results and decision problem.

For the purposes of this chapter, it was felt that the contribution by Saramago and colleagues lacked a crucial focal point in the process of utilizing HRQoL evidence for cost-effectiveness analysis; namely, the type of instrument or process for deriving preference-based HRQoL scores. In the interest of promoting the comparability of evidence across technology appraisals, NICE has specified a preference for the valuation of HRQoL using the EQ-5D measurement in cost-effectiveness submissions (NICE, 2008, 2013). The taxonomy has been extended to investigate whether statistical methods are being used to promote consistency in the use of HRQoL evidence. Table 2.1 shows the modified taxonomy for the purposes of the review in this chapter. Note that it no longer accounts for the number of parameters to be estimated; whilst this is still a relevant factor with regards to the analysis of HRQoL data for cost-effectiveness analysis, it has been excluded to prioritize the key factors without overcomplicating the taxonomy.

Data Extraction

The following documents outlining cost-effectiveness studies submitted to NICE were downloaded from its website (www.nice.org.uk) for review:

- **Manufacturer's Submission** This provides detailed information about the cost-effectiveness model submitted by the sponsor to NICE as part of the STA process.
- **Full Guidance** This document outlines the whole appraisal process including the evidence submitted, assessment of this evidence by the nominated ERG and the reimbursement decision taken by the NICE appraisal committee.

Table 2.1: Modified Taxonomy

	Single Source		Multiple Sources	
	EQ-5D	Alternative	Homogeneous Outcomes	Heterogeneous Outcomes
Aggregate Data	A1	B1	C1	D1
Individual Patient Data	A2	B2	C2	D2
Combination	-	-	C3	D3

- **ERG Report** This provides an overview of the ERG's assessment of the evidence submitted to NICE and includes any additional analyses or modifications requested by the group.

Data was extracted to classify the evidence used to derive model inputs according to the gallery of scenarios in table 2.1. Information about the statistical methods employed within each of those scenarios was also extracted. The review was primarily concerned with the data and methods at the parameter level in line with the taxonomy although a series of items were extracted at the model level, particularly with regards to the availability of patient-level data for each appraisal and the consistency of HRQoL evidence in a given model. Table 2.2 outlines the items to be identified and extracted chiefly from the Manufacturer's Submission documents although the other two documents were checked to see if there were any subsequent modifications in the analytical techniques employed. The process of identifying information for extraction was relatively straightforward due to the consistent and well-defined formatting in the NICE documentation.

2.2 Results

Sample Description

A total of 19 NICE STAs are reviewed in this chapter, the details of which are provided in table 2.3. All of the cost-effectiveness analyses identified in the review were model-based rather than trial-based, reflecting NICE methods guidance (NICE, 2013). One appraisal surveyed two models corresponding to different indications (TA252). The models differed in terms of the HRQoL evidence used so each was considered separately. The

Table 2.2: Information Extracted

Model-level findings: issues concerned with format
- Evidence available: was there HRQoL data collected in a clinical trial?
- Evidence employed: was IPD used when available? If not, justification?
- Evidence employed: was the same source of evidence used to populate all of the parameters in the model?
Model-level findings: HRQoL instruments
- Evidence available: where IPD was available, which instrument was collected?
- Evidence employed: where IPD was used, what was the HRQoL instrument?
- Evidence employed: was there consistency in the valuation methods for a given model?
Parameter-level findings: issues concerned with format
- Evidence employed: how many sources of evidence were used to derive the parameters?
- Evidence employed: what was the format of the available evidence?
- Evidence employed: provide a description of the methods employed to analyse the available evidence and derive parameter estimates.
- Evidence available: where available evidence wasn't used, was a justification provided?
Parameter-level findings: HRQoL instruments
- Evidence employed: which instruments were used to measure and value HRQoL?

generalizability of the sample seems to be reasonable given that it covers a range of disease areas, with cancer treatments being the most prevalent.

Model-Level Results

Table 2.4 shows a number of statistics related to the availability and use of HRQoL evidence at the model-level. Almost all of the appraisals acknowledged the availability of HRQoL data in the clinical trial associated with the treatments under investigation; however, only around sixty percent of those submissions with such data utilised it to derive HRQoL model input parameters. Where HRQoL data from clinical trials was disregarded, the following justifications were provided: the data was not collected using the reference case instrument, the EQ-5D (TA250), or similarly that the data collected did not include a preference-based measurement (TA269); there were insufficient numbers of patients experiencing the defined health states set out in the cost-effectiveness model (TA248); there was a low proportion of patients completing the questionnaire (TA269); EQ-5D available but only available in patients representing a subset of health states encountered in the cost-effectiveness model

Table 2.3: List of STAs in the review

STA	Disease Area
244	Respiratory
245	Cardiovascular
248	Endocrine, nutritional and metabolic
249	Cardiovascular
250	Cancer
252	Digestive system/Infectious diseases
253	Digestive system/Infectious diseases
254	Central nervous system
255	Cancer/Urogenital
256	Cardiovascular
258	Cancer/Respiratory
259	Cancer/Urogenital
260	Central nervous system
261	Cardiovascular
263	Cancer
266	Digestive system/Respiratory
267	Cardiovascular
268	Cancer
269	Cancer

(TA254). Two additional studies did not provide any justification, although in both cases it would appear that the data was disregarded on the grounds that the HRQoL outcomes were non-preference-based (TA258, TA261). Interestingly, even when trial evidence was used, it was rarely suitable for estimating all of the HRQoL input parameters in a given model; only two submissions relied entirely upon trial data for the estimation of the HRQoL parameters (TA260, TA267). One notable issue was the inability of clinical trials to capture the HRQoL effects of acute health events represented in the cost-effectiveness model, e.g. stroke or myocardial infarction (TA244, TA249, TA266). Clinical trials developed to capture data at defined time points may struggle to capture the HRQoL impact of acute health events if the effects have elapsed between fixed follow-up visits (Bansback et al., 2008).

Parameter-Level Results

Table 2.5 shows the use of HRQoL evidence in the sample categorised according to the taxonomy. In the majority of cases, individual parameters (e.g. HRQoL values for a given

Table 2.4: Results at the model level

Evidence available	**Frequency**
Was there HRQoL data collected in a clinical trial?	16/19 (84%)
Where IPD was available, which instrument was collected?	
- EQ-5D	8/16 (50%)
- Alternative PBM	1/16 (6%)
- Disease-specific patient-reported outcome	7/16 (44%)
Evidence employed	
Was IPD used when available?	10/16 (63%)
Where IPD was used, which instrument was used?	
- EQ-5D	6/10 (60%)
- Alternative PBM	1/10 (10%)
- Disease-specific patient-reported outcome	3/10 (30%)
Was the same source of evidence used to populate all of the parameters in the model?	4/19 (21%)
Was there homogeneity in the valuation methods for a given model?	
- Yes	7/19 (37%)
- No	10/19 (53%)
- Unclear	2/19 (11%)

health state in the model) were informed by a single source of evidence and heavily reliant upon aggregate data from existing studies. Table 2.6 provides a more detailed breakdown of the evidence according to the method of valuation. Despite NICE having specified a preference for EQ-5D data, a large proportion of parameters employed alternative outcome measures.

2.3 Taxonomy

This section outlines the analytical methods employed for the derivation of HRQoL inputs across the range of scenarios set out the taxonomy in Table 2.1.

Scenarios A1 and B1

This section includes those scenarios where only one relevant source of aggregate data has been identified for the purposes of informing a model input parameter. The review

Table 2.5: Categorization of parameters by scenario

	Single Source			Multiple Sources			Total
	EQ-5D	Alternative	Unclear	Homogeneous Outcomes	Heterogeneous Outcomes	Unclear	
AD	73/193 (38%)	56/193 (29%)	4/193 (2%)	4/193 (2%)	7/193 (4%)	2/193 (1%)	146/193 (76%)
IPD	11/193 (6%)	5/193 (3%)	0/193 (0%)	2/193 (1%)	15/193 (8%)	0/193 (0%)	33/193 (17%)
Comb	-	-	-	12/193 (6%)	2/193 (1%)	0/193 (0%)	14/193 (7%)
Total	149/193 (77%)			44/193 (23%)			

found, perhaps unsurprisingly, that there were no analytical methods employed in such scenarios for the derivation of model estimates. Typically, mean estimates of HRQoL associated with health states in the model (along with evidence to account for uncertainty surrounding these values) were selected from an existing study, usually in the published literature, and used directly in the cost-effectiveness model. For example, one model in the review (TA249), evaluating treatments for the prevention of stroke and systemic atrial fibrillation, used HRQoL estimates for three health states (mild, moderate and major post-stroke health states) that were directly extracted from one study. Ideally, estimates selected for use in the cost-effectiveness model should be based upon mean values in line with a rationale set out in the cost literature which says that summary statistics other than the arithmetic mean are inappropriate for health policy decisions because they don't take all of the observations in the sample into account (Thompson and Barber, 2000). Moreover, the population average impact of a strategy is required for policy decisions, which can only be obtained by multiplying the mean treatment effect by the number of individuals that need treatment (Briggs and Gray, 1999).

Table 2.6: Parameter-level results

Instrument	
EQ-5D (reference case)	84/193 (44%)
EQ-5D (Non-UK value set)	1/193 (<1%)
EQ-5D (mapping)	17/193 (9%)
Alternative Generic PBM	5/193 (3%)
Disease specific PBM	1/193 (<1%)
Valuation exercise (TTO, SG, VAS)	
- General public valuing vignettes (UK population)	19/193 (10%)
- General public valuing vignettes (non-UK population)	5/193 (3%)
- Patient values	16/193 (8%)
- Health professional values	2/193 (1%)
- Lack of information	12/193 (6%)
Combination	
- Homogeneous (EQ-5D)	15/193 (8%)
- Homogeneous (Other)	3/193 (2%)
- Heterogeneous	7/193 (4%)
- Unclear	2/193 (1%)
Not explicitly stated	4/193 (2%)

Scenarios A2 and B2

Evidence available at the IPD level might come from a randomised controlled trial (RCT), observational study or clinical registries. These data may have either been collected at a single point in time (cross-sectional data) or across multiple time points for each individual (repeated measures data). In total, there were sixteen cases in the review, relating to four separate cost-effectiveness models that could be categorised as being either in scenario A2 or B2 (TA249, TA255, TA266, TA267).

The simplest approach in these scenarios involved the calculation of mean preference-based HRQoL scores in a patient population or in patients groups relevant to the model input of interest. One model (TA266), evaluating treatments for cystic fibrosis, used patient-level HRQoL data from a clinical trial to populate model input parameters. First, baseline HRQoL in the model was estimated on the basis of the mean score at baseline in the trial. In addition, the data was used to provide estimates of change in HRQoL at week 14 relative to baseline for each of the treatments under consideration. Mean estimates of change in HRQoL were calculated for those patients who experienced an improvement in respiratory symptoms as well as for those patients who did not experience an improvement.

There was one example identified in the review where regression methods were applied to IPD to estimate multiple HRQoL input parameters (TA267). The cost-effectiveness model, evaluating treatments for chronic heart failure, had a Markov cohort structure with two health states (alive and dead) and captured HRQoL using a measure of disease severity (the New York Heart Association classification or NYHA). All of the HRQoL evidence used in the model came from one study, a subset of a clinical trial. In the study, where patients were treated with either of the therapies under evaluation in the model, EQ-5D data was collected at repeated time points.

Model inputs were derived from this dataset by analysing the impact of NYHA upon EQ-5D using a mixed regression model specifically designed to handle multilevel data. The analysis also accounted for the effect of patients' baseline characteristics, the rate of hospitalisation and the impact of the treatment group. The methodology employed for the derivation of model inputs in this particular model was notable because it accounted for the impact of treatment upon HRQoL over and above that explained by the model structure. Typically, HRQoL inputs are associated with health states or events and assume that these estimates are independent of the treatment that patients are receiving. The methods employed in NICE appraisal 267 show that, with suitable evidence, this assumption can be tested. Furthermore, regression methods might be considered for the analysis of HRQoL

data at the IPD level for a number of other reasons. As previously mentioned, one of the benefits of having access to IPD is the potential to explore heterogeneity in cost-effectiveness estimates across patient subgroups. Whilst there were no such examples in the review, others can be found in the published literature (Henriksson et al., 2008). Alternatively, one might use regression methods to analyse HRQoL data in order to identify and control for confounding factors that mask the true estimates of interest.

Scenarios C1 - C3 and D1 - D3

The review identified three types of methods involving the use of multiple sources of evidence for the estimation of HRQoL parameters. Despite the extensive use of meta-analytic methods for the synthesis of clinical evidence (Cooper et al., 2005), only two models in the review were found to employ similar methods for the synthesis of HRQoL evidence (TA244, TA255). Four of these cases pertained to a model evaluating treatments for metastatic hormone-refractory prostate cancer (TA255) that accounted for the impact of several adverse events upon HRQoL, including pulmonary embolism, febrile neutropenia, fatigue and nausea. A literature search identified two studies with HRQoL estimates associated with the impact of pulmonary embolism and two separate studies both providing HRQoL estimates associated with the impact of febrile neutropenia, fatigue and nausea. For three of the health states (febrile neutropenia, fatigue and nausea), the decision to synthesise evidence from multiple studies looks justifiable on the grounds that the valuation methods were homogeneous. Unfortunately, it was not possible to ascertain the homogeneity of the evidence for the pulmonary embolism health state due to a lack of reporting.

Provided that patient-level datasets are sufficiently homogeneous, they can be pooled using evidence synthesis methods for the derivation of HRQoL model inputs (either scenario C2 or D2). As with a single source of IPD, pooled IPD requires some form of analysis to derive model inputs with the same analytical techniques being applicable. There were two cases in the review where model inputs were derived using patient-level datasets that pooled multiple sources of data, although unfortunately the analytical methods used for the derivation of the model inputs were not specified (TA244).

There were a number of cases where a single source of evidence was employed when in fact multiple sources were available with no obvious grounds for the restricted approach. One noteworthy example of this can be found in TA254, an appraisal evaluating treatments for multiple sclerosis. In here, EQ-5D data was collected in the clinical trial but this was discarded from the model in the original submission on the grounds that the study population only captured a subset of the health states set out in the model; instead, evidence was

derived from a published study. However, experts reviewing the cost-effectiveness model on behalf of NICE (the ERG) felt that the original submission had excluded potentially relevant evidence, including the trial data, and so conducted an additional analysis that involved taking the average of estimates from the trial data combined with those derived from the published study. There were further examples in the review where analysts appeared to have access to multiple sources of evidence but only opted to select a single source with no discernable justification for doing so (TA244, TA245, TA249, TA256, TA263). In many cases, alternative sources of evidence were utilized in sensitivity analyses.

Another method involving the use of multiple sources of evidence for the estimation of HRQoL parameters is the adjustment of HRQoL values (twenty-three cases in total). Given that HRQoL evidence is rarely collected specifically for the purposes of informing a cost-effectiveness model, analysts often rely upon evidence collected in patients whose characteristics do not exactly match those set out in the model. In such circumstances, analysts might consider combining multiple HRQoL scores to adjust for factors such as comorbidities or age (Ara and Wailoo, 2011). In most cases of adjustment, the evidence being combined appeared to be homogeneous in terms of the valuation methods employed. However, there were five cases of adjustment involving heterogeneous HRQoL values, all in TA 256, two of which combined EQ-5D and standard gamble valuation and three that combined EQ-5D valuations derived using different value sets.

There were nine cases in the review which employed adjustment techniques to account for comorbidities, five of which applied a multiplicative approach, two that applied an additive approach, whilst the remaining cases were unclear. For example, there was one case in the review where literature searching was unable to identify HRQoL values for patients in a post myocardial infarction (MI) health state with atrial fibrillation (AF). However, there was one estimate for post-MI patients without AF (data in AD format) and another for patients with AF without an MI (data also in AD format). An adjusted value was then estimated by multiplying these two values together, also known as the multiplicative approach. In contrast, the additive approach involves applying the estimated decrement in HRQoL associated with one of the comorbidities to the absolute HRQoL value associated with the other health state. Unfortunately, the choice of approach was not justified by the authors. There were a further twelve cases of adjustment to account for the direct effect of different treatments (all in TA252). In each case, the baseline value in the cost-effectiveness model was adjusted (using the additive approach) according to the treatment received. In addition, there were two cases of adjustment for age but unfortunately the methods weren't clearly specified (TA268).

A third methodology observed in the review that involved the combination of evidence from multiple sources was the use of mapping techniques for the indirect estimation of model input parameters. Mapping, also known as cross-walking, refers to the prediction of HRQoL values from an alternative outcome measure (Longworth and Rowen, 2011). Relevant PBM values are not always available for the derivation of model inputs, especially where a reference case measurement has been nominated, as is the case with NICE. In this type of situation, analysts may turn to using evidence measured with an alternative outcome measure in combination with a mapping algorithm, derived using an external dataset, that links the reference case measure (as the dependent variable) to the outcome measure available (the explanatory variable(s)). This algorithm might have either been derived by the analyst using patient-level data or have come from an existing study in the aggregate data format.

In total, there were seventeen model parameters spanning three separate models derived via mapping techniques in the review. Owing to the fact that there is more than one outcome measure involved when mapping, all of these cases have been categorised in scenarios D1-D3. In two of the models, the input parameters were derived by applying mapping algorithms to disease-specific patient-reported outcomes collected in clinical trials (TA259 and TA260). In each case, the algorithms were predicted from IPD using ordinary least squares (OLS) estimation techniques. The third model utilized a disease-specific patient-reported outcome measure from a clinical trial (IPD) but the mapping process was slightly different (TA268). Instead, a set of values (AD) associated with the disease-specific measure was identified in the published literature (Rowen et al., 2011). This last case does not constitute mapping in the traditional sense but rather the use of disease-specific preference weights.

2.4 Discussion

The review in this chapter set out to identify the statistical methods employed to derive HRQoL parameters for cost-effectiveness models and explore the factors influencing the approach employed. The review identified four separate applications involving the use of statistical methods to analyse HRQoL evidence for the purposes of cost-effectiveness analysis. The first application concerns the methods employed when the analyst has IPD, specifically in scenarios A2 or B2 of the taxonomy. Unfortunately, the methods employed to derive model parameters in these scenarios were rarely set out explicitly, although there was one notable exception, the NICE appraisal TA267. Despite the majority of

appraisals having accessible trial data with HRQoL outcome measures, IPD played a limited role in the population of the associated parameters in the cost-effectiveness model. Two issues related to the collection of HRQoL data in clinical trials underpin this finding. Firstly, just under half of the trials with HRQoL evidence collected measures without associated preference weights, necessary for the estimation of HRQoL values. Whilst some of the submissions overcame this problem by employing mapping procedures, for others this provided sufficient grounds for discarding the evidence. Another observed flaw with trial data contributing to this low usage was the inability of trials to capture HRQoL effects – those typically associated with defined health states and events – captured within the cost-effectiveness models. More generally, this points to a broader problem regarding the collection of HRQoL data for cost-effectiveness models; such data is typically collected as a secondary outcome measure, along the lines of clinical outcome measures, with the intention to estimate the effects of the treatments under evaluation. These findings would suggest that more consideration should be paid to the HRQoL requirements of cost-effectiveness models in practice.

Another application requiring statistical methods to derive HRQoL parameters for cost-effectiveness analysis is the use of pooling or meta-analytic techniques to combine HRQoL evidence from multiple sources. As previously mentioned, these methods enjoy widespread usage for the synthesis of clinical evidence (Cooper et al., 2005). Only a handful of cases were identified in this review and those conducted synthesised evidence either by pooling or calculating the mean. The lack of more formal synthesis techniques applied in this context could be interpreted in two ways. Firstly, this could be viewed as being a consequence of a limited evidence base; that is, meta-analytic techniques are rarely ever necessary because analysts typically only have access to a single source of HRQoL evidence, at best. However, there were many cases in the review where multiple sources of evidence were available and synthesis techniques were not considered. This points to a second possible explanation – a lack of recognition that HRQoL evidence, as with all other parameters, should be identified and selected in a comprehensive manner to avoid bias. This second point would also indicate a lack of exemplars and guidance with regards to the synthesis of HRQoL evidence.

The third and fourth applications requiring statistical methods to derive HRQoL parameters – mapping and adjustment techniques – are similar in that they both seek to derive inputs where there is no evidence of direct relevance to the cost-effectiveness model. The relevance of the evidence depends upon a number of factors including: (a) the instrument used to value HRQoL, especially where a reference case measurement has been specified;

(b) whose values were used (patients, general population, other) and whether they reflect the jurisdiction in question (e.g. country-specific values); (c) the degree to which the patient population (or health state description) in the study corresponds to that in the model. Whilst analysts might not have access to HRQoL evidence fulfilling all of the desired criteria, statistical methods may be employed to derive indirect estimates using evidence satisfying at least some of the criteria.

One approach involved the adjustment of HRQoL values from studies where the patient characteristics did not correspond to those set out in the model, particularly with regards to factors such as age and comorbidities. The findings of the review would suggest that there is no consistency with regards to the application of adjustment techniques. Where adjustment techniques were used there were no justifications provided for the specific approach selected (either additive or multiplicative). More generally, these techniques did not appear to be employed routinely throughout the appraisals in the review. Many of the models reviewed accounted for the effects of changes in health status by applying a 'disutility' – that is, a decrement in HRQoL – to the baseline HRQoL value. Despite this, only there were only a handful of cases recognizing that found to have explicitly acknowledged this issue.

Another approach involving the indirect estimation of model inputs is the combination of evidence from multiple sources using of mapping techniques. As with the utilization of adjustment techniques, there did not appear to be consistency regarding the usage of mapping techniques. Of the seven appraisals that reported having access to a disease-specific measure of HRQoL in the IPD format, only three made use of this evidence using mapping techniques. For the remaining appraisals, this evidence was typically discarded despite the fact that there appeared to be relevant mapping algorithms available in two of these cases.[1] However, neither of the two appraisals in question utilized values derived by the reference case measurement, the EQ-5D. Whilst NICE has stated that mapped EQ-5D values are regarded as being "second-best" when compared to EQ-5D scores derived directly, it is unclear whether they are also inferior to values obtained via other means.

[1] TA250 reported having access to data for the EORTC-QLQ C30 and two studies were found containing relevant mapping algorithms (Crott and Briggs, 2010; McKenzie and Van Der Pol, 2009). TA 269 reported having access to data for the FACT-M and one study was found containing a relevant mapping algorithm (Askew et al., 2011).

Limitations

Whilst the sample selected for review in this chapter would appear to be generalizable in the sense that it covers a range of disease areas, it is not without limitations. Firstly, the review only considers cost-effectiveness evidence developed by pharmaceutical manufacturers seeking a positive adoption decision from NICE with regards the health technology under review. It may be the case that manufacturers have a propensity to use different methods from those that might be used by an independent assessment group; in particular, one could argue that manufacturers may have an incentive not to be comprehensive in the selection of evidence. Whilst the review might only provide a partial viewpoint of methodological practice in terms of the groups developing cost-effectiveness models, this is not to say that it is unusable. In fact, the main finding of this chapter – the inconsistent application of statistical methods – transcends the issue of which group was responsible for submissions in the review and points to a broader issue regarding a lack of methodological guidance.

Another issue concerning the generalizability of the findings in this chapter is the exclusive focus upon NICE and cost-effectiveness evidence in the UK setting. As such, these results are not necessarily indicative of practice globally. Nevertheless, NICE appraisals were viewed as representing the best resource for the purposes of the review given the comprehensiveness in the reporting of the methods. Even so, the reporting of the evidence review process was limited and this meant that it was difficult to ascertain whether all relevant evidence was put into practice.

Finally, one might question whether the scope of the review is limited given that it deals exclusively with model-based cost-effectiveness studies and there may be plausible scenarios involving trial-based cost-effectiveness studies. The fact that all of the cost-effectiveness submissions in the review employed a model-based approach is unsurprising given that this method is considered to be the best method for adequate representation of costs and outcomes of interest to policy makers.

Priorities for Future Research

The review of NICE technology appraisals outlined in this chapter has highlighted a number of fundamental inconsistencies regarding the application of statistical methods for the analysis of HRQoL evidence. Synthesis, mapping and adjustment techniques were employed on an ad hoc basis and this raises concerns over the degree of comparability across cost-effectiveness studies. Although HRQoL evidence should be identified and synthesised

comprehensively to avoid biased selection of evidence, as well as adequately characterizing the associated uncertainty, the results in this chapter would suggest that this is not the case in practice. Regardless of whether there are AD or IPD, from a single study or multiple ones, researchers will often have to deal with a HRQoL evidence base that is heterogeneous in terms of the patient-reported outcomes available. As such, the development of methods for the comprehensive utilization of HRQoL evidence, particularly in the face of heterogeneous outcome measures, should be a priority for future research. The next chapter looks at the existing methodological literature in this area to explore the statistical techniques available for the synthesis of HRQoL evidence. A second issue established in the review is the problem of data collection. One of the key factors complicating the use of HRQoL data is the fact that evidence is rarely collected specifically for the purposes of populating a cost-effectiveness model, i.e. so that one could estimate the HRQoL effects associated with defined health states and events. Although further work is required in relation to this issue, it will not be a research priority in this thesis.

Chapter 3

Review of the Methodological Literature

3.1 Introduction

The previous chapter identified a number of inconsistencies in the use of statistical methods to derive HRQoL parameters for CEA. One trend of particular concern was the utilization of a single source of evidence in scenarios where multiple sources were in fact available, with no obvious grounds for the restricted approach. This approach implies a partial representation of the evidence base and increases the risk of obtaining misleading results. This chapter aims to evaluate the methods and guidance relating to the synthesis of HRQoL evidence in HEE, to offer guidance where possible and identify areas where future research would be worthwhile. A citation-search strategy is employed in order to identify any relevant studies in this area. In addition, documents containing methodological guidance issued on behalf of bodies employing health economic evaluation are also selected.

3.2 Literature Review: Methods

Following recommendations from Hinde and Spackman, a citation searching approach was considered to be most appropriate approach for identifying relevant literature given the discursive nature of the review (Hinde and Spackman, 2015). This recommendation was based upon an illustrative methodological review showing that a citation-searching approach, specifically bidirectional citation searching to completion (BCSC), was able to identify a larger number of relevant studies that the standard Boolean logical approach, based upon a keyword search.

For the purposes of this review, the BCSC approach was employed, starting with four studies as the initial pearls (Ara and Wailoo, 2011; Papaioannou et al., 2011; Peasgood and Brazier, 2015; Saramago et al., 2012). These studies were selected based on prior knowledge of their relevance to the review. The next step involved appraising the studies referenced by the initial pearls, as well as the studies found to cite the initial pearls, and selecting those studies identified as being relevant to the review. This process was repeated until no additional studies were identified.

Additional search techniques were required in order to capture relevant information in the grey literature (i.e. information outside of academic publishing), specifically any methodological guidance issued by policy-making bodies utilising economic evidence to inform health technology adoption recommendations. The documents were identified from an existing review conducted on behalf of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) (Eldessouki and Dix Smith, 2012). This was considered to be a credible source given the important role that ISPOR has played globally in advancing the science and practice of CEA in health policy over the past twenty years.

3.3 Literature Review: Results

In total, the review found nineteen potentially relevant studies, in addition to the four references used as the initial pearls. These studies can be divided into two categories: (i) applied meta-analytic studies involving the synthesis of HRQoL evidence, and (ii) methodological studies providing discussion without any empirical evidence. The grey literature review identified thirty-eight methodological guidance documents.

3.3.1 *Applied Studies*

A total of seventeen applied meta-analytic studies were identified as being potentially relevant. One study was excluded on the grounds that it only provided a protocol outlining plans for a systematic review and meta-analysis of HRQoL values associated with diabetic retinopathy health states (Sampson et al., 2015). Another study was excluded given that it did not pertain to the synthesis of preference-based HRQoL values (Kinney et al., 1996). It would appear that all of the remaining studies synthesised aggregate data only (i.e. no IPD), given that they referred to the identification and selection of evidence from the published literature. Only two out of the fifteen remaining studies examined the direct effects of health care interventions upon HRQoL (Cheng and Niparko, 1999; Wyld et al., 2012). The remainder of the studies instead focused upon the estimation of HRQoL

values associated with defined health states or events; for example, Lung and colleagues combined evidence using a meta-regression approach to explore the impact of diabetes and its related complications upon HRQoL values (Lung et al., 2011). The fact that the majority of studies deal with defined health states and events (rather than treatment effects) is unsurprising given that most cost-effectiveness studies are model-based and estimate QALYs via surrogate health outcomes (see Chapter 2).

A descriptive outline of each of the applied studies is provided in table A.1 (see Appendix A). One of the most notable observations in this review is the issue of how to handle heterogeneity when dealing with different outcome measures. A selection of studies might be considered to be homogeneous if they are all capturing a common parameter estimate. However, this assumption relies upon the studies being identical in terms of the factors that may influence the dependent variable, including the patient populations involved and the exposures or interventions under investigation. In practice, some degree of heterogeneity is inevitable and, provided that the studies are assumed to be sufficiently similar for meta-analysis to be considered viable, two methods have been proposed to deal with this: (i) the inclusion of study-level covariates, also known as meta-regression, and (ii) embracing a random effects modeling approach (Higgins et al., 2009).

Meta-regressions seek to explain heterogeneity in parameter estimates obtained through the synthesis of evidence from multiple studies in relation to one or more characteristics of the studies involved (Thompson and Higgins, 2002). This approach featured in the majority of applied studies identified in the review as a means to explore variations in HRQoL values according to disease stage or health state. Covariates were also included to control for imbalances in study-level characteristics such as age and gender. A large proportion of studies also included dummy-variable covariates in an effort to overcome the lack of comparability existing between alternative instruments for valuing HRQoL (Conner-Spady and Suarez-Almazor, 2003). Two studies took a different approach and selected a reference case measurement, the EQ-5D, and exclusively synthesised evidence pertaining to this measure (Doth et al., 2010; Peasgood et al., 2009). Although this method might alleviate concerns about a lack of comparability between instruments, restricting searches to a reference case measure may provide an incomplete representation of HRQoL effects or, in some cases, no evidence at all. Two further studies opted to undertake separate analyses for different outcome measures (Liem et al., 2008; Mohiuddin and Payne, 2014).

The random-effects approach to evidence synthesis is often utilized as a means of controlling for unobservable heterogeneity between studies (Higgins et al., 2009). In a Bayesian

context, the concept of exchangeability is often used to consider the a priori judgments regarding the similarity of evidence (Lunn et al., 2012). A large proportion of applied studies identified in the review were found to employ random-effects models to synthesise HRQoL evidence. Another trend with important methodological implications is the fact that many of the applied meta-analytic studies used values obtained from the same study. The standard approach to meta-analysis, which assumes that all data points are independent of one another, is inappropriate when observations are in fact correlated; specifically, this can lead to biased estimation of the standard errors associated with the parameter estimates. As such, it would seem incorrect to assume independence between values when some have been obtained from the same study and so different assumptions are necessary. A number of studies employed a clustering approach that treats the data clusters (i.e. values from the same study) as being independent of one another (Djalalov et al., 2014; Peasgood et al., 2010; Wyld et al., 2012). Other studies chose to employ hierarchical modelling techniques (McLernon et al., 2008; Sturza, 2010; Tengs and Lin, 2002, 2003). This method is considered to be more flexible than the clustering method given that it allows for some exchangeability between the clusters (Baio, 2012).

3.3.2 *Methodological Studies*

The BCSC search strategy identified two methodological studies (Ara and Wailoo, 2012; Papaioannou et al., 2013), although both of these were adapted versions of references used as the initial pearls (Ara and Wailoo, 2011; Papaioannou et al., 2011). There were two additional methodological studies also used as initial pearls (Peasgood and Brazier, 2015; Saramago et al., 2012). A report commissioned by NICE recognised that whilst NICE does not require a formal quantitative synthesis of HRQoL evidence, there may be situations where the use of such methods is warranted (Ara and Wailoo, 2012). The same report recommended that where HRQoL values are sufficiently homogeneous then pooling should be considered as a way to improve the precision of the estimates of the mean utility values and their variances. Although the report does not provide an explicit clarification regarding the necessary degree of homogeneity, it does suggest that this could refer to those scenarios where values have been "collected in the same patient population using the same instrument and using the same UK value (set)", which may only apply in a selected number of scenarios. Another report issued by the DSU identified the challenge of synthesizing HRQoL evidence in practice given the heterogeneity that is often observed between studies (Papaioannou et al., 2011). The authors noted that the methods for undertaking this type of synthesis are unclear and that further research is necessary.

Much of the methodological discourse regarding the synthesis of HRQoL evidence focuses upon the issue of between-study heterogeneity. Saramago and colleagues identified three challenges in this regard: (i) variation in the instrument used to derive HRQoL values; (ii) variation in the population whose values are used; (iii) inadequacy of the available statistical methods in the handling of these issues (Saramago et al., 2012). The authors recommended the prioritisation of future research efforts concentrated on the development of methods for synthesising heterogeneous PRO measures. This call is echoed by a recent study by Peasgood and Brazier (2015); they argue that the inclusion of covariates to account for differences in valuation methods (in a meta-regression) is unlikely to pick up the relative weights attributed to the different domains. The differences between alternative preference-based measures of HRQoL can be attributed to a number of factors – differences in their descriptive systems, as well as the techniques used to obtain the associated value sets – and it is unlikely that a dummy variable will be able to characterise these differences.

Peasgood and Brazier (2015) point to the use of mapping techniques as a potential means of resolving the differences between values from heterogeneous outcome measures (Peasgood and Brazier, 2015). As previously mentioned, mapping involves the prediction of HRQoL values for a reference case instrument using scores or responses to some other outcome measure in combination with a mapping algorithm linking the two outcome measures. Another appealing prospect associated with the use of mapping techniques would be the ability to draw upon non-preference-based measures of HRQoL, which could uncover a much greater evidence base for researchers to draw upon. Despite the potential of the mapping approach, Peasgood and Brazier (2015) are careful to acknowledge that these techniques have a number of limitations, especially the prediction errors that occur as a result of their usage. There is already a significant body of literature devoted to methodological issues associated with the development of mapping algorithms (Longworth and Rowen, 2011). The next section outlines some of these issues and considers them in relation to a methodology for evidence synthesis.

3.3.3 *Grey Literature*

There was only one document found in the grey literature that made a reference to the synthesis of HRQoL evidence, issued on behalf of the Health Intervention and Technology Assessment Program in Thailand (Sakthong, 2008). However, the reference was limited to a single remark recognising the need for "a systemic approach including meta-analysis... to combine utilities taken from different studies". Whilst NICE makes no reference to the synthesis of HRQoL evidence in its 'Guide to the methods of technology appraisal' (NICE,

2013), further guidance developed by the Decision Support Unit (DSU) – a body commissioned by NICE to provide research and educational activities – makes recommendations with regards to this matter. One of the reports written by the DSU recognised that whilst NICE does not require a formal quantitative synthesis of HRQoL evidence, there may be situations where the use of such methods is warranted (Ara and Wailoo, 2011). The same report recommended that where HRQoL values are sufficiently homogeneous then pooling should be considered as a way to improve the precision of the estimates of the mean utility values and their variances. Although the report does not provide an explicit clarification regarding the necessary degree of homogeneity, it does suggest that this could refer to those scenarios where values have been "*collected in the same patient population using the same instrument and using the same UK value* (set)", which may only apply in a selected number of scenarios. Another report issued by the DSU identified the challenge of synthesising HRQoL evidence in practice given the heterogeneity that is often observed between studies (Papaioannou et al., 2011). The authors noted that the methods for undertaking this type of synthesis are unclear and that further research is necessary.

3.4 Discussion

Researchers undertaking economic evaluations may find that they have more than one relevant source of HRQoL evidence for the estimation of QALYs. Ideally, HRQoL evidence should be identified and synthesised from all available studies in order to avoid bias in the selection of evidence and to ensure that uncertainty surrounding the estimates is fully characterised (Sculpher et al., 2006). However, chapter 2 found that there are fundamental inconsistencies regarding the use of statistical methods for the purposes of achieving this objective. Guidance on this matter – typically issued on behalf of the policy-making bodies that use economic evaluation to inform health technology adoption decisions – was found to be severely lacking. In the case of NICE guidance, synthesis of HRQoL evidence from multiple sources is recommended but only where studies are sufficiently homogeneous (Ara and Wailoo, 2011). Observing this recommendation in the strictest sense would imply synthesis should only be considered in HRQoL values captured using the same instrument given the lack of comparability in values derived via different instruments (Conner-Spady and Suarez-Almazor, 2003).

A review of the methodological literature found that researchers looking to synthesise HRQoL evidence have struggled contend with two competing objectives. A review of the methodological literature found that researchers looking to synthesise HRQoL evidence

have struggled to strike a balance between achieving both completeness – that is, making use of all available HRQoL evidence, irrespective of the outcome measure – and comparability – where HRQoL effects can be compared on a common scale. This situation is far from ideal as it brings a potential risk of obtaining misleading results as a consequence of the partial representation of the evidence base. A recent study has suggested that this problem might be circumvented through the use of mapping techniques (Peasgood and Brazier, 2015). There is already a significant body of literature devoted to methodological issues associated with the development of mapping algorithms (Longworth and Rowen, 2011). The next chapter outlines some of these issues and considers them in relation to a methodology for evidence synthesis.

Chapter 4

Methods for Synthesising Heterogeneous HRQoL Evidence

4.1 Introduction

Traditionally, mapping has been seen as a second-best option for the estimation of preference-based HRQoL values, when compared to the collection of values first-hand (NICE, 2008). As such, the application of mapping techniques has only been advocated in those scenarios where direct evidence is unavailable and, even with these restrictions, concerns exist about the validity of mapped estimates of HRQoL (Longworth and Rowen, 2011). However, a recent study has proposed their usage more broadly as a means of making comprehensive use of available evidence whilst also capturing HRQoL effects on a comparable scale (Peasgood and Brazier, 2015). This chapter considers the potential role of mapping techniques in combination with evidence synthesis techniques. It starts by reviewing some of the statistical issues associated with the development of mapping algorithms. It then considers the bias introduced when mapping algorithms do not account for measurement error across all outcome measures and how this can be avoided by employed structural equation modeling techniques. Finally, the structural equation modeling framework is described in relation to the challenges involved in the analysis of preference-based measures of HRQoL

4.2 Statistical Methods for Mapping

Much of the methodological debate in this area has revolved around the development of statistical methods capable of overcoming the shortcomings of standard regression methods for the analysis of preference-based measures of HRQoL. Despite the popularity and simplicity of the ordinary least squares (OLS) approach to regression analysis, the main

drawback of this method is that it does not guarantee that predictions will lie within a plausible range – OLS can lead to the prediction of values greater than 1, the highest value achievable (Chuang and Kind, 2009).

The methodological literature aimed at overcoming this problem can be broadly categorised into two groups. The first of these groups has focused upon the development of methods better suited to characterising the distribution of HRQoL index values. Regression models based upon the beta distribution have been put forward to account for the bounded nature of health index values (Basu and Manca, 2012; Hunger et al., 2011). The rationale for the beta regression is that the dependent variable is assumed to be restricted between 0 and 1. A simulation exercise showed these methods to be more robust for the estimation of covariate effects compared to OLS (Basu and Manca, 2012). However, empirical studies have shown this approach to be comparable to OLS in terms of predictive performance (Basu and Manca, 2012; Hunger et al., 2011).

Hernandez Alava and colleagues developed a model to capture some specific distributional features of the EQ-5D (with a UK value set) (Alava et al., 2012, 2014). This model incorporated an adjusted limited dependent variable model to deal with the gap in EQ-5D values between 0.883 and 1, as well as restricting values above 1. For the remainder of the distribution, a mixture model was employed to account for the multimodality observed in the EQ-5D values, in part explained by the weightings assigned to the N3 term in the UK value set. This model was defined as the adjusted limited dependent variable mixture model (ALDVMM) and the initial studies indicate that it performs well in terms of predictive performance (Alava et al., 2012, 2014). Whilst other methods have been employed to analyse HRQoL index values, on-going research in this area has indicated that the Beta and ALDVMM approaches hold the most potential.[1].

An alternative strand of mapping research, known as response mapping, analyses responses to the health state descriptions of preference-based measures as opposed to the health index values predicted using such responses. Response mapping has an intuitive appeal because it deals directly with subjects' responses and therefore ensures the prediction of feasible health index values. Another advantage is that the predicted response values can be used in different countries with country-specific valuations (Rivero-Arias et al., 2010).

The modelling approach required for response mapping is fundamentally different from the previous techniques dealing with index values given that it involves analysing categorical outcomes. In many of the existing response mapping studies, a separate analysis was

[1]See research grant http://gtr.rcuk.ac.uk/project/5A46044C-DC5B-4A0D-B21E-863FCFF87114

conducted for each of the different dimensions of the health state descriptive system with modelling techniques including multinomial logit (Gray et al., 2006), ordered logistic (van Hout et al., 2012) and ordered probit (Alava et al., 2014). There have been concerns raised over whether or not response mapping should account for correlations between dimensions (Alava et al., 2014). A recent study by Coniglian and colleagues (2015) has demonstrated that dependences between dimensions can be modelled using a multivariate approach (Conigliani et al., 2015).

The outputs of these models are not especially intuitive given that the outcomes have to be transformed to facilitate their analysis. In the case of the logistic models, parameter estimates show the impact of predictors upon the log-odds of a given response level whilst the ordered probit captures effects in terms of the standard normal distribution. However, these outputs can be converted into probabilities such that there is a probability associated with each of the potential responses within a given dimension according to a given set of predictors. With these probabilities, one can predict HRQoL values by combining them with the associated value sets either using an expected utility approach, using the response with the highest probability or using Monte Carlo procedure (Gray et al., 2006).

The evidence pertaining to the predictive performance of response mapping techniques is mixed but overall comparable performance to OLS (Chuang and Kind, 2009; Gray et al., 2006; McKenzie and Van Der Pol, 2009; van Hout et al., 2012). Hernandez-Alava and colleagues found that predictive performance of the generalized ordered probit model was worse than the ALDVMM (Alava et al., 2014). Despite this, one advantage of the response mapping approach over the ALDVMM is that, unlike the ALDVMM, it does not apply to a single instrument and value set. Furthermore, response mapping ensures the prediction of feasible health index values whilst also providing a more detailed understanding of the relationships between different outcome measures (Dakin et al., 2013).

4.2.1 *Measurement Error and Mapping*

Although conventional wisdom holds that mapped estimates of HRQoL are second-best to observed estimates for use in cost-effectiveness studies, recent studies have proposed a rationale for the use of mapped estimates of HRQoL in cost-effectiveness analysis (Lu et al., 2013). This concept relies upon an alternative mapping approach, known as the common factor model (CFM), which characterises the relationship between outcome measures in terms of a shared, latent factor. Not only does this approach potentially pave the way towards the synthesis of heterogeneous HRQoL outcome measures (Lu et al., 2014), it is also claimed to potentially deliver more efficient HRQoL estimates than those derived

directly (Ades et al., 2013).

Lu et al. (2013) were the first to put forward the CFM for the purposes of mapping in CEA. Their study developed a mapping algorithm to predict generic HRQoL scores from disease-specific measures (DSM) based upon the assumptions (a) that both measures share some underlying construct, known as the common factor model, and (b) that the DSM is a pure measure of the variability of interest with the exception of measurement error. These assumptions were set out in formulaic terms by splitting the measures into separate components, as represented by the equation below:

$$DSM_i = \alpha_1 + \beta_1 CF_i + \varepsilon_{1i} \tag{4.1}$$

$$Gen_i = \alpha_2 + \beta_2 CF_i + \beta_3 Other_i + \varepsilon_{2i} \tag{4.2}$$

Based upon the equations above, the authors postulated that a coherent mapping between the two measures should be realized in terms of the common factor, specifically the ratio of the coefficients upon this factor, β_1/β_2. The authors demonstrated that this term can be derived by re-parameterizing the common factor model in terms of reliability and responsiveness. In practice, estimation of this term requires external information about the reliability of the DSM as well as the covariance structure associated with the DSM and generic measure (or patient-level data from which this can be estimated).[2]

The authors proposed three key properties for a mapping system to be considered coherent: invertability, transitivity and scale invariance. A case study was conducted using trial data with three outcome measures – EQ-5D, SF-12 and the Beck Depression Inventory – that demonstrated the adherence of the CFM to each of these properties. A range of other regression techniques were conducted, including OLS, but none were shown to consistently possess all of these properties. The geometric mean regression was the only method, other than the CFM, shown to be capable of possessing all three properties, although only in specific circumstances.

Lu and colleagues have recently claimed that the CFM framework would be well placed to synthesise evidence from multiple heterogeneous HRQoL outcome measures (Lu et al., 2014). The CFM approach assumes that these different measures capture alternative realisations of the same underlying construct(s). In doing so, this method avoids disregarding

[2]Note that the authors show another method that does not require external information about reliability but instead uses information for two generic measures and one DSM. However, the methods associated with this approach are not considered in this chapter.

potentially relevant evidence, simply because it has not been produced using the benchmark PRO instrument chosen by the decision maker (e.g. in the case of NICE, the EQ-5D) or a commonly reported outcome in the disease area.

The CFM developed by Lu and colleagues (2013), whilst promising, is not without limitations; thus far, research has been conceptual and only a handful of case studies have been conducted (Lu et al., 2013). First, the CFM applications to date have been concerned with the synthesis of HRQoL evidence relating to treatment effects. In practice, most cost-effectiveness models require HRQoL parameters associated with defined health states and events, rather than just treatment effects and, as a consequence, require baseline HRQoL values associated with not experiencing the event or health condition (Ara and Wailoo, 2012). Given that the CFM in its current form is only capable of synthesising evidence capturing relative effects, further research is required to determine if and how one could go about synthesising evidence associated with absolute values.

Another limitation of the methods currently available is that they assume that the common factor representing treatment effects can be captured by a unitary construct. Lu et al. (2013) noted that certain disease areas might not be characterised by a single construct. The paper does not suggest how one might go about extending the CFM to deal with multiple health domains. This would seem particularly important given that many DSMs are composed of several subscales capturing different dimensions of health.

Finally, the CFM developed by Lu and colleagues (2013) employed a simple linear estimation procedure that does not take into account some of the unique features of HRQoL data. This approach does not guarantee that predictions will lie within a feasible range, nor does it adequately account for the unique distribution of health index values. As such, the CFM would benefit from further research to generate methods better suited to the handling of these features.

The next section considers many of the issues raised in the context of the structural equation model framework given that the CFM developed by Lu and colleagues (2013) can be formulated as a structural equation model (SEM). SEMs have enjoyed extensive application in other areas of research, particularly psychometrics research, which has seen significant methodological advances to contend with the complex relationships often hypothesized in terms of latent variables. The descriptive systems of preference-based measures of health share a likeness with the outcome measures used in psychometrics in that they are both attempting to measure human perceptions of subjective concepts. As such, the capability of SEMs opens up a range of opportunities for effectively contending with

the challenges of analysing preference-based measures to derive HSVs for the estimation of QALYs.

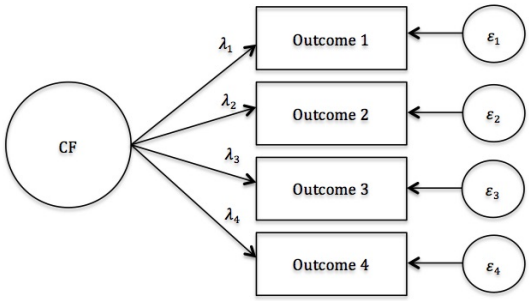
4.3 Structural Equation Models

The term structural equation model refers to a broad family of related statistical models rather than a single specific technique (Kline, 2006). In fact, many well-known statistical techniques, such as variance and regression analysis, can be undertaken within the SEM framework. One of the distinguishing features of SEM's is that they allow an explicit representation of observable and latent variables. These relationships are commonly expressed in algebraic form or in a graphical format (Moon-Ho et al., 2012). An abstract example is provided in Figure 3.1 to explain some of the concepts further.

SEM diagrams typically employ circles to represent latent variables and rectangles to represent observed variables (Moon-Ho et al., 2012). In Figure 4.1, there is one variable representing a hypothetical construct of interest, labelled "CF" representing a "common factor", and a further four latent variables representing error terms. The relationships between variables are represented using arrows, with each starting at the independent variable and ending at the dependent variable.

The model in Figure 4.1 assumes that the observed variables, indicators 1 to 4, are related due to the action of the unobserved construct, CF. Any observed correlations between the indicators are assumed to occur as a result of this construct. The relationship between the construct, CF, and the indicators is represented by the coefficients on each of the arrows (λ_1,λ_2,λ_3 and λ_4), known as factor loadings. It is also assumed that the observed variables are partly determined by other factors and this is captured in the model with the error terms (ε_1,ε_2,ε_3 and ε_4). As such, the model has decomposed the variance of each of the observed variables into two parts: a shared part (CF) and a unique part (ε_i).

Figure 4.1: Graphical Example of a Structural Equation Model



The model in Figure 4.1 can be tested provided that there is data available containing each of the variables of interest, specifically a variance-covariance matrix. The analysis undertaken, known as a confirmatory factor analysis (CFA), aims to identify a set of parameters, or more specifically the factor loadings, that minimize the difference between the observed and estimated data (Lei and Wu, 2012). This method for deriving parameter estimates is very similar to that with an OLS regression analysis.

An important assumption made in SEMs, such as the one in Figure 4.1, is that of model identification. There are two key conditions that must be satisfied for a model to be identified. First, the amount of unknown information in the model should be less than or equal to the amount of observed information available. The term unknown information relates to the number of free parameters in the model, which may include factor loadings, factor variances and error terms[3]. The amount of observed information refers to the number of elements in the variance-covariance matrix. For example, the model in Figure 4.1 has ten pieces of information – four observed variables with four variances and six covariances[4].

The second condition for model identification is that every factor must have a scale. Latent variables are not directly measured and require a scale so that their variances are interpretable. This can be achieved in two ways. The first involves fixing the variance of a factor to equal a constant (e.g. equal to 1). Alternatively, one might fix a factor loading for one of the indicators to equal 1, effectively scaling the variance of the latent variable to equal that of the indicator variable.

The two aforementioned conditions represent the minimum theoretical requirements for a model to be correctly identified[5]. Where these conditions are not met, the model is said to be under-identified and the analyst should consider an alternative specification. The first condition implies that a standard CFA requires at least three factors in order to be correctly identified. Interestingly, the CFM developed by Lu and colleagues (2013) would have been under-identified were it not for the addition of external information relating to the reliability of the DSM. This model, represented in equations (1) and (2), is set out graphically in Figure 4.2. Note that the latent factor, $Other_i$, has been incorporated within the error term, ε_i. The following sections will consider the extension of this model

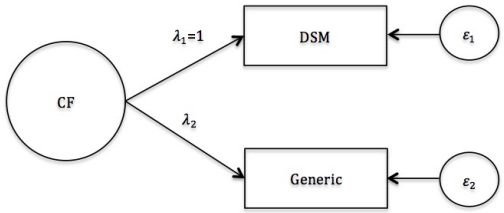
[3]It is worth noting that not all model parameters will necessarily be free; sometimes parameters are constrained to a specific value, also known as fixed parameters.

[4]In general, for k observed variables, there are $k(k+1)/2$ pieces of information (Kenny and Milan, 2011).

[5]Note: models that are theoretically identified are still prone to empirical under-indentification due to data related problems (Kline, 2006).

to handle DSMs with multiple constructs and dealing with item-level responses rather than the index values from generic instruments.

Figure 4.2: Graphical Representation of the Common Factor Model



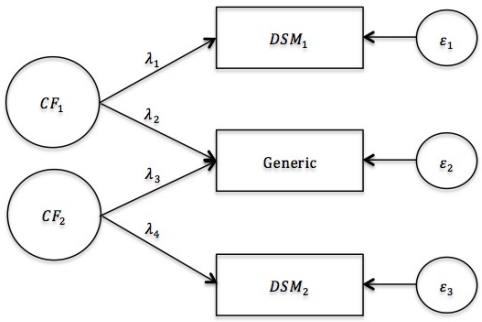
4.3.1 *Dealing with Multi-Construct Outcome Measures*

The model set out in Figure 4.2 assumes that the observed DSM is represented by a single score. In practice, DSMs are often composed of many multi-choice questions, known as items, and these items may be organised into groups that measure separate constructs. For example the cancer-specific questionnaire, the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 (EORTC QLQ C-30), has 30 items that cover several health domains including five functioning scales, a global QoL scale and nine symptom scales (Fayers et al., 1999). As such, analysts may wish to characterise DSMs within SEMs at the item-level or sub-score level and, in some cases, this might be necessary where there is no overall summary score.

First, let us consider the scenario of a DSM with several sub-scores, each measured on a continuous scale, and the summary scores for a generic measure, also measured on a continuous scale. In practice, the sub-scores of DSMs are often designed to capture different aspects of the condition of interest. For example, the EORTC QLQ-C30 covers five functioning scales, including physical, role, social, emotional and cognitive functioning. For this reason, it will often be more appropriate to assume that the separate sub-scores are explained by different factors. Figure 4.3 illustrates a model with separate factors corresponding with the different sub-scores.

This model assumes that the sub-scores, DSM_1 and DSM_2, are explained by separate underlying constructs. As with the CFM set out in Figure 4.2, the sub-scores here can be assumed to be "pure" measures of these constructs with the inclusion of external information about the reliability of the sub-scores. However, the model in Figure 4.3 relies upon a strong assumption that the error terms (ε_1,ε_2,ε_3) are uncorrelated. In practice, this might be too strong an assumption given that there may be remaining shared variance between the sub-scores of the DSM, beyond that explained by the model. Where there

Figure 4.3: Graphical Example of a Structural Equation Model with a Multi-Construct Disease-Specific Measurement

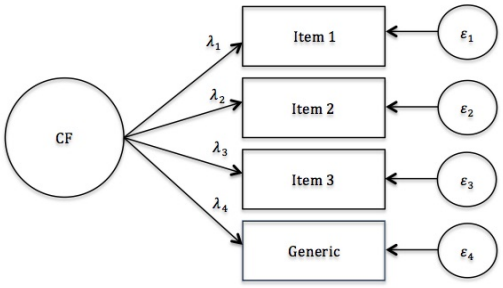


are concerns that this might be the case, one solution would be to introduce an additional factor into the model that accounts for the shared variance between the sub-scores of the DSM.

4.3.2 *Dealing with Item-Level Responses*

SEMs are capable of analysing item-level responses to both generic measures and DSMs thanks to methodological advances in the way that categorical data is handled (Edwards et al., 2012). However, this raises a number of challenges in relation to the application of the CFM, all of which relate to the question of how many latent factors should be specified. First, suppose that we have item-level responses for the DSM and index values for the generic measure; should we specify a different latent factor for each of the items, in a similar fashion to the model in Figure 4.3? It will often be the case that different items on the same measure will be specified to capture different aspects of the same construct. As such, one would expect to specify a model such as the one in Figure 4.4.

Figure 4.4: Graphical Example of a Structural Equation Model with Multiple Categorical Items Loading on the Same Factor



This model represents an important departure from the original CFM because it no longer assumes that we have one pure measure of the construct of interest. Instead, it assumes that there are several items that capture slightly different aspects of the construct and

that we are interested in the variability that is shared between these items[6]. This raises the question of whether we really need external information about the reliability of the items. Unlike the models in Figures 4.2 and 4.3, this model is correctly identified thanks to the additional items on the latent factor.

The benefits of having access to IPD for the purposes of synthesising HRQoL data with structural equation modeling techniques – in particular, the ability to exploit item-level information – have yet to demonstrated in the context of HEE. Only one study has been found to employ an SEM-type approach, specifically a multidimensional item response theory (IRT) method, to analyse preference-based measures of HRQoL at the item level (Gibbons et al., 2014). This study showed that the multidimensional IRT model performed well in terms of its ability to predict EQ-5D scores compared to traditional regression methods. Despite showing early promise, the predictive performance of this approach has yet to be externally validated. These issues are explored further in the case study presented in chapter 6.

4.4 Priorities for Future Research

Although promising, early research involving SEM techniques for the synthesis of HRQoL evidence has been conceptual and lacking in generalisability. Critically, most of these studies do not even involve preference-based measures of HRQoL. Further research exploring the application of SEM methods in a greater range of scenarios is required to inform policy makers and the research community alike about the likely implications of adopting such methods. The primary objective for the remainder of the thesis is to investigate the plausibility of the SEM framework serving as a generalised framework for the handling of HRQoL evidence. SEM methods are tested across each of the scenarios illustrated in the taxonomy in Chapter 2; in each case, a comprehensive synthesis of heterogeneous HRQoL outcomes using the SEM approach is compared against a restrictive synthesis involving a reference case measurement. Whilst the methods for identifying and selecting HRQoL evidence for CEA are not the primary concern of this thesis, it is important to distinguish between the alternative inclusion criteria for study selection that may be specified. For convenience, the term "reference case" approach is used to refer to a systematic review of HRQoL evidence with the inclusion criteria restricted to select only those studies that collecting the reference case measurement (e.g. the EQ-5D for NICE).

[6]This model is said to be uni-dimensional because it assumes that there is a single factor loading on all of the item responses.

The term "all-inclusive" approach is used to refer to a systematic review that incorporates a range of outcomes beyond the reference case measurement. The additional measures could potentially include any of the following: alternative generic preference-based measures, condition-specific preference-based measures, generic PROs (non-preference-based) or disease-specific PROs (non-preference-based). Assuming that the "reference case" and "all-inclusive" approaches are identical in all other aspects of the inclusion criteria (i.e. the population characteristics of interest), the studies selected in the former approach should always either be the same as or a subset of those studies selected in the latter approach. In addition, the implications of these alternative approaches will be explored in the context of a cost-effectiveness model. The thesis also looks to investigate the advantage for researchers having access to patient-level data, rather than aggregate data.

Chapter 5

Case Study I: Synthesis of Aggregate Data

5.1 Introduction

Chapter 3 identified the SEM framework as a potential avenue for the synthesis of heterogeneous HRQoL in CEA. It was decided that further empirical research was required in order to evaluate the feasibility of this methodology in practice. This chapter focuses upon scenario D1 from the taxonomy in Table 2.1 where researchers, seeking to estimate HRQoL parameters for CEA, are faced with an evidence base involving multiple sources of evidence in the AD format. One study has been identified to date as having used SEM techniques in this context although it did not involve preference-based HRQoL values (Lu et al., 2014). Furthermore, this study did not consider the implications of using the SEM approach for CEA in comparison to current methods. As such, this chapter seeks to (a) identify the methodological challenges unique to scenario D1, and (b) to compare HRQoL parameter estimates, derived using an "all-inclusive" approach to evidence synthesis with SEM methods, against those obtained using a "reference case" approach to evidence synthesis with standard meta-analytic methods.

5.2 Background

This section provides context in relation to the application of SEM techniques using evidence in the AD format.

5.2.1 *Evidence Synthesis via Standard Meta-Analytic Methods*

In this chapter, the meta-analytic approach used to synthesise the "reference case" evidence includes both covariates, for explained heterogeneity, and random effects, for residual heterogeneity. This method is implemented using a Bayesian approach and the model specification is illustrated in Equations 5.1 – 5.5. The Bayesian approach to evidence synthesis assumed that model parameters are random quantities and a likelihood function is defined to reflect the plausibility of the data given the model parameter values (Sutton and Abrams, 2001).

Equations 5.1 – 5.2 represent the likelihood function. The term $HRQoL_i$ represents the observed mean HRQoL value in study i with the associated standard error represented by the term SE_i. A covariate, $Event_i$, has been included in order to explore differences between patients' reported HRQoL values depending upon whether or not they have experienced a health event. In addition, study specific random effects have been included in order to capture unexplained heterogeneity between values coming from different studies. These random effects are captured by the term θ_z and are assumed to be samples from a normal distribution with mean μ and and variance τ_z^2

$$HRQoL_i \sim Normal(\theta[study_i] + \beta \cdot Event_i, SE_i) \tag{5.1}$$

$$\theta_z \sim Normal(\mu, \tau_z^2) \tag{5.2}$$

A defining feature of the Bayesian approach to evidence synthesis is the specification of prior distributions for the unknown model parameters. Prior distributions can be based upon external evidence (so-called "informed" priors) or subjective a priori beliefs (Lung et al., 2011). Equations 5.3 – 5.5 illustrate the prior distributions for all of the unknown parameters in this model.

$$\mu \sim Uniform(0, 1) \tag{5.3}$$

$$\beta \sim Normal(0, 0.1) \tag{5.4}$$

$$\tau_z^2 \sim Uniform(0.001, 0.2) \tag{5.5}$$

In this chapter, it is assumed that there is no external evidence available to inform the prior distributions. Instead, "reference" priors have been specified which are assumed to fulfill

some basic desirable properties associated with the parameters under evaluation (Lunn et al., 2012). The constant term has been assigned a uniform distribution, ranging between 0 and 1, based on the fact that this typically reflects the plausible range of HRQoL values.[1] The coefficient β has been assigned a normal distribution with a mean at zero to reflect the fact that the impact of the health event is unknown. Finally, the variance component τ_z^2 requires a prior distribution that guarantees non-negative values and a decision was made to use the uniform distribution.

5.2.2 *Evidence Synthesis via SEM Methods*

The linear structural equation modeling approach, also known as the LISREL method, readily lends itself to the analysis of data in the AD format. This is because the parameter estimation procedure typically involves minimizing some form of discrepancy between a sample variance-covariance matrix and a model-implied covariance matrix (Lei and Wu, 2012). In addition to the sample variance-covariance matrix, researchers might consider the inclusion of sample means for all of the outcome measures involved as intercepts in the model. If this doesn't happen, as was the case in the study by Lu and colleagues (Lu et al., 2014), then there are no intercepts and the means scores for each of the outcomes are assumed to equal zero. This approach was reasonable in the study by Lu and colleagues given that the main objective was to synthesise evidence relating to treatment effects for six measures of HRQoL. However, supposing that one wants to synthesise evidence pertaining to *baseline* or *absolute* HRQoL values, then inclusion of the sample means is necessary.

Aside from the work by Lu and colleagues (Lu et al., 2014), the intuition behind the intersection of SEM and meta-analysis has only come about very recently and the early work in this area has put forward two distinctive applications: (i) the implementation of standard multivariate meta-analyses (MVMA) in a SEM framework (Cheung, 2013, 2014b); (ii) a meta-analytic SEM approach that involves pooling correlation (or covariance) matrices (Cheung, 2014a). The former deals only with the structural part of the SEM framework, rather than the measurement part. The motivation behind this approach is the promise of 'borrowing of strength' across mean outcomes and the potential for subsequent reductions in uncertainty surrounding parameter estimates. MVMA has been put forward as a means of incorporating evidence from several correlated outcome measures for the

[1] Note that this range will vary depending upon the instrument under evaluation. For example, a range between -0.594 and 1 would be appropriate for the EQ-5D with population weights from the study by Dolan (1997).

purposes of populating input parameters in cost-effectiveness models (Bujkiewicz et al., 2013). Although very similar to the model by Lu and colleagues in terms of the evidence requirements, the MVMA approach does not produce coherent mappings between the various outcomes under evaluation.

The meta-analytic SEM approach involves two steps: first, a synthesis of multiple correlation or covariance matrices is performed and then an SEM is fitted to the output (Cheung, 2014a). This approach is potentially useful for researchers interested in the synthesis of multiple sample covariance matrices involving HRQoL outcome measures. However, applications involving the meta-analytic SEM method are not investigated in this chapter for two reasons. First, the likelihood of researchers managing to obtain a single sample covariance matrix, let alone multiple matrices, from the published literature is contested in this chapter. Second, the meta-analytic SEM method is currently unable to accommodate the synthesis of mean scores in addition to covariance matrices.

The chapter considers two separate approaches involving SEM techniques for the synthesis of heterogeneous HRQoL evidence: (i) a two-step method with mapping undertaken separately from evidence synthesis, and (ii) an integrated approach that undertakes these steps simultaneously. These methods are implemented in conjunction with the "all-inclusive" approach to evidence identification and selection, i.e. there will be HRQoL values from multiple instruments. A prerequisite for the implementation of either method is the availability of covariance matrix data, capturing all of the relevant outcome measures. For the purposes of this case study, we will consider a scenario where researchers obtain a sample covariance matrix for one of the studies identified in the review.[2]

Equations 5.6 – 5.16 illustrate the first step of the two-step method for evidence synthesis involving SEM techniques. A confirmatory factor analysis (CFA) modeling approach is shown in equations 5.6 – 5.9; this model specifies the relationships between observed values for different HRQoL instruments according to an underlying latent factor, LF. The CFA falls within the broad family of SEM methods and estimation of this model provides the parameter estimates (λ_2 and λ_3) necessary for mapping between the different instruments. Equations 5.10 – 5.12 show the model-implied covariance matrices, which are a product of the factor loadings and factor variance score (Brown and Moore, 2012). The parameter estimation procedure typically involves minimizing some form of discrepancy between the sample covariance matrix and the model-implied covariance matrix.

[2]That is, one study provides mean HRQoL values associated with one of the defined health states for all of the relevant instruments, as well as providing a sample covariance matrix.

$$EQ5D_i = \mu_{EQ5D} + LF_i \tag{5.6}$$

$$SF6D_i = \mu_{SF6D} + \lambda_2 \cdot LF_i \tag{5.7}$$

$$HUI3_i = \mu_{HUI3} + \lambda_3 \cdot LF_i \tag{5.8}$$

$$LF_i \sim Normal(0, 1) \tag{5.9}$$

$$Cov(EQ5D, SF6D) = 1 \cdot \lambda_2 \cdot 1 \tag{5.10}$$

$$Cov(EQ5D, HUI3) = 1 \cdot \lambda_3 \cdot 1 \tag{5.11}$$

$$Cov(SF6D, HUI3) = \lambda_2 \cdot \lambda_3 \cdot 1 \tag{5.12}$$

Once the factor loadings (λ_2 and λ_3) have been obtained, HRQoL values can be predicted on the scale of the reference case instrument (in this case, the EQ-5D) for all of the studies available. Equations 5.13 – 5.16 show how mean SF-6D and HUI-3 values, along with the associated standard errors, can be mapped onto the EQ-5D scale. Once all of the evidence has been mapped onto the EQ5D scale (if it wasn't already), the evidence synthesis step can be undertaken using the methods from equation 5.1 – 5.5.

$$Mapped\ EQ5D_i = \mu_{EQ5D} + \frac{SF6D_i - \mu_{SF6D}}{\lambda_2} \tag{5.13}$$

$$Mapped\ EQ5D_i = \mu_{EQ5D} + \frac{HUI3_i - \mu_{HUI3}}{\lambda_3} \tag{5.14}$$

$$SE(Mapped\ EQ5D_i) = \frac{SE(SF6D_i)}{\lambda_2} \tag{5.15}$$

$$SE(Mapped\ EQ5D_i) = \frac{SE(HUI3_i)}{\lambda_3} \tag{5.16}$$

The so-called *integrated* method aims to use a single model to explain all aspects of the available data. The motivation behind this method is that it should, theoretically, provide a more logical approach to evidence synthesis compared to the two-step method given that

all of the observed evidence is interpreted in a consistent manner. However, the process of developing a more logical approach to evidence synthesis brings greater complexity in terms of the model specification required to implement this technique.

Equations 5.17 - 5.40 illustrate the mathematical notation for the *integrated* method. This approach relies heavily upon an assumption of fixed factor loadings across studies given that there is only a single study containing the sample covariance information (Equations 5.17 and 5.18). The factor loadings (λ_2, λ_3) are estimated using the steps outlined in Equations 5.6 – 5.12, along with the associated standard errors (SE(λ_2), SE(λ_3)). Estimation of the standard errors is an important step as it allows the uncertainty surrounding the factor loadings to be captured.

$$\lambda_{2t} \sim Normal(\lambda_2, SE(\lambda_2)) \tag{5.17}$$

$$\lambda_{3t} \sim Normal(\lambda_3, SE(\lambda_3)) \tag{5.18}$$

Although most of the studies are missing information for some of the outcome measures, this problem is circumvented by the fact that all of the outcomes are related to each other via the latent factor. As such, if a study is missing sample statistics for one outcome (e.g. EQ-5D) then information can be borrowed from another measure (e.g. SF-6D) through the relationships set out in Equations 5.25 – 5.30. Note that this borrowing of information applies to standard errors (5.28 – 5.30), in addition to mean values (5.25 – 5.27).

$$EQ5D_i \sim Normal(\theta_{1i}, SE(EQ5D_i)) \tag{5.19}$$

$$SF6D_i \sim Normal(\theta_{2i}, SE(SF6D_i)) \tag{5.20}$$

$$HUI3_i \sim Normal(\theta_{3i}, SE(HUI3_i)) \tag{5.21}$$

$$SE(EQ5D_i) \sim Normal(\phi_{1i}, \psi_1) \tag{5.22}$$

$$SE(SF6D_i) \sim Normal(\phi_{2i}, \psi_2) \tag{5.23}$$

$$SE(HUI3_i) \sim Normal(\phi_{3i}, \psi_3) \tag{5.24}$$

The scale of latent factor is fixed to that of the EQ-5D, which is illustrated by the fact that there is no factor loading in Equation 5.25 (i.e. the factor loading is equal to one). Random effects are included to capture any unexplained heterogeneity between latent factor scores coming from different studies (Equation 5.31). The term υ captures the mean population HRQoL for patients with no previous experience of an MI. This parameter is scaled in a way such that deviations from zero on the latent factor scale reflect deviations from μ_{EQ5D} on the EQ-5D scale. As such, we should add μ_{EQ5D} to our estimate of υ in order to interpret it on the EQ-5D scale.

$$\theta_{1i} = \mu_{EQ5D} + LF_i \tag{5.25}$$

$$\theta_{2i} = \mu_{SF6D} + \lambda_{2t} \cdot LF_i \tag{5.26}$$

$$\theta_{3i} = \mu_{HUI3} + \lambda_{3t} \cdot LF_i \tag{5.27}$$

$$\phi_{1i} = SE(LF_i) \tag{5.28}$$

$$\phi_{2i} = \lambda_{2t} \cdot SE(LF_i) \tag{5.29}$$

$$\phi_{3i} = \lambda_{3t} \cdot SE(LF_i) \tag{5.30}$$

$$LF_i \sim Normal(\delta[study_i] + \beta \cdot Event_i, \varepsilon) \tag{5.31}$$

$$\delta_z \sim Normal(\upsilon, \tau_z^2) \tag{5.32}$$

$$\upsilon \sim Normal(0, 0.1) \tag{5.33}$$

$$\beta \sim Normal(0, 0.1) \tag{5.34}$$

$$\varepsilon \sim Uniform(0.001, 0.2) \tag{5.35}$$

$$SE(LF_i) \sim Uniform(0.001, 0.2) \tag{5.36}$$

$$\psi_1 \sim Uniform(0.001, 0.2) \tag{5.37}$$

$$\psi_2 \sim Uniform(0.001, 0.2) \tag{5.38}$$

$$\psi_3 \sim Uniform(0.001, 0.2) \tag{5.39}$$

$$\tau_z^2 \sim Uniform(0.001, 0.2) \tag{5.40}$$

5.3 Methods

An empirical case study was conducted in order to compare HRQoL parameter estimates obtained using standard meta-analytic methods against those obtained using each of the proposed SEM methods. The standard meta-analytic approach draws solely upon HRQoL evidence for a reference case measurement, the EQ-5D. In contract, the SEM methods exploit HRQoL evidence from a broader range of measurements. The case study also explores the implications of using the different methods in terms of the impact upon cost-effectiveness results. The case study selected investigates the cost-effectiveness of an early surgical intervention compared to medical management for the treatment of acute coronary syndrome in a patient subgroup with diabetes.

5.3.1 *HRQoL Evidence*

Health state valuations were extracted from an existing meta-analytic study exploring the effects of diabetes and related complications upon HRQoL. Lung and colleagues performed a search and review of published databases to identify studies containing relevant values, the details of which can be found in the original study (Lung et al., 2011). The criteria for study selection was not confined to any one HRQoL instrument and no efforts were made to adjust values from alternative instruments onto the same scale. The studies were grouped according to the type of complications experienced by patients that included the following: no complications, stroke, myocardial infarction, end-stage renal disease, blindness, amputation and ulcers. In total, Lung and colleagues identified forty-six health state values, along with associated standard errors. Not all of these values were used for the purposes of this chapter given that they varied not only in terms of the instruments used but also the population index values for a given instrument. It is well established that different population index values for a given same instrument can result in different

valuations and, hence, undermine the comparability (Kharroubi et al., 2010, 2014). For this reason, specific preference weights were defined for each instrument and health state values were extracted if they were met any of the following criteria:

- EQ-5D values derived using population weights from the study by Dolan (1997).
- SF-6D values derived using population weights from the study by Brazier et al. (2002).
- HUI-3 values derived using population weights from the study by Feeny et al. (2002).

For those values meeting the inclusion criteria, the following information is extracted: mean value; standard error; publication author; publication year; preference-based instrument; information regarding any diabetes-related complications; the number of patients in the sample.

A variance-covariance matrix relating to the three instruments was obtained from the National Health Measurement Survey (NHMS). The NHMS was conducted in the United States and collected a variety of measures of HRQoL, including the EQ-5D, HUI-3 and the SF-36, in a sample of the general population. Further details about this dataset can be found elsewhere (Fryback et al., 2007). The variance-covariance matrix was derived in a subset of patients recorded as having diabetes. In addition, the mean values for each of the measures are extracted so that they can be included in the synthesis. Whilst the individual-patient data might have been available to estimate this matrix, we assume that the statistics have been identified in the aggregate format.

5.3.2 *Health Economic Model*

The health economic model has been derived from an existing study, the details of which can be found elsewhere (Henriksson et al., 2008). The expected costs and QALYs for competing interventions are estimated indirectly as follows: (i) modeling the impact of treatments upon the probability of a given patient either experiencing a myocardial infarction or dying; (ii) linking these intermediate endpoints to the 'final' endpoints, namely length of life, HRQoL and costs. This chapter is solely concerned with the estimation of HRQoL inputs for this model, referred to as the RITA-3 model, so the remaining parameters have been left unchanged from those in the original study.

The HRQoL input parameters employed in this case study have been simplified from the original study. Specifically, the model not longer accounts for the impact of a myocardial infarction in the last twelve months. Instead, the model in this study only contains three

health states that differentiate between HRQoL – one for patients with diabetes who have yet to experience a myocardial infarction; another for patients with diabetes who have experienced a myocardial infarction; and finally, death (assumed to equal zero).

5.3.3 *Estimation of Parameter Inputs*

The following statistical procedures for evidence synthesis will be compared:

- Hierarchical Meta-Regression (Model 5.1): health state values derived using the EQ-5D are analysed using the model set out in Equations 5.1 – 5.5. Note that the Equation 5.3 is modified to account for the range of EQ-5D values as follows:

$$\mu \sim Uniform(-0.594, 1) \tag{5.41}$$

- CFA-Approach to Evidence Synthesis (Model 5.2): the confirmatory factor analysis, set out in Equations 5.6 – 5.12, is fitted using the sample statistics from the NHMS. The parameter estimates from this model (λ_2 and λ_3) and the mean values (μ_{EQ5D}, μ_{SF6D} and μ_{HUI3}) are then used to map health values for the HUI-3 and SF-6D onto the EQ-5D scale. Health state values derived using the EQ-5D, including those that have been mapped onto the EQ-5D scale, are then combined using the hierarchical meta-regression method.

- SEM-Approach to Evidence Synthesis (Model 5.3): the confirmatory factor analysis set out in Equations 5.6 – 5.12 is fitted using the sample statistics from the NHMS. Using the parameter estimates for the factor loadings, the health state values derived using the EQ-5D, SF-6D and HUI-3 are synthesised using the model set out in Equations 5.17 – 5.40.

For any statistical analysis, it is important to assess how well the estimated model fits the observed data. As such, the posterior mean of the residual deviance has been estimated, given that this method is recommended for the assessment of model fit in evidence synthesis applications involving Bayesian methods (Sutton et al., 2012). In addition, cross-validation of the results has been undertaken in order to assess the degree to which the data from different evidence sources are consistent. If there are prominent differences between the results from different studies, then this may reduce confidence in the conclusions draw from a synthesis of these results. Cross-validation involves omitting an individual study from the analysis and then comparing the findings from the omitted study against the predictive distribution from the analysis (Sutton et al., 2012). If the omitted data point

is shown to lie within the random effects distribution obtained from the analysis, then it is reasonable to assume that it is consistent with the remaining data.

5.3.4 *Statistical Software*

The meta-analytic models are estimated using the JAGS software, which is run through the R program using the *R2jags* package (Su and Yajima, 2012). Each model is run with three Markov chains over total of 10,000 iterations, the first 1,000 of which are discarded as the burn-in period. For each method, the posterior distribution is fed directly into the health economic model for the probabilistic sensitivity analysis. The factor loadings are estimated using the *Lavaan* package in R (Rosseel, 2012). The relevant code for each of the different methods can be found in Appendix B.

5.4 Results

Of the original forty-six values with health state description in the original study by Lung and colleagues, five values were identified as meeting the inclusion criteria for the "reference case" approach. An additional four values were identified as meeting the inclusion criteria for the "all-inclusive" approach. These values are shown in Table 5.1. Note that additional values were extracted from the NHMS study.

5.4.1 *Evidence Synthesis*

Three alternative methods for synthesising evidence were implemented to estimate HRQoL parameter inputs. Model 5.1 combined evidence from six separate studies, all of which collected EQ-5D values (the "reference case" approach). Models 5.2 and 5.3 used evidence from an additional four studies that collected alternative PBMs (the "all inclusive" approach). The forest plot in Figure 5.1 provides a graphical representation of the summary statistics from all of the studies selected for synthesis, as well as summary measures for each of the synthesis methods evaluated. All of the estimates are on the EQ-5D scale, including those from Glasziou *et al.* (2007), Maddigan *et al.* (2005) and Wee *et al.* (2005).

The parameter estimates obtained using the "all-inclusive" evidence (Models 5.2 and 5.3) are consistently higher than those obtained using the "reference case" evidence (Model 5.1). In addition, the estimated impact of a MI upon HRQoL is reduced when the "all-inclusive" evidence is used instead of the "reference case" evidence. This is likely to be explained by the additional evidence from the study by Maddigan *et al.* (2005). The

Table 5.1: HRQoL Values extracted from the study by Lung and colleagues (2011)

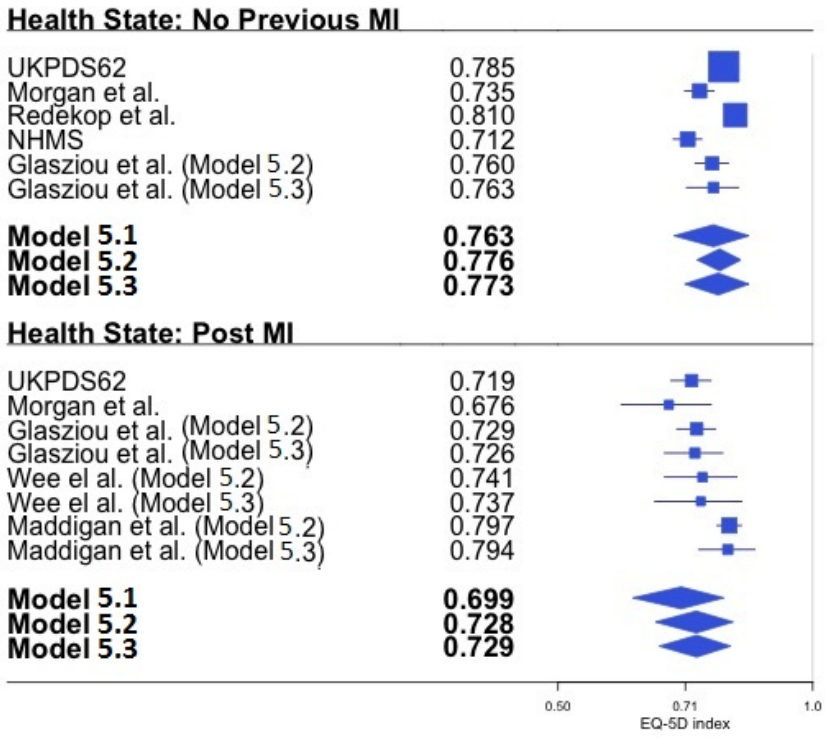
Study	Sample Size	Patient Subgroup	Instrument	Mean Estimate	Standard Error
UKPDS62 (2002)	2,636	Diabetes, no complications	EQ-5D	0.785	0.005
Morgan *et al.* (2006)	360	Diabetes, no complications	EQ-5D	0.735	0.015
Redekop *et al.* (2002)	1,136	Diabetes, no complications	EQ-5D	0.810	0.007
Glasziou *et al.* (2007)	210	Diabetes, no complications	SF-6D	0.780	0.009
UKPDS62 (2002)	200	Diabetes, Post-MI	EQ-5D	0.719	0.020
Morgan *et al.* (2006)	172	Diabetes, Post-MI	EQ-5D	0.676	0.042
Glasziou *et al.* (2007)	183	Diabetes, Post-MI	SF-6D	0.764	0.010
Wee *et al.* (2005)	46	Diabetes, Post-MI	SF-6D	0.770	0.019
Maddigan *et al.* (2005)	172	Diabetes, Post-MI	HUI-3	0.770	0.015
NHMS	479	Diabetes, no complications	EQ-5D SF-6D HUI-3	0.712 0.755 0.675	0.014 0.007 0.015

parameter estimates obtained using the "all-inclusive" evidence also exhibit reduced uncertainty when compared to those obtained using the "reference case" evidence. This finding is particularly encouraging given that one of the main reasons for synthesising evidence is the possibility of attaining increased statistical power for the estimation of parameters (Higgins et al., 2009). Models 5.2 and 5.3 produce similar results with one exception being the uncertainty surrounding the parameter representing the health state 'No Previous MI'. One possible explanation for this finding could be the fact that Model 5.3 has a larger number of unknown parameters.

The residual deviance for Model 5.1 has a posterior mean equal to 5.09 which is close to the number unconstrained data points (N=5) used in the estimation of Model 5.1. Similarly, the posterior mean of the residual deviance for Model 5.2 ($\bar{D}_{RES}$=9.12) is close to the number unconstrained data points (N=10) used in the estimation of Model 5.2.

These results do not indicate a lack of fit for Models 5.1 and 5.2. Residual deviance was not estimated for Model 5.3 given that there are three dependent variables with lots of missing data. For all three models, cross-validation of the results found the data from different evidence sources to be consistent.

Figure 5.1: Forest Plot



5.4.2 *Cost-Effectiveness Results*

Tables 5.2 – 5.4 show the implications of using the different methods in terms of the impact upon the expected cost-effectiveness results. Ideally, HEE for decision-making purposes should investigate (a) whether a health intervention is expected to be cost-effective based on existing evidence, and (b) whether or not additional evidence is required to support its use (Griffin et al., 2011). In order to inform the first objective, researchers can present probabilistic results from their cost-effectiveness model in a cost-effectiveness acceptability curve (CEAC) (Briggs et al., 2006). The CEAC shows the probability of an intervention being cost-effective for a given threshold value when compared to other strategies. Figure 5.2 shows a series of CEACs associated with the various parameter estimation methods assessed in this chapter. Comparison of the curves shows that the choice of methodology has a negligible impact upon decision uncertainty for this particular case study. This information is also presented in terms of error probabilities in Table 5.5,

assuming threshold values of £20,000 per QALY and £30,000 per QALY respectively. These results present the probability of making an incorrect decision, for a given patient, on the basis of the expected cost-effectiveness results.

Table 5.2: Cost-Effectiveness Results Using Model 5.1 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£36,940	£32,188	£4,752
QALYs (discounted)	12.83	12.70	0.12
Cost-per-QALY	-	-	£38,306

Table 5.3: Cost-Effectiveness Results Using Model 5.2 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£36,940	£32,188	£4,752
QALYs (discounted)	13.07	12.95	0.12
Cost-per-QALY	-	-	£38,179

Table 5.4: Cost-Effectiveness Results Using Model 5.3 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£36,940	£32,188	£4,752
QALYs (discounted)	13.03	12.91	0.12
Cost-per-QALY	-	-	£38,444

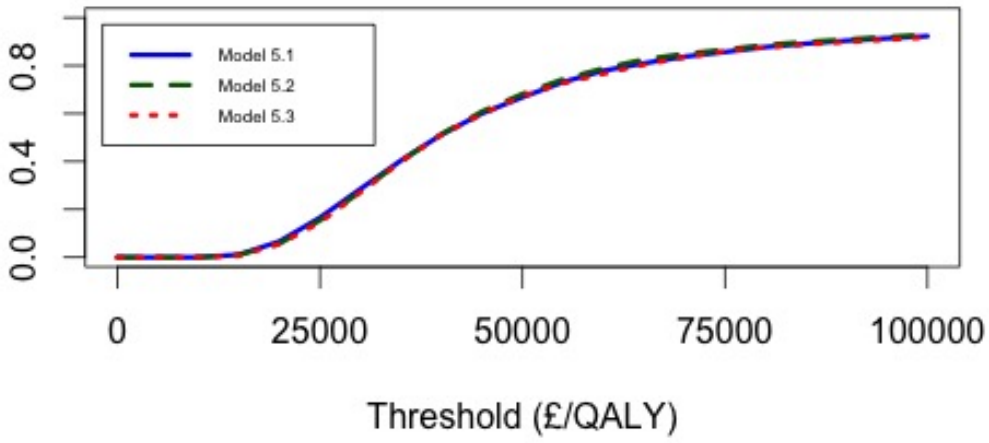
In order to inform the second objective, researchers can use their probabilistic results to estimate the expected value of perfect information (EVPI). The EVPI puts an upper bound estimate on the value of resolving the decision uncertainty. Figure 5.3 shows that the choice of methodology also has little impact upon the EVPI. However, the additional evidence exploited with the "all-inclusive" approach appears to have slightly reduced the cost of the decision uncertainty over the threshold range associated with NICE decisions (i.e. $20,000-30,000$ per QALY (NICE, 2013)). Despite the lack of variability in the cost-effectiveness results, it is important to remember that these findings are case study dependent. Additional analyses in Appendix C show that the impact of HRQoL is rela-

tively modest in the RITA-3 model.

Table 5.5: Error Probabilities

	Threshold = £20K	Threshold = £30K
Model 5.1	0.070	0.262
Model 5.2	0.057	0.265
M0del 5.3	0.051	0.276

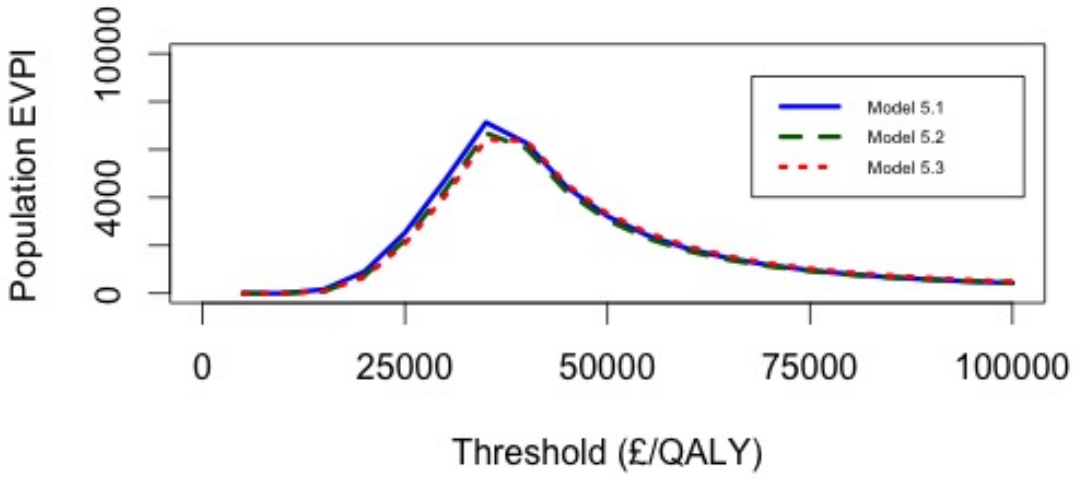
Figure 5.2: Cost-Effectiveness Acceptability Curves



5.5 Discussion

The aims of this chapter were to (a) identify the methodological challenges unique to scenario D1, and (b) to compare HRQoL parameter estimates, derived using an "all-inclusive" approach to evidence synthesis with SEM methods, against those obtained using a "reference case" approach to evidence synthesis with standard meta-analytic methods. Two separate approaches involving SEM techniques for the synthesis of heterogeneous HRQoL evidence were proposed, both of which rely upon the assumption of fixed factor loadings across studies. The two methods differed in terms of the point at which these factor loadings were applied to map between outcomes. A two-step method separated the mapping and evidence synthesis procedures, while an integrated approach conducted these tasks simultaneously. Of the two methods proposed, the *integrated* method would

Figure 5.3: Expected Value of Perfect Information



appear to be the logically correct approach insofaras it employs a single model to explain all aspects of the available data. The implementation of this method was made possible, in large part, thanks to the flexibility allowed by the Bayesian framework. However, the *integrated* approach has a greater number of unknown parameters and so is more likely to be affected by the choice of priors. Moreover, the additional complexity with the *integrated* method may create problems when it comes to the interpretation of results.

The additional evidence incorporated with the "all-inclusive" approach in this study did not have a substantial impact upon the parameter estimates compared to the "reference case" approach. However, it is important to acknowledge that these findings are case study dependent and, thus, should not be used to make conclusions about the benefits of the SEM evidence synthesis methods in general. These methods are likely to be particularly important where generic instruments, such as the EQ-5D, are subject to measurement error in the estimation of population parameters. For instance, HRQoL parameter estimates may be susceptible to noise occurring as a result of unrelated comorbidities. In scenarios such as these, parameter estimates derived solely using evidence pertaining to the reference case measure may exhibit a high degree of imprecision depending upon the combined sample size of patients. Increasing the sample size, through the inclusion of evidence relating to alternative HRQoL measures, provides greater statistical power to detect the true population values. Consequently, any improvements in the precision with which these parameters are estimated could potentially result in reduced decision uncertainty.

A key priority for future research will be to undertake simulation exercises to fully understand the mechanisms at work. The benefit of simulation exercises is that the researcher has full control over all of the underlying parameters feeding into the model. However, even if the "all-inclusive" approach had shown a more substantial impact upon the parameter estimates compared to the "reference case" approach, this would not have necessarily had an impact upon the cost-effectiveness results. The EVPPI results presented in Appendix C show that the impact of HRQoL is modest in the RITA-3 model. This means that the choice of method for the estimation of HRQoL parameters is unlikely to have a substantial impact upon the decision uncertainty for this study and, thus, would be unlikely to have any implications for policy decision-making. It is important to note that the influence of HRQoL parameters within a given decision model is a function of the model structure, as well as the other model parameters, and is therefore also case study dependent. In practice, researchers are unlikely to know the extent to which HRQoL parameters will influence cost-effectiveness prior to construction of a model. As such, this issue would not constitute grounds for assessing whether or not to pursue implementation of these methods.

Thus far, the discussion around uncertainty has focused upon the implications for parameter estimation of conducting evidence synthesis with SEM methods. The adoption of these methods also has important implications for structural uncertainty and methodological uncertainty in decision modeling for health economic evaluation. Methodological uncertainty occurs when there is disagreement over the most appropriate methodological approach (Briggs and Gray, 1998). One issue contributing to methodological uncertainty regarding the use of the SEM approach to evidence synthesis is the lack of consensus regarding the indirect estimation of HRQoL effects (i.e. mapping). As previously mentioned, mapping has traditionally been seen as a second-best option for the estimation of preference-based HQRoL values, when compared to the collection of values first hand (NICE, 2008). Taking this stance would imply that non-reference case HRQoL evidence should only be considered where directly relevant reference case evidence is not available. Of course, there are counterarguments that could be made against this position, not least the potential benefits of estimating parameters with greater precision. Moreover, one could point to the fact that there are other parameter estimation techniques employed in health economic evaluation that rely upon indirect evidence, i.e. treatment comparisons (Sutton et al., 2012).

Structural uncertainty refers to forms of uncertainty characterised in other ways from those already mentioned (i.e. not parameter or methodological uncertainty), including the sta-

tistical methods selected to estimate model parameters (Bojke et al., 2009). The LISREL approach, employed in Models 5.2 and 5.3, assumes that outcome measures are linearly related to one another and that the outcomes are normally distributed. It is important to recognise that the validity of these structural assumptions has been questioned in the context of HRQoL research (Alava et al., 2012; Basu and Manca, 2012). However, there is currently a lack of clarity as to the methods that should be used to evaluate structural uncertainties (Bojke et al., 2009). Future research should investigate the plausibility of alternative model specifications that are better suited to the handling of HRQoL outcome measures.

It is important to consider that the evidence selected for synthesis in this chapter was obtained from an existing review. This limited the case study to a synthesis exercise involving three preference-based measures of HRQoL. In theory, the SEM methods for evidence synthesis would allow researchers to draw upon the wealth of non-preference-based HRQoL measures available. As well as synthesising a broader range of outcome measures, the SEM methodology is capable of estimating parameters with greater precision by incorporating disease-specific outcome measures. A recent study by Ades and colleagues illustrated how this might be achieved using SEM-type mapping techniques (Ades et al., 2013). Of course, the inclusion of disease-specific measures would have important implications for the identification of HRQoL evidence given that the range of outcomes would vary according to the patient population of interest. Further empirical research is needed to demonstrate this potential and also establish how these considerations would be integrated within a search strategy. For instance, this could encompass a scoping process prior to the commencement of the literature search to identify all relevant outcome measures. For certain conditions, research may have even been conducted already to determine which patient-reported outcomes are appropriate for the measurement of HRQoL (Gibbons et al., 2014; Hadi et al., 2010; Mackintosh et al., 2009).

The empirical work outlined in this chapter is not without limitations. The case study in this chapter assumes a scenario where researchers can obtain a sample covariance matrix from an existing study. The plausibility of such information being readily accessible in the published literature has been debated elsewhere (Riley et al., 2014; Wei and Higgins, 2013). In practice, implementation of the SEM meta-analytic techniques is likely to rely upon the availability of IPD containing all of the relevant outcomes of interest. The availability of evidence in IPD format affords researchers analytical opportunities beyond the LISREL method. These opportunities are considered further in chapters 5 and 6.

Chapter 6

Case Study II: Synthesis of Individual Patient Data

6.1 Introduction

The previous chapter demonstrated how to use structural equation modeling techniques for the synthesis of heterogeneous HRQoL evidence from the published literature. An empirical case study showed that the SEM approach was able to exploit additional evidence compared to the 'reference case' approach, and, consequently, reduce the uncertainty surrounding the HRQoL parameter estimates. This finding was encouraging given that one of the main reasons for synthesising evidence is the possibility of attaining increased statistical power for the estimation of parameters (Higgins et al., 2009). This work raises important questions regarding the appropriate range of outcome measures to consider when searching for HRQoL evidence. Given that Chapter 5 only considered scenarios involving aggregate data, the analytical techniques were limited to the use of summary scores or index values.

The benefits of having access to individual-patient data for the purposes of synthesising HRQoL data with SEM techniques – in particular, the ability to exploit item-level information – have yet to be demonstrated in the context of health economic evaluation. Only one study has been found to employ an SEM-type approach, specifically a multidimensional item response theory (IRT) method, to analyse preference-based measures of HRQoL at the item level (Gibbons et al., 2014). This study showed that the multidimensional IRT model performed well in terms of its ability to predict EQ-5D scores compared to traditional regression methods. Despite showing early promise, the predictive performance of this approach has yet to be externally validated.

This chapter presents an empirical case study involving the estimation of HRQoL input parameters for a decision model derived through the analysis of heterogeneous PROs at the item-level data using SEM techniques. The aims of the chapter are threefold: first, to provide an outline of the factors that require consideration when utilizing these techniques specifically for the purposes of CEA. The second aim is to demonstrate how parameter estimates derived using heterogeneous outcome measures differ from those solely relying upon the reference case. And finally, this chapter looks to explore how parameter estimates derived via item-level analyses compare to those from analyses involving index scores at the aggregate level.

6.2 Background

There are several factors motivating the decision to focus on the development of methods that make use of item-level responses to PROs rather than index or summary scores. First, it has been suggested that item-level analyses are superior on the grounds of their interpretability; in contrast, analyses involving index values have been criticized because they conflate information about patient responses with the preference weightings (Parkin et al., 2010). Second, and related to the previous point, is the fact that item-level analyses exploit the detail of the available data. In particular, it would seem preferable to acknowledge impacts occurring at the item level that could be easily overlooked when dealing with index or summary scores. Finally, item-level analyses would be preferable for the purposes of this chapter in the interests of developing a generalizable framework. Although a range of sophisticated analytical approaches have been developed for the analysis of index values, these may need to be tailored to handle the specific characteristics of the instrument and value set under evaluation (Alava et al., 2012).

The primary interest of this chapter lies in the development of methods capable of handling 'Likert-type' responses to descriptive questionnaires that ask individuals to select one category form several options available (e.g. strongly disagree, agree, neither agree nor disagree, disagree, strongly disagree). With this in mind, it would seem logical to investigate methods for analysing ordinal variables given that these responses exhibit a natural ordering. The assumption of linear dependence between observed variables and a latent factor, made in the previous chapter, is no longer appropriate in this context since it does not account for the discrete nature of the data. This problem can be circumvented, however, by looking at the item responses in terms of the probability or odds of a particular response being selected.

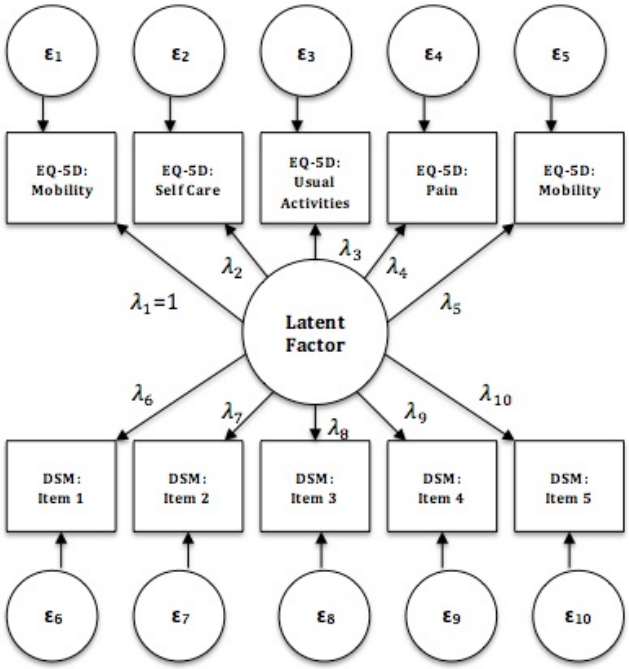
When thinking about the development of structural equation models for the investigation of ordinal, categorical variables, a useful starting point is to consider the latent factor. Despite the observed variables being categorical, it is often more convenient to treat the latent factor as being a continuous variable and to assume a non-linear relationship with the observed variables. This relationship is built upon an assumption that each of the observed categorical variables, known as *indicators*, has a separate underlying continuous scale (referred to here as the *latent response scale*). For a given indicator, each individual is assumed to lie somewhere on the latent response scale and it is the location on this scale that determines the observed category selected. Each of the latent response scales are assumed to have an associated distribution and the aim of the model is to investigate the extent to which a hypothesized latent factor accounts for correlations between these distributions.

Correlations between ordinal variables, such as those involving Likert-type responses to questionnaires, can be estimated using polychoric methods (Edwards et al., 2012). The latent response scales are assumed to follow a normal distribution, where the area under the curve represents the proportion of patients realizing a given response level. Distributions are constructed for each of these scales by assigning threshold values defining the point at which people change their selected response category. These values typically represent points on the standard normal distribution and can be used to approximate the proportion of patients in each category using the cumulative distribution function.

6.2.1 *Model Specification*

In the previous chapter, the process of model specification was simple to the extent that it only entailed a single latent factor. With the additional item-level information that is gained from having access to patient-level data, there is potential for greater complexity when specifying the underlying constructs. In order to consider the issue of latent factor specification, let us consider an illustrative example. Supposing that we are interested in developing an SEM to capture the shared variance between two outcome measures, the EQ-5D and a hypothetical disease-specific measure with five indicators, each with three category responses. To start with, consider the assumptions implied by a model with containing a single latent factor defined by all of the indicators across the two instruments, shown in Figure 6.1. The variability of each indicator is separated into two components: 1) the variability due to the latent factor and 2) unexplained error variance (represented by the ε terms).

Figure 6.1: Graphical Representation of a Unidimensional Model



For estimation involving a probit link, the model can be represented using the following formula from Bovaird and Koziol (2012):

$$\Pr(y_{ij} = 1) = \Phi(-\tau_{1,j} + \lambda_j \cdot f_i) \tag{6.1}$$

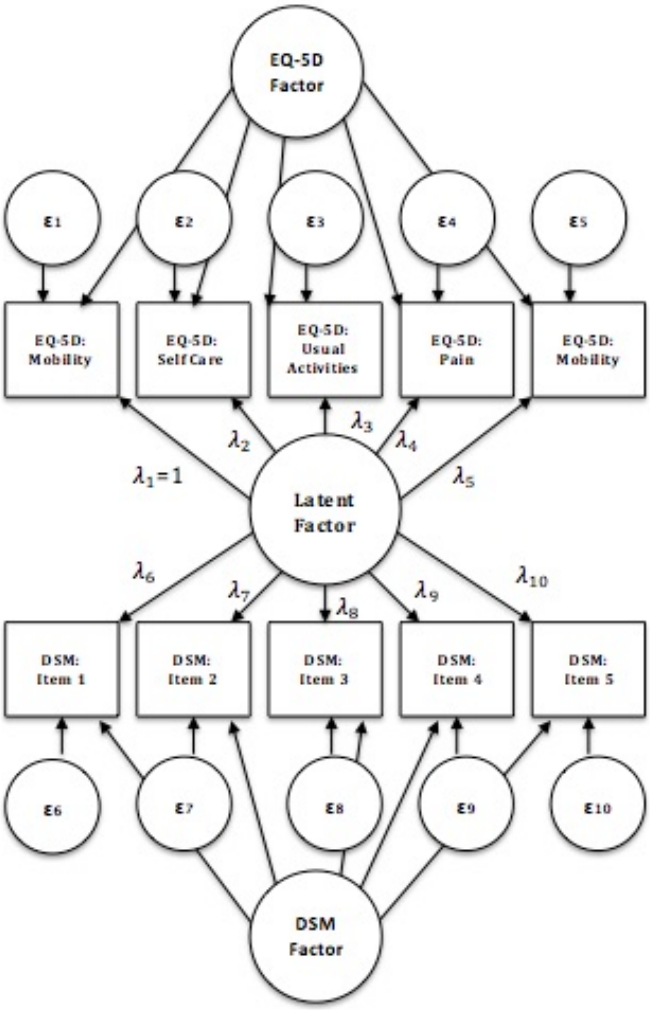
$$\Pr(y_{ij} = 2) = \Phi(-\tau_{2,j} + \lambda_j \cdot f_i) - \Phi(-\tau_{1,j} + \lambda_j \cdot f_i) \tag{6.2}$$

$$\Pr(y_{ij} = 3) = 1 - \Phi(-\tau_{2,j} + \lambda_j \cdot f_i) \tag{6.3}$$

Where y_{ij} represents individual i's response to item j, f_i is individual i's latent factor score and λ_j is the factor loading for item j. The $\tau_{k,j}$ terms are thresholds indicating the point on the latent factor response scale where individuals switch their item response. Φ is the standard cumulative distribution function. These formulae show the probability of an individual selecting each of the possible item responses.

Crucially, this model in Figure 6.1 assumes that the error terms are independent of one another. One might question this assumption if there are concerns about multidimensionality especially in the presence of well-defined clusters of items. An alternative model specification that researchers might consider for scenarios involving multidimensionality in addition to a common trait is the bi-factor model (Reise, 2012). An example is provided in Figure 6.2 that assumes clustering is present at the instrument level. Conceptually,

Figure 6.2: Graphical Representation of a Bifactor Model



this model posits that one general variable explains common variance in all of the items but beyond that there is some instrument-specific variability left over. The variability of each indicator is now separated into three components: 1) the variability due to the general factor, 2) the variability due to the measure-specific factor (EQ5D or DSM), and 3) unexplained error variance.

In the study by Gibbons and colleagues, a bi-factor IRT model was specified containing one primary dimension, loading on all of the items, and two scale-specific dimensions, loading on EQ-5D and SF-12 items respectively. The same study also conducted a separate analysis using a unidimensional model (i.e. without the scale-specific latent factors) and compared the fit of the two models using a likelihood ratio Chi-squared statistic. This test found an improvement in fit with the bi-factor model compared to the unidimensional model.

6.2.2 *Estimation of Parameter Inputs*

As with the previous chapter, there are two plausible approaches for the estimation of CEA inputs, a CFA method and a SEM method. Once again, the former is essentially a mapping procedure that aims to predict the reference case outcome measure in datasets where this measure has not been collected. However, the task of deriving predictions with a CFA mapping algorithm is far more complex when dealing with item-level responses, compared to index/summary scores. This is because the availability of multiple observed items, representing a given construct of interest, introduces a dimensionality problem. Latent variable scoring methods are required in order to reduce the dimensionality and obtain a unique factor score for each individual. Using the parameter estimates from the CFA model, one can predict the position of a patient on the latent factor scale even if there is only a subset of the items in the original model. Subsequently, one can then use the factor score estimate in combination with the parameter estimates for the missing items to predict responses for those items.

While there are a range of alternative methods for assigning values to latent variables, the most popular is the Empirical Bayes (EB) method (see for instance, Skrondal and Rabe-Hesketh (2004)), also known as the 'expected a posteriori' (EAP) predicator. This method can be implemented in the Mplus software package (Muthén and Muthén, 2010), using the GSEM package for Stata (StataCorp, 2013) and using the Lavaan package for R (Rosseel, 2012). The EB method derives a posterior mean prediction of the latent variable ($\hat{f}_i$) for each individual in the sample using their observed responses to each of the indicators (y_{ij}) in combination with parameter estimates from the CFA (λ_j,$\tau_{k,j}$) as follows:

$$\hat{f}_i = \mathrm{E}(f_i|y_{ij};\lambda_j;\tau_{k,j}) \tag{6.4}$$

Whilst this method is actually a *Frequentist* technique, its name refers to the fact that it utilizes Bayesian principles when interpreting the latent variable in terms of a posterior distribution. A *fully* Bayesian approach differs from the EB method given that it would interpret both the latent variable and the parameter estimates in terms of a posterior distribution.

Another challenge associated with the CFA prediction method is the fact that most software packages are set up to obtain factor scores in the same sample used to fit the original model. Although this makes the task of obtaining predictions for individuals missing the reference case items more complicated, there are two ways in which this problem can be circumvented. The first, implemented by Gibbons and colleagues, involves combining all

of the datasets of interest before fitting the CFA under the assumption of missing data for those individuals for whom the reference case items have not been collected. The full information maximum likelihood (FIML) estimation method ensures an efficient and unbiased utilization of all of the available information when fitting the CFA (Graham and Coffman, 2012).[1] By fitting the model with data that includes observations missing the reference case items, it is possible to estimate factor scores for these cases.

Although the FIML-estimation method provides an elegant and efficient solution for the utilization of observations with missing data, the software options currently available do not enable an adequate characterization of parameter uncertainty. Cost-effectiveness models are typically expected to use probabilistic sensitivity analysis (PSA) to capture the impact of parameter uncertainty upon the cost-effectiveness results (Briggs et al., 2006). HRQoL parameter estimates derived using the techniques considered in this chapter are subject to two sources of uncertainty: the unexplained heterogeneity surrounding the underlying latent factor, and the uncertainty surrounding the factor loadings. In principle, both forms of uncertainty could be captured within a PSA but, unfortunately, none of the software options available are capable of incorporating the latter.

Given the current software available, a second-best approach for obtaining factor scores is required to capture all of the possible sources of parameter uncertainty. An alternative method would be to fit the CFA using only the complete case observations and then, with the incomplete data, fit subsequent models with parameter estimates fixed to those from the first analysis. Crucially, the subsequent models need to have free parameters associated with the latent variable (mean and variance) in order to permit the estimation of factor scores. Once factor scores have been obtained for those observations missing the reference case items, they can be used to predict the expected responses to these items using the relevant factor loadings and threshold values from the complete case analysis.

The previously mentioned issues in dealing with categorical data mean that item-level responses are predicted in terms of the probability of a particular response being selected. These probabilities can then be used, in combination with the index weightings of the associated instrument, to estimate the expected HRQoL index scores for each patient (Le and Doctor, 2011). In order to capture uncertainty surrounding the factor loadings and threshold values, the process of obtaining factor scores and then using these scores to predict the expected HRQoL index values must be incorporated within the probabilistic sensitivity analysis. A stochastic process is employed using the standard errors associated with the CFA parameter estimates.

[1] Assuming that data are missing completely at random

The SEM approach to parameter estimation differs from the CFA approach in that it is not concerned with the prediction of factor scores at the observation level. It employs almost exactly the same model specification as the CFA approach but with the addition of a structural component capturing the impact of covariates upon the latent variable (also known as a path analysis). As such, the SEM approach to HRQoL parameter estimation is possible via prediction of the expected latent factor distribution conditional upon patient characteristics. Given that the SEM approach is not concerned with the prediction of scores at the observation level, the model can be fit to a combined dataset that includes incomplete cases using the FIML estimation procedure. This approach allows the prediction of a latent factor distribution for a given patient profile, as specified by the covariates in the SEM.

Using the latent factor distribution, item-level responses can be predicted for a given patient profile. In this chapter, a Monte Carlo simulation procedure will be employed in order to predict item-responses using the parameter estimates obtained from the SEM. As with the CFA approach, it is important to ensure that the uncertainty surrounding the parameter estimates in the statistical model has been acknowledged. A stochastic process is implemented involving all of the parameters involved in the SEM. The simulation sample should be sufficiently large to ensure that the latent factor distribution is adequately represented. Once the simulation sample has been predicted the valuation weights can be assigned in order to derive HRQoL index values, after which one can calculate the associate mean and standard deviation.

6.3 Methods

The case study in this chapter is composed of two stages. In the first stage, HRQoL parameter inputs are derived using one of the alternative estimation techniques. In the second stage, the parameter estimates are fed into the model and cost-effectiveness estimates are obtained for each of the alternative methods. The case study selected investigates the cost-effectiveness of an early surgical intervention compared to medical management for the treatment of acute coronary syndrome in a patient subgroup with diabetes.

6.3.1 *HRQoL Evidence*

As with the previous chapter, this case study aims to estimate HRQoL values associated with two defined health states: one for patients with diabetes who have yet to experience a myocardial infarction and another for patients with diabetes who have experienced a

myocardial infarction. Data from four studies that collected PROs for the measurement of HRQoL in relevant patient subgroups were obtained.

The National Health Measurement Survey (NHMS) was conducted in the United States and collected a variety of measures of HRQoL, including the EQ-5D, HUI-3 and the SF-36, in a sample representative of the general population (Fryback et al., 2007). The Medical Expenditure Panel Survey (MEPS) data collected the EQ-5D and the SF-12 in a nationally representative sample of the United States in 2003 (Sullivan et al., 2011). The Welsh Health Survey (WHS) collected the SF-36 in a representative sample of people living in private households in Wales in 2013 (NatCen Social Research, 2013). Finally, the Randomized Intervention Trial of unstable Angina (RITA-3) collected the EQ-5D in a sample of patients diagnosed with acute coronary syndrome (Kim et al., 2005).

Whilst the EQ-5D may have been collected in the MEPS dataset, we will be assuming that it is not available for illustrative purposes. Instead, EQ-5D responses will be predicted using the available SF-12 data via the SEM approach. The reasoning behind this approach is that it allows us to test the external validity of the SEM predictions (i.e. compare the predicted EQ-5D scores against the actual scores).

6.3.2 *Health Economic Model*

The health economic model employed is the same as that used in Chapter 4, where further details describing this model can be found.

6.3.3 *Estimation of Parameter Inputs: CFA Approach*

The CFA approach is implemented by first fitting a bi-factor model to the NHMS data. This dataset is selected on the grounds that it is the only one containing all of the PRO items of interest. The bi-factor model contains three latent variables: a common factor explaining all of the items for the EQ-5D and the SF-12; an EQ-5D-specific factor explaining the EQ-5D items; and a SF-12-specific factor explaining the SF-12 items. Once this model has been fitted, the resulting parameter estimates are used in subsequent analyses to obtain latent factor scores in the datasets without the EQ-5D as follows:

- In order to estimate factor scores for patients in the MEPS dataset a single factor model explaining all of the SF-12 items is specified. Estimated factor loadings that explain SF-12 item responses in terms of the common factor are taken from the NHMS analysis and used as fixed parameters in this model, which is then fitted to the MEPS dataset. Once this model has been fitted, latent factor scores can be predicted using the Empirical

Bayes (EB) method. These factor scores, in combination with the parameter estimates from the NHMS analysis, can be used to predict the expected EQ-5D responses and index values.

- In order to estimate factor scores for patients in the WHS dataset a single factor model explaining all of the SF-12 items is specified. Although SF-36 responses were collected in this dataset, there were concerns that differences in the items responses between the NHMS and WHS datasets would lead to invalid predictions. There were substantial differences between the NHMS and WHS datasets in terms of subjects' responses to questions 33 through to 36 on the SF-36. In the NHMS data, fewer than 2% of subjects were found to select a "Don't know" response in these questions; in contrast, between 15 – 35% of subjects selected this response in the WHS data. For this reason, it was felt that it would be more appropriate to select the items corresponding to the SF-12 instrument and employ the same method as that with the MEPS data.

The predicted EQ-5D values can be used to derive inputs for the cost-effectiveness model. These are the sample moments (mean and standard deviation) associated with the two health states captured in the cost-effectiveness model: one for patients with diabetes who have yet to experience a myocardial infarction and another for patients with diabetes who have experienced a myocardial infarction. Note that the aforementioned methods only provide parameter estimates for a deterministic cost-effectiveness analysis. For the purposes of a probabilistic analysis, the steps undertaken to estimate latent factor scores and, subsequently, predict EQ-5D responses need to be implemented via a stochastic process.

6.3.4 *Estimation of Parameter Inputs: SEM Approach*

The model specification of the SEM approach is essentially the same as that in the CFA approach with the addition of a structural component capturing the impact of patients having experienced a MI. In contrast to the CFA approach, this model is fit to a sample combining the all of the datasets. In order to make use of observations that have missing data, a FIML estimation procedure is employed. The estimated model captures a latent factor distribution associated with each of the health states of interest, which can be subsequently used to predict item-level responses using Monte Carlo simulation methods. As with the CFA approach, a stochastic process is necessary in order to capture the uncertainty surrounding the factor loadings and threshold values. Moreover, to ensure that the predicted item responses reflect the latent factor distribution, a large number of

simulations are run (100,000 for each health state). Model inputs can then be obtained by calculating the means and standard deviations for each of the simulated samples.

6.3.5 *Estimation of Parameter Inputs: Reference Case Approach*

An additional analysis was undertaken to estimate HRQoL parameter inputs solely using EQ-5D. Estimation of the HRQoL input parameters simply involved calculating the means and standard deviations for the two groups of patients in a sample comprised of data from the RITA-3 and NHMS studies.

6.3.6 *Statistical Software*

The CFA approach to parameter estimation was implemented in the R software program using the Lavaan package. For the SEM approach, the initial model was fitted using Mplus software package and implemented using the FIML procedure for handling missing data. The cost-effectiveness model inputs were subsequently obtained by simulating the expected responses to the EQ-5D based upon the initial model using the Simsem package in R (Pornprasertmanit et al., 2013). For each of the different methods, a stochastic process was implemented to capture the uncertainty surrounding these estimates. The subsequent parameter estimates are then fed into the RITA-3 health economic model, also built in R. A total of 3,000 simulations are run in the probabilistic sensitivity analysis to derive the probability of the intervention being cost-effective for a given threshold. Measures for comparison of model fit, such as the Akaike information criterion, are unsuitable in this case study given that these methods are used to compare model fit for a given dataset. Even though the alternative SEM approaches are applied to the same data, information criteria comparisons would not be of any use given that one approach uses a single model to simultaneously map and synthesise evidence whilst the other does this in two separate stages.

6.4 Results

6.4.1 *Descriptive Statistics*

Table 6.1 provides descriptive statistics for each of the datasets used in this empirical exercise. There are noticeable discrepancies between the samples grouped according to health state description (i.e. whether or not patients have experienced a previous MI), both in terms of the age and gender composition. If there is reason to believe that such imbalances may bias the parameter estimates, it is important that statistical techniques are

employed to control for them. Previous research has shown that HRQoL varies according to both age and gender (Ara and Brazier, 2010; Kind et al., 1999).

Table 6.1: Descriptive Statistics

	NHMS	MEPS		WHS		RITA-3	
	No MI	No MI	Post MI	No MI	Post MI	No MI	Post MI
Total Sample	552	1,335	189	875	150	156	88
Analysis Sample	520	1,228	175	729	105	156	88
Age	63.6	58.9	67.9	61.8	68.6	64.2	63.9
% Male	0.41	0.40	0.58	0.56	0.64	0.57	0.66

6.4.2 *Model Results: The CFA Approach*

The factor loadings for the bi-factor CFA model, fitted to the NHMS data, are presented in Table 6.2. These values show the extent to which each of the items are correlated with the associated latent factor; a higher factor loading estimate corresponds to a higher degree of correlation with the factor loading. For the purposes of HRQoL parameter estimation, the loadings on the CF latent variable are of primary interest. Increases on the latent factor scale reflect decreases in HRQoL. As such, items with responses ordered in a way to reflect diminishing health – e.g. the items of the EQ-5D having the best health states coded as 1 and the worst as 3 – are shown to have positive factor loadings. Conversely, those items with responses ordered to reflect improved health have negative factor loadings.

The item threshold values for the bi-factor CFA model can be found in Appendix E. These estimates can be interpreted as reflecting the item difficulty; how far away from the mean on the latent factor scale, at zero, does an observation need to be to give rise to a change in the item response. Whilst interpreting these estimates might be challenging, especially given that they reflect values on the standard normal distribution, they can be converted into probabilities using the cumulative normal distribution function.

A chi-squared test of model fit was employed to explore whether or not the multi-dimensional assumption seems reasonable. The result (p=0) suggests that it would be reasonable to

assume that the bi-factor model fits the data better than a uni-dimensional approach.[2] To ensure comparability of the modelling techniques explored in this chapter, the bi-factor CFA model was run in Mplus, as well as R, and the results were shown to be equivalent.

Table 6.2: CFA Approach - Factor Loadings

Indicator	Latent Factor 1 (CF)	Latent Factor 2 (SF-12)	Latent Factor 3 (EQ-5D)
SF36 Q1	0.55	0.14	0.00
SF36 Q2	-0.84	-0.24	0.00
SF36 Q3	-0.77	-0.19	0.00
SF36 Q4	-0.72	-0.34	0.00
SF36 Q5	-0.82	-0.37	0.00
SF36 Q6	-0.81	0.30	0.00
SF36 Q7	-0.76	0.29	0.00
SF36 Q8	0.83	0.06	0.00
SF36 Q9	0.58	-0.35	0.00
SF36 Q10	0.69	-0.05	0.00
SF36 Q11	-0.76	0.40	0.00
SF36 Q12	-0.77	0.80	0.00
EQ5D Q1	0.75	0.00	0.41
EQ5D Q2	0.77	0.00	0.25
EQ5D Q3	0.87	0.00	0.38
EQ5D Q4	0.74	0.00	0.37
EQ5D Q5	0.81	0.00	-0.25

6.4.3 *Model Results: The SEM Approach*

The model specification for SEM approach is essentially the same as that for the CFA approach except for the addition of a structural component included to capture the impact of a myocardial infarction upon the common factor CF. Following initial concerns that imbalances in patient characteristics might bias the estimated impact of a MI, three additional control variables were included in the structural component: one for gender and two dummy variables for age categories (one for subjects between 50 and 80 years old and another for those patients over 80 years old).

Table 6.3 shows the factor loadings and regression coefficients capturing the impact of the aforementioned covariates upon the latent factor CF. Increases on the scale of the

[2]We can reject a null hypothesis that inclusion of instrument-specific factors does not significantly improve model fit.

latent variable CF reflect improvements in HRQoL. The factor loadings have remained remarkably similar to those obtained with the CFA approach. The latent factor CF is estimated to be over half a standard deviation lower in those subjects who have previously had an MI compared to those subjects who have never experienced an MI. Moreover, factor scores are estimated to be higher in men and lower with increasing age. An additional model was fitted that did not control for age and gender to compare the coefficients for the MI variable.

Table 6.3: SEM Approach - Factor Loadings

Indicator	Latent Factor 1 (CF)	Latent Factor 2 (SF-12)	Latent Factor 3 (EQ-5D)
SF36 Q1	-0.70	0.07	0.00
SF36 Q2	0.86	-0.23	0.00
SF36 Q3	0.83	-0.26	0.00
SF36 Q4	0.89	-0.19	0.00
SF36 Q5	0.93	-0.20	0.00
SF36 Q6	0.80	0.48	0.00
SF36 Q7	0.78	0.39	0.00
SF36 Q8	-0.81	0.11	0.00
SF36 Q9	-0.59	-0.30	0.00
SF36 Q10	-0.71	-0.04	0.00
SF36 Q11	0.66	0.39	0.00
SF36 Q12	0.82	0.11	0.00
EQ5D Q1	-0.72	0.00	-0.52
EQ5D Q2	-0.71	0.00	-0.31
EQ5D Q3	-0.80	0.00	-0.40
EQ5D Q4	-0.61	0.00	-0.08
EQ5D Q5	-0.68	0.00	0.24
MI	-0.67	0.00	0.00
Age (50 - 80)	-0.16	0.00	0.00
Age (80+)	-0.41	0.00	0.00
Gender (Male)	0.36	0.00	0.00

As with the CFA approach, a chi-squared test of model fit was employed to explore whether or not the multi-dimensional assumption seems reasonable. Once again, the result (p=0) suggests that it would be reasonable to assume that the bi-factor model fits the data better than a uni-dimensional approach.

6.4.4 *Parameter Estimates*

Table 6.4 shows the estimated HRQoL parameter inputs for the cost-effectiveness model derived. There are clear differences between the estimates derived using heterogeneous outcome measures (Models 6.2 and 6.3) compared to those that solely relied upon the reference case measurement (Model 6.1). The parameter inputs derived for Model 6.1 show that increases in age are associated with improvements in HRQoL. This counterintuitive finding is likely to be caused by the unusual associations between age and HRQoL in the RITA-3 data (Mahon, 2014). In addition, this finding may go some way in explaining why these estimates are so much higher than those for Models 6.2 and 6.3.

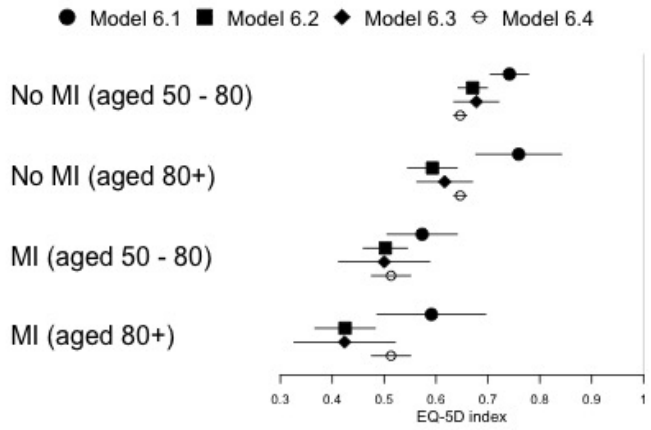
Although similar, the discrepancies between estimates from Models 6.2 and 6.3 are likely to be driven by the different methods used to account for covariates. In Model 6.2, EQ-5D predictions were obtained for the MEPS and WHS datasets before the impact of the covariates (MI, age and gender) were derived using a pooled sample using these predicted values in combination with the observed EQ-5D values (from the RITA-3 and NHMS datasets). In contrast, Model 6.3 accounted for the impact of the covariates upon the latent factor first and then EQ-5D predictions were derived for each patient profile.

Table 6.4: Mean HRQoL Parameter Inputs

	No Previous MI		Post MI	
	50 - 80 years	80 years +	50 - 80 years	80 years +
Model 5.1	0.742	0.759	0.574	0.592
Model 5.2	0.671	0.593	0.502	0.424
Model 5.3	0.640	0.605	0.533	0.480
Model 5.4	0.647		0.513	

In addition to looking at the expected EQ-5D values for model inputs, it is important to also consider the impact of the alternative synthesis methods upon the uncertainty associated with these values. In the case of the inputs derived via Models 6.1, 6.2 and 6.4, the uncertainty associated with the expected values is a function of the standard errors surrounding the parameter estimates listed in Tables 6.2 and 6.3 (i.e. those for the factor loadings, threshold values and structural components). The forest plot in Figure 6.3 provides a graphical representation of the estimated EQ-5D distributions for all of the patient profiles defined in the RITA-3 model, which were obtained in the probabilistic sensitivity analysis.

Figure 6.3: HRQoL Parameter Estimates



Whilst the figures suggest that the uncertainty associated with the model inputs is lowest when statistical Model 6.4 is employed, the results from this model are not directly comparable to the other methods given that there were fewer covariates utilized (age and gender). It may be that the lack of adjustment for these covariates partially explains the reduced uncertainty compared with the other methods presented. The simulations obtained from statistical Model 6.2 exhibited reduced uncertainty compared to those from the reference case approach (Model 6.1). This was particularly pronounced for the inputs associated with patients aged 80 years or older. However, comparison of the results for Models 5.1 and 5.3 would suggest that the impact of additional, non-reference case evidence has an ambiguous effect upon the uncertainty surrounding the HRQoL inputs.

6.4.5 *Predictive performance*

The external validity of Model 6.2, estimated using the NHMS data, was verified by comparing EQ-5D predictions against observed EQ-5D values in the MEPS dataset. Table 6.5 shows that the predictive performance of Model 6.2, measured using the mean squared error, is superior overall to that of the common factor model approach, first proposed by Lu and colleagues (2013), across most of the EQ-5D distribution.

6.4.6 *Cost-Effectiveness Results*

The differences between parameter estimates obtained via the different synthesis techniques only really matter if they give rise to substantial differences in the cost-effectiveness estimates used to inform the associated reimbursement decisions. In view of that, parameter estimates from statistical Models 6.1 to 6.4 were utilized in the RITA-3 model in order to explore the impact of the methodological differences upon the cost-effectiveness results.

Table 6.5: Mean Predictive Performance (MEPS data)

	CFM	Bi-factor CFA
Whole Sample	0.059	0.046
EQ-5D ≤ 0	0.096	0.084
$0 <$ EQ-5D ≤ 0.25	0.053	0.075
$0.25 <$ EQ-5D ≤ 0.5	0.063	0.072
$0.5 <$ EQ-5D ≤ 0.75	0.064	0.049
$0.75 <$ EQ-5D ≤ 1	0.049	0.028

Tables 6.6 – 6.9 show the implications of using the different methods in terms of the impact upon the expected cost-effectiveness results. In addition, four separate probabilistic sensitivity analyses were run, one for each of the statistical models, the results of which have been presented in the form of cost-effectiveness acceptability curves in Figure 6.4. . This information is also presented in terms of error probabilities in Table 6.10, assuming threshold values of £20,000 per QALY and £30,000 per QALY respectively. These results present the probability of making an incorrect decision, for a given patient, on the basis of the expected cost-effectiveness results. Finally, the expected value of perfect information was estimated for each of the statistical models and curves presenting this information can be found in Figure 6.5.

Table 6.6: Cost-Effectiveness Results Using Model 6.1 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£39,940	£32,188	£4,752
QALYs (discounted)	12.34	12.21	0.13
Cost-per-QALY	-	-	£36,163

Table 6.7: Cost-Effectiveness Results Using Model 6.2 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£39,940	£32,188	£4,752
QALYs (discounted)	11.00	10.88	0.12
Cost-per-QALY	-	-	£39,960

Table 6.8: Cost-Effectiveness Results Using Model 6.3 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£39,940	£32,188	£4,752
QALYs (discounted)	11.12	11.00	0.12
Cost-per-QALY	-	-	£39,198

Table 6.9: Cost-Effectiveness Results Using Model 6.4 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£39,940	£32,188	£4,752
QALYs (discounted)	10.76	10.65	0.11
Cost-per-QALY	-	-	£42,054

One notable observation in each of these figures is the close proximity of the curves involving heterogeneous HRQoL evidence (Models 6.2, 6.3 and 6.4). Moreover, Figure 6.4 shows that the uncertainty surrounding the decision in question is reduced in these cases when compared to the results obtained using the homogeneous HRQoL evidence (statistical Model 6.3). This is because the probability of the early intervention being cost-effective is reduced, i.e. we can be more confident that it is not cost-effective. Figure 6.5 also shows that, over a threshold range between £20,000 and £30,000 per QALY, the cost associated with the decision uncertainty is greater for the scenario involving homogeneous HRQoL evidence compared to the scenarios with heterogeneous evidence.

6.5 Discussion

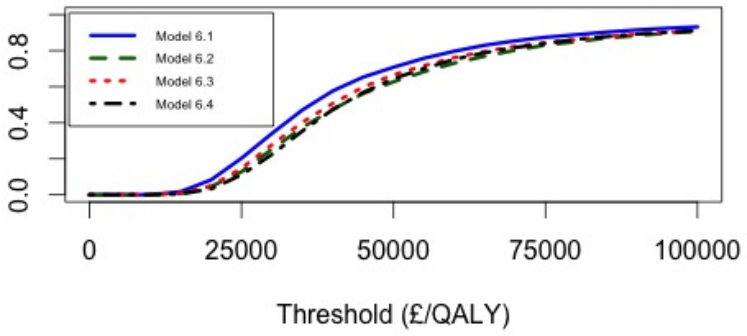
The primary objective of this chapter was to explore the benefits of having access to IPD – specifically, item-level data – when synthesising HRQoL evidence for the purposes of cost-effectiveness analysis. Investigations were focused on the use of methods capable of handling item responses given that this would apply to all PRO measures involving likert-type questionnaire responses. Moreover, item-level analyses allow researchers to make greater use of the available information and, unlike analyses involving summary scores, avoid overlooking item-level effects.

Structural equation modelling with categorical item variables is possible thanks to the development of polychoric methods. These methods are extremely flexible and typically

Table 6.10: Error Probabilities

	Threshold = £20K	Threshold = £30K
Model 6.1	0.083	0.340
Model 6.2	0.046	0.253
Model 6.3	0.049	0.277
Model 6.4	0.033	0.224

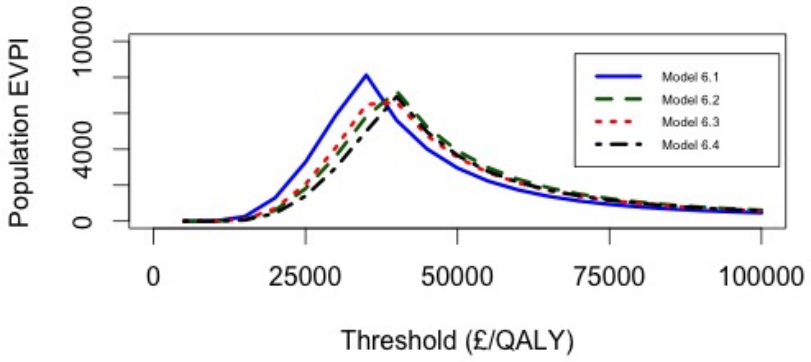
Figure 6.4: Cost-Effectiveness Acceptability Curves



assume that item responses correspond to points on an underlying latent factor with a continuous scale. There is only one existing study known to have implemented these methods in the context of HRQoL evidence synthesis for CEA. Gibbons and colleagues employed a bi-factor IRT model and found that this approach showed good predictive performance when compared to other mapping techniques (Gibbons et al., 2014). Unfortunately, the study by Gibbons and colleagues only demonstrated the predictive performance of the bi-factor IRT approach in terms of its internal validity. Moreover, their study did not consider some of the challenges likely to be encountered by researchers implementing these methods in conjunction with a cost-effectiveness model, principally of how to account for the uncertainty surrounding the parameter estimates.

Whilst fitting SEM models is relatively straightforward, the prediction of PRO responses via latent factor variables is made more challenging by the need to predict latent factors. Furthermore, there is additional complexity involved in the process of accounting for parameter uncertainty. In this chapter, two alternative approaches have been put forward for the prediction of PRO responses. The first method involved fitting a CFA-type model to data containing all of the items of interest and then using the output to estimate latent factor scores in data that only collected a subset of the items. The second method involved

Figure 6.5: Expected Value of Perfect Information



fitting an SEM-type model with an additional structural component in order to control for the variables of interest in the cost-effectiveness model. Once this model was fitted, the predictors assigned values to reflect the patient profiles of interest in the CE model. For each of the aforementioned prediction methods, it is vital that researchers take account of the uncertainty surrounding the factor loadings and threshold values. Otherwise, they might risk underestimating the uncertainty that is introduced by using indirect estimates of HRQoL.

Of the two item-level methods put forward in this chapter, the SEM approach offers a more computationally efficient approach to the prediction of PRO responses. This method simply involves fitting the specified model to a dataset combining all of the relevant data and then using the output of this analysis to predict PRO responses. The CFM approach involves two stages before PRO responses can be predicted: first, fitting a model to the data that collected all of the items of interest and, second, using the output of this analysis in a subsequent model that is then fitted to the dataset(s) containing a subset of the items in the original model. Unfortunately, the SEM approach cannot currently be implemented in the R software package. Instead, this approach requires software, such as Mplus, capable of deriving latent factor scores following the implementation of a FIML-estimation procedure.

The case study in this chapter did not show any clear impact of incorporating a broader range of HRQoL measures in terms of either parameter or decision uncertainty. However, it is important to acknowledge that these findings are case study dependent. While the 'all-inclusive' approach can potentially lead to increased statistical power for the estimation of HRQoL parameters compared to the 'reference case' approach, this does not guarantee that the decision uncertainty will be reduced. For instance, the gains from having an increased sample size might be offset by the uncertainty introduced as a result of the additional

parameters included to estimate factor scores and to predict HRQoL responses (i.e. factor loadings and threshold values). Several of the issues pertaining to uncertainty around cost-effectiveness results obtained using SEM methods from the previous chapter also apply to this chapter. The 'all-inclusive' approach to evidence synthesis with SEM methods is likely to be most beneficial when the reference case measure (e.g. EQ-5D in the UK) is subject to large amounts of measurement error in the estimation of population parameters. There is also methodological uncertainty stemming from the indirect estimation of HRQoL effects (i.e. mapping).

As with the methods in the previous chapter, there is uncertainty regarding the structural assumptions made in the specification of the SEM models in this chapter. One might question whether it is appropriate to assume that all of the HRQoL effects of interest should be captured on a single health domain (i.e. CF). Depending upon the condition under evaluation and the outcome measures available, it may be necessary to employ more complex model specifications in order to disentangle HRQoL effects. For instance, it may be more appropriate to model physical and mental health effects separately from one another (Gibbons et al., 2014). Although this potential complexity may seem overly burdensome, it could also present additional opportunities in the modeling of HRQoL effects. For certain disease areas, such as psoriatic arthritis (Kavanaugh et al., 2016), patients may experience several distinctive morbidity effects that need to be captured in a cost-effectiveness model. The flexibility of SEM methods means that it would be possible to explicitly account for these morbidity effects in the modeling approach. Future research is required to explore the issue of structural uncertainty in the specification of SEM models for the purposes of evidence synthesis. Existing research from the field of psychometrics would be well placed to inform these investigations given that there have been significant methodological advances to contend with the complex relationships often hypothesized in this field.

The empirical exercise conducted in this chapter demonstrated that parameter estimation involving the synthesis of heterogeneous HRQoL evidence with SEM techniques is preferable using IPD, rather than AD, for several reasons. First, IPD allows researchers to control for covariate imbalances in the data that would otherwise bias the parameter estimates of interest. Related to this is the fact that IPD facilitates the exploration of heterogeneity according to patient characteristics, which can reduce the uncertainty surrounding cost-effectiveness results. Although covariate adjustment can also be conducted using aggregate data, there is always a danger of bias (Piantadosi et al., 1988). Furthermore, access to IPD ensures that any correlations between covariates can be accounted for

in the probabilistic sensitivity analysis. IPD also allows researchers to conduct analyses at the item level, rather than having to rely upon linear analyses of summary scores, which have been shown to be superior in terms of predictive performance.

The models employed in this chapter assume that observations obtained from different studies can be treated as being random samples from a single population. Given that the datasets in this chapter differ in a number of respects – in terms of patient demographics, inclusion criteria, study setting and study design – it may be more appropriate to treat the data as being heterogeneous with observations being drawn from different groups, i.e. a hierarchical structure. Unfortunately, the fact that the datasets also differed from one another in terms of the item variables collected meant that a hierarchical model could not be implemented in Mplus.[3] Further research around this issue is necessary, particularly in relation to the use of a Bayesian multi-level approach (Lee, 2007).

Finally, the results of this chapter cannot be extrapolated should not be extrapolated to scenarios beyond the synthesis of heterogeneous generic PRO measures. While it is anticipated that the same methods could be employed in other scenarios, the implications of incorporating disease-specific measures in addition are unclear. Existing research has shown that mapping via disease-specific measures generally tends to exhibit a worse predictive performance than that via generic measures (Brazier et al., 2010). Comparatively, recent work has proposed that mapping via disease-specific measures has the potential to improve the precision of parameter estimates (Ades et al., 2013). Further research should prioritize the application of these methods in scenarios involving disease-specific measures.

[3]Lavaan does not currently include a hierarchical SEM estimation procedure.

Chapter 7

Case Study III: Synthesis of Aggregate and Individual Patient Data

7.1 Introduction

Chapter 6 demonstrated that the availability of HRQoL evidence at the IPD level offers substantial benefits for researchers seeking to synthesise that evidence with SEM techniques. Access to IPD permits the analysis of PBMs in terms of their item responses - rather than their index values - yielding parameter estimates that exhibit lower levels of prediction error compared to the estimates that would have been obtained from AD. Furthermore, IPD allows researchers to adjust for covariate imbalances and explore heterogeneity according to patient characteristics. Unfortunately, it is unlikely that researchers will be able to obtain raw data for all of the studies that are relevant to their cost-effectiveness model and instead will have to rely upon a combination of IPD and AD. This poses the question over how such evidence should be combined given that it will be composed of summary scores and/or index scores (in the case of AD), as well as item-level evidence (in the case of IPD). A simple solution would be to obtain summary scores and/or index values from the IPD studies available and to synthesise these in the AD format, along with the remaining AD studies (using one of the methods outlined in Chapter 5). However, the downside of this approach is that it fails to exploit the aforementioned advantages of having access to IPD. The objective of this chapter is to explore how researchers might simultaneously address the challenges encountered in this scenario: on the one hand, using methods that exploit the benefits of having access to IPD, whilst on the

other, ensuring that parameter estimates make comprehensive use of the available data. An empirical case study is conducted that investigates the use of a multistage Bayesian approach to evidence synthesis for the estimation of HRQoL inputs.

7.2 Background

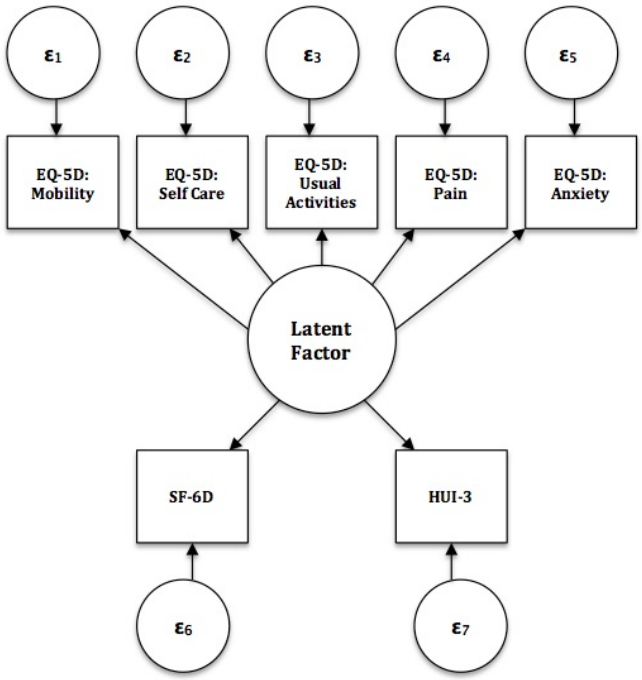
The aforementioned challenges associated with the synthesis of evidence comprised of both IPD and AD can be circumvented through the application of Bayes' theorem. There are two steps involved in this approach, which is illustrated in Equation 7.1: (1) the evidence in the IPD format is used to estimate the unknown HRQoL parameter estimates, θ, which are subsequently used as informed priors; (2) the evidence in the AD format is used in the likelihood function, $L(\theta; y)$, to obtain an updated posterior distribution, $p(\theta|y)$, for the HRQoL parameter estimates. The main strength of this approach is that it allows researchers to exploit the SEM techniques capable of handling item-level responses and to combine the output with evidence in the AD format.

$$p(\theta|y) = L(\theta; y) \cdot p(\theta) \tag{7.1}$$

Whilst the Bayesian approach might allow researchers to exploit item-level responses for the synthesis of IPD, it does not get away from the fact that the subsequent synthesis of evidence at the AD level involves HRQoL index values. However, any concerns regarding the plausibility of the LISREL method can be avoided if researchers have access to a dataset containing all of the PRO measures found in evidence available in the AD format. The availability of IPD containing all such measures enables the prediction of the reference case PRO measure at the item level using evidence from alternative measures in the AD format, thus ensuring the prediction of feasible HRQoL index values. As such, an SEM model is specified to include a mixture of both categorical and continuous indicator variables.

An illustrative example is shown in Figure 7.1 depicting a SEM incorporating information from three PRO measures. Two of the measures are represented in terms of their index values (SF-6D and HUI-3), and thus captured as continuous variables. The reference case measurement, the EQ-5D, is represented in terms of the item-level responses to its descriptive system (i.e. categorical variables).

Figure 7.1: Graphical Representation of a Mixed Outcome SEM Model



7.3 Methods

The case study in this chapter seeks to combine evidence from the motivating examples used in Chapters 5 and 6 for the estimation of HRQoL input parameters in the RITA-3 cost-effectiveness model. The primary aim of the case study is to demonstrate how Bayesian methods can help to facilitate the synthesis of heterogeneous HRQoL evidence in both IPD and AD formats, via SEM techniques. Two alternative techniques for combining AD will be compared: one using the LISREL method (see Chapter 5) to map between outcome measures and another using a novel mixed outcomes SEM approach. In addition, parameter estimates obtained using the two methods for combining IPD and AD evidence will be compared with those derived using the EQ5D IPD alone (i.e. the methods observed in the previous chapter). Once again, the parameter estimates are fed into the RITA-3 model and cost-effectiveness estimates are obtained for each of the alternative methods.

7.3.1 *HRQoL Evidence*

PRO evidence in the IPD format was obtained from four studies: the National Health Measurement Survey (NHMS), the Medical Expenditure Panel Survey (MEPS), the Welsh Health Survey (WHS) and the Randomized Intervention Trial of unstable Angina (RITA-3) trial (Fryback et al., 2007; Kim et al., 2005; NatCen Social Research, 2013; Sullivan et al., 2011). Further details describing these datasets can be found in chapter 6. Evidence

in the AD format was extracted from an existing meta-analysis that assessed the effects of diabetes and related complications upon HRQoL (Lung et al., 2011). Further details describing this data can be found in Chapter 5.

7.3.2 *Health Economic Model*

The health economic model employed is the same as that used in Chapter 5, where further details describing this model can be found.

7.3.3 *Estimation of Parameter Inputs*

As with the previous two chapters, the aim of this empirical exercise is to estimate HRQoL values associated with two defined health states: one for patients with diabetes who have yet to experience a myocardial infarction and another for patients with diabetes who have experienced a myocardial infarction. A "reference case" approach is conducted to estimate HRQoL parameter inputs solely using the EQ-5D evidence in both IPD and AD formats (Method 7.1). This involves aggregating the IPD evidence and then synthesising it in combination with the AD evidence using Model 5.1.[1]

An alternative method is considered for the synthesis of evidence selected as part of the "all-inclusive" approach (i.e. including evidence beyond the reference case measurement). This method (Method 7.2) conducts the mapping and evidence synthsis procedures separately. HRQoL evidence in the IPD format is mapped onto the EQ-5D scale using Model 6.2 from Chapter 6. A regression analysis is then implemented using all of the IPD observations available to obtain two parameter estimates, an intercept and a coefficient capturing the impact of an MI on HRQoL.[2] The parameter estimates obtained from the IPD analysis will then be combined with evidence in the AD format through the specification of a Bayesian meta-regression model. However, before this task can be undertaken all of the evidence in the AD format has to be mapped onto the scale of the reference case measurement.

The availability of IPD linking all of the relevant outcomes means that we are no longer restricted to using the LISREL method as in Chapter 5. A mixed outcomes SEM approach, illustrated in Figure 7.1, is fitted to the NHMS dataset and the resulting output used to

[1] Note that sample statistics obtained from the IPD data are split according to the patient health state (i.e. no MI versus post MI) and study.

[2] Although the case study in Chapter 5 also included covariates for age and gender, these are removed for illustrative purposes to show a scenario with corresponding parameters across the evidence formats.

convert SF-6D and HUI-3 scores in the AD format onto the latent factor (CF) scale. This model is represented in Equations 7.2 - 7.6.

$$\Pr(EQ5D_{ij} = 1) = \Phi(-\tau_{1,j} + \lambda_j \cdot f_i) \tag{7.2}$$

$$\Pr(EQ5D_{ij} = 2) = \Phi(-\tau_{2,j} + \lambda_j \cdot f_i) - \Phi(-\tau_{1,j} + \lambda_j \cdot f_i) \tag{7.3}$$

$$\Pr(EQ5D_{ij} = 3) = 1 - \Phi(-\tau_{2,j} + \lambda_j \cdot f_i) \tag{7.4}$$

$$SF6D_i = \alpha_{SF6D} - \lambda_{SF6D} \cdot f_i \tag{7.5}$$

$$HUI3_i = \alpha_{HUI3} - \lambda_{HUI3} \cdot f_i \tag{7.6}$$

Where $EQ - 5D_{ij}$ represents individual i's response to the jth item of the EQ-5D, f_i is individual i's latent factor score and λ_j is the factor loading for EQ-5D item j. The $\tau_{k,j}$ terms are thresholds indicating the point on the latent factor response scale where individuals switch their item response. Φ is the standard cumulative distribution function. The $SF6D_i$ and $HUI3_i$ terms represents individual i's valuations for the SF-6D and HUI-3 instruments. These measures have intercepts represented by the terms α_{SF6D} and α_{HUI3}, and factor loadings represented by the terms λ_{SF6D} and λ_{HUI3}.

EQ-5D responses first need to be predicted before values can be obtained. A simulation exercise is implemented to this extent that predicts EQ-5D responses using the latent factor statistics in combination with the relevant factor loadings and threshold values from Equations 7.2 - 7.4. Once a sufficiently large number of simulations have been run, the UK EQ-5D population value set is applied to the predicted responses and sample statistics derived for the resulting values. The Bayesian meta-regression approach used to synthesise evidence in the AD format is the same as Model 5.2 (see Equations 5.1 - 5.5) with exception of the priors specified. Whilst Model 5.2 used plausible but uninformed priors (Equations 5.3 and 5.4) for the parameter estimates of interest, the priors used in this chapter are informed by the IPD analysis. A different assumption is needed regarding the distribution of the constant term, μ, given that the uniform distribution is no longer applicable. Instead, a beta distribution is assumed and, as such, the parameter estimates

obtained from the synthesis of IPD evidence need to be modified via methods of moments calculations (see Briggs et al. (2006)).[3]

7.3.4 *Statistical Software*

The meta-analytic models are estimated using the JAGS software, which is run through the R program using the *R2jags* package (Su and Yajima, 2012). Each model is run with three Markov chains over total of 10,000 iterations, the first 1,000 of which are discarded as the burn-in period. For each method, the posterior distribution is fed directly into the health economic model for the probabilistic sensitivity analysis. The factor loadings are estimated using the *Lavaan* package in R (Rosseel, 2012).

7.4 Results

7.4.1 *Parameter Estimates*

Table 7.1 shows the estimated HRQoL parameter inputs for the cost-effectiveness model. This shows that there are substantial differences between the estimates derived using heterogeneous outcome measures (Model 7.2) compared to those that solely relied upon the reference case instrument (Model 7.1). The forest plot in Figure 7.2 provides a graphical representation of the estimated EQ-5D distributions for all of the patient profiles defined in the RITA-3 model, which were obtained in the probabilistic sensitivity analysis. The parameters derived by taking an "all-inclusive" approach to evidence synthesis exhibit significantly reduced uncertainty.

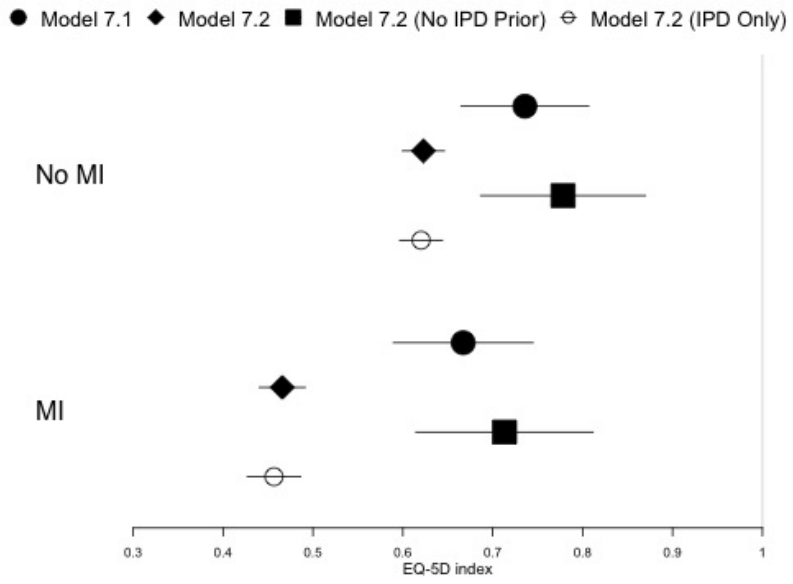
Table 7.1: Mean HRQoL Parameter Inputs

	No Previous MI	Post MI
Model 6.1	0.736	0.667
Model 6.2	0.623	0.466
Model 6.2 (No IPD Prior)	0.778	0.713
Model 6.2 (IPD Only)	0.620	0.457

[3]Note that there is still a discrepancy between the EQ-5D and the beta distribution in terms of their ranges; transformations are required to account for these differences in scale.

The evidence used in Model 7.2 has been split up according to the format (i.e. IPD and AD) in an effort to understand the parameter estimates. The estimates labelled 'Model 7.2 (No IPD Prior)' show the results form an additional scenario analysis performed using Model 7.2 but with the priors set to be the same as those in Model 5.1. The estimates labelled 'Model 7.2 (IPD only)' show the parameter estimates obtained using the IPD only. These results clearly demonstrate that the IPD evidence is the main driver behind the results. This can be explained by the fact that the parameter estimates derived from the IPD exhibit far greater precision when compared to those derived from the AD.

Figure 7.2: HRQoL Parameter Estimates



7.4.2 *Cost-Effectiveness Results*

HRQoL parameter inputs obtained via Models 7.1 and 7.2 were utilized in the RITA-3 model in order to explore the impact of the methodological differences upon the cost-effectiveness results. Tables 7.2 – 7.3 show the implications of using the different methods in terms of the impact upon the expected cost-effectiveness results. In addition, probabilistic sensitivity analyses were run, the results of which are presented as CEACs in Figure 7.3. This information is also presented in terms of error probabilities in Table 7.4, assuming threshold values of £20,000 per QALY and £30,000 per QALY respectively. These results present the probability of making an incorrect decision, for a given patient, on the basis of the expected cost-effectiveness results. Finally, the EVPI was estimated for each method and curves presenting this information can be found in Figure 7.4.

Table 7.2: Cost-Effectiveness Results Using Model 7.1 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£36,940	£32,188	£4,752
QALYs (discounted)	12.36	12.24	0.12
Cost-per-QALY	-	-	£39,495

Table 7.3: Cost-Effectiveness Results Using Model 7.2 Parameter Estimates

	Treatment	Comparator	Incremental Difference
Costs (discounted)	£36,940	£32,188	£4,752
QALYs (discounted)	10.32	10.21	0.11
Cost-per-QALY	-	-	£42,549

Table 7.4: Error Probabilities

	Threshold = £20K	Threshold = £30K
Model 7.1	0.006	0.139
Model 7.2	0.005	0.098

Figure 7.3 shows that the uncertainty surrounding the decision in question is reduced when the "all-inclusive" approach to evidence selection is employed as opposed to the "reference case" approach. This is because the probability of the early intervention being cost-effective is reduced, i.e. we can be more confident that it is not cost-effective. Figure 7.4 shows that, over a threshold range between £20,000 and £30,000 per QALY, the cost associated with the decision uncertainty is greater for the "reference case" approach compared to the "all-inclusive" approach.

Figure 7.3: Cost Effectiveness Acceptability Curves

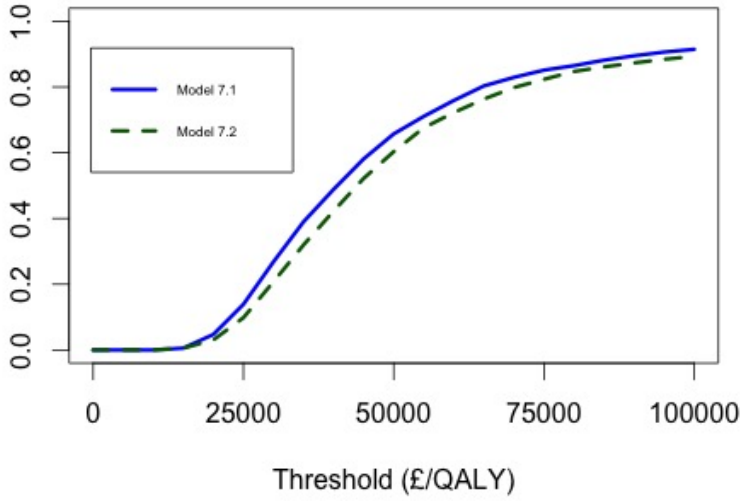
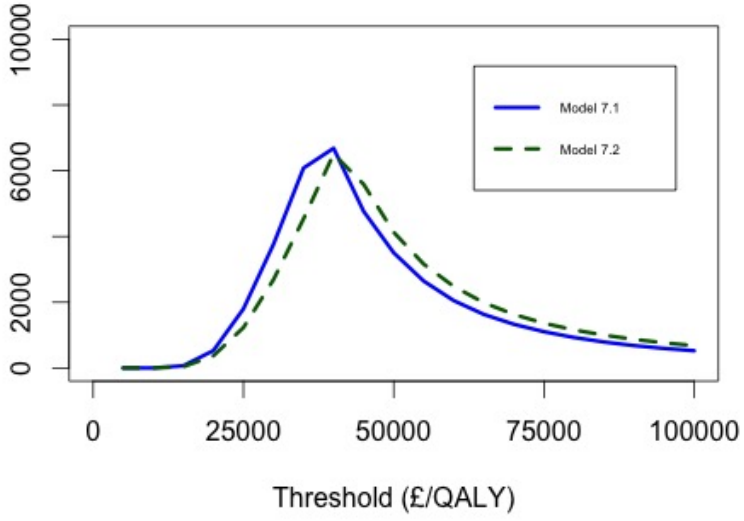


Figure 7.4: Expected Value of Perfect Information



7.5 Discussion

The primary objective of this chapter was to explore methods for the synthesis of heterogeneous HRQoL evidence in both IPD and AD formats, via SEM techniques. The main challenge in this task is finding a means to exploit the benefits of having access to item-level responses in the IPD, whilst also ensuring that the AD evidence is used to inform parameter estimates. This chapter proposes the combination of evidence in multiple steps through a process of Bayesian updating and an empirical case study was also conducted to explore the use of this method for the estimation of HRQoL inputs. The case study showed this multi-stage method to be successful in making optimal use of the available evidence and found that parameter uncertainty was dramatically reduced when a combination of IPD and AD evidence was used compared to IPD alone.

In addition to the multi-stage approach to parameter estimation, the case study in this chapter also illustrated how IPD can be used for the prediction of item-level responses on the reference case measurement using evidence from alternative measures in the AD format. This method is preferable for mapping from AD given that it ensures the prediction of plausible item responses. Given the concerns regarding use of LISREL methods in Chapter 4, this exercise illustrates the benefits of having access to IPD in terms of the opportunities it affords researchers seeking to synthesise heterogeneous AD evidence. The methods considered in this chapter are particularly relevant given that a scenario involving some combination of IPD and AD probably reflects that most plausible situation in practice. This is partly due to the fact that implementing SEM methods for the synthesis of HRQoL evidence is likely to rely upon the availability of IPD in order to estimate factor loadings (see chapter 5 for further details). However, researchers are unlikely to be able to obtain all relevant evidence in IPD format and, thus, will have to make do with results in AD format for many studies.

The case study in this chapter found clear evidence that parameter uncertainty was reduced using the 'all-inclusive' approach to evidence synthesis (Model 7.2) compared to the 'reference case' approach (Model 7.1). Furthermore, HRQoL estimates obtained with the 'all-inclusive' approach were considerably reduced compared to the 'reference case' approach'. However, these differences would appear to be largely explained by the additional IPD employed in the 'all-inclusive' approach, given the similar parameter estimates obtained from Model 7.2 with and without the AD evidence. This raises doubts over the consistency of the results from the non-reference case IPD data compared to the remaining evidence, which begs the question of whether or not they should be included in the synthe-

sis. It may be reasonable to distinguish the non-reference case IPD data, taken from the WHS and MEPS studies, from the remaining studies given that the patient populations were selected on the basis of their self-reported comorbidities. As such, these patients may not be directly comparable to patients identified by clinicians.

It is important to acknowledge that the case study in this chapter is subject to a number of limitations. Unlike the previous chapter, the model specified for the meta-regression in this chapter did not include covariates for age and gender. Although the age and gender covariates were removed in this chapter for illustrative purposes, the findings from Chapter 6 indicated that these are important predictors. Given that the decision problem of interest relates to a male patient population aged 52 years old, a failure to control for these variables may distort the estimated cost-effectiveness results. There are likely to be many cases, such as this one, where it would be more appropriate to disregard the AD evidence for the sake of exploiting additional covariates from the IPD.

Even where covariates are found to be important predictors of cost-effectiveness, this does not necessarily imply that the results should be used to make stratified health care decisions. This is because there may be transaction costs associated with the stratification of patients that outweigh the benefits accrued from having controlled for additional covariates (Basu and Meltzer, 2007; Espinoza et al., 2014). Espinoza and colleagues developed a framework to estimate the value of heterogeneity in order to determine the optimal level of stratification (Espinoza et al., 2014). This framework could potentially be used to determine whether it would be better to use IPD evidence alone or in combination with AD evidence depending upon the value of heterogeneity. Note that this concept is not unique to HRQoL parameters – the same idea applies to other parameters including clinical evidence, resource use and costs.

Another issue regarding the inclusion of covariates is that of omitted variable bias. The validity of the approach in this chapter, as with any model, relies upon an assumption that there are no confounding variables that could give rise to biased parameter estimates. Researchers will often find that the existing studies from which their AD evidence was obtained did not analyse the raw data using a method in-keeping with the needs of their decision model, e.g. controlling for variables such as age or gender. There are several options to consider pursuing in this type of situation. First, one could consider extending the meta-regression to control for additional covariates although there is an additional risk of obtaining biased estimates when exploring statistical relationships between variables at the aggregate level (Piantadosi et al., 1988). Instead, one could employ a strict study selection strategy and disregard such evidence; however, this could prove to be inappropriate

if the IPD is subject to some alternative form of bias.

A third possibility would be to employ an evidence discounting method (Lunn et al., 2012). This method allows researchers to discount biased evidence – specifically, where there are concerns about bias either in terms of relevance or methodological rigor – such that it does not carry a full weighting compared to evidence assumed to be free of bias. However, this approach has yet to be explored in the context of HRQoL parameter estimation and it is unclear as to how one would go about quantifying these biases. Further research is required for this method to be considered for use in practice.

Although the specification of Model 7.2 might be preferable to the LISREL method in terms of its ability to predict plausible responses for the reference case measurement, this approach is still subject to flaws regarding the specification of the remaining outcome measures. Model 7.2 assumes that the SF-6D and HUI-3 are both continuous variables, and are linearly related to the latent factor. Given that both instruments exhibit a bounded range of plausible values and non-normal error terms, the assumptions necessary for this approach to be valid are likely to be violated. Further research is necessary to develop novel methods for handling this issue. However, it should be noted that – as with the empirical applications in Chapters 5 and 6 – there is additional structural uncertainty introduced with the 'all-inclusive' approach to evidence synthesis given that it requires a modelling approach to incorporate multiple HRQoL outcomes.

Chapter 8

Discussion and Conclusions

"...and all the pieces matter"

Lester Freamon, The Wire

8.1 Summary of the thesis

One of the main tenets of the decision analytical modeling approach to CEA is that researchers should aim to make comprehensive use of all relevant evidence to minimize the potential risk of obtaining misleading results. In this regard, there has been a large body of methodological research devoted to the statistical methods used to synthesizing clinical evidence in CEA (Sutton et al., 2012). However, this issue has been largely overlooked in the context of HRQoL evidence. As such, the main purpose of this thesis was to contribute to the understanding of statistical methods for incorporating HRQoL evidence within the decision-analytic modelling framework. The thesis addressed this overall aim by exploring four research questions, set out in Chapter 1.

The first research question sought to evaluate the current state of practice with regards to the use of statistical methods to incorporate HRQoL data in applied cost-effectiveness studies. A review of NICE technology appraisals was conducted to evaluate the use of statistical methods to incorporate HRQoL evidence in applied cost-effectiveness studies (Chapter 2). The results of the review found that the following statistical procedures were used on an irregular basis: pooling techniques for the synthesis of evidence from multiple studies; mapping techniques for the prediction of reference case HRQoL values using alternative outcome measures; adjustment techniques used to combine HRQoL values for comorbidities. These findings suggest that there are fundamental inconsistencies regarding the utilization of HRQoL evidence for the purposes of informing technology adoption

decisions. In light of these findings, a decision was made to focus research efforts upon the identification of statistical methods capable of utilizing HRQoL evidence in both a comprehensive and consistent manner.

The second research question of the thesis sought to evaluate the guidance currently available with respect to the use of statistical methods for the synthesis of HRQoL evidence in CEA. A review of the policy guidance and published literature was conducted to evaluate the methodological guidance currently available (Chapter 3). Recommendations from policy-makers regarding the synthesis of HRQoL evidence were found to be lacking. Chapter 3 concluded that taking a vague stance on this issue might actually risk promoting future inconsistencies in the utilisation of HRQoL evidence. This conclusion has important implications for policy makers as it suggests that more prescriptive methodological guidance may be required.

A frequently observed issue in the published literature was the predicament of how to deal with between-study heterogeneity (i.e. a lack of comparability amongst the alternative instruments for measuring and valuing HRQoL). Some studies attempted to get around this problem by restricting syntheses to only include evidence for a reference case instrument to ensure that the evidence being incorporated is comparable. However, this approach may provide an incomplete representation of HRQoL effects or, in some cases, no evidence at all. One of the studies identified in the review in Chapter 3 proposed the use of mapping techniques as a means to resolve the issue of between-instrument heterogeneity (Peasgood and Brazier, 2015). Chapter 3 concluded that incorporating mapping procedures within a broader evidence synthesis framework could potentially pave the way towards a more appropriate utilisation of HRQoL evidence for CEA. In particular, it was felt that this suggested approach would avoid researchers having to make a compromise between thoroughness and comparability in the selection of HRQoL evidence.

Chapter 4 explored the potential role of mapping techniques in combination with evidence synthesis techniques. It began by reviewing some of the statistical issues associated with the development of mapping algorithms. SEM techniques were identified as exhibiting a number of important characteristics with regards to the analysis of HRQoL data, most notably their ability to account for measurement error in applications involving multiple outcomes (Lu et al., 2013). Next, a detailed outline of the SEM framework was provided with particular emphasis upon the opportunities and challenges involved in the analysis of preference-based measures of HRQoL. This description constitutes a valuable contribution on the grounds that it provides the health economic evaluation community with an introduction to the use of the SEM framework, a methodology not traditionally associated

with this discipline.

Despite the early promise exhibited by the SEM methodology however, it was felt that further empirical research should be undertaken before any recommendation with regards to their use in applied research could be made. It was decided that this should include the following: (i) validation of this approach across the variety of plausible scenarios that researchers might encounter, (ii) comparison of the methods against those employed in the present circumstances, and (iii) demonstration in the context of a cost-effectiveness model involving HRQoL parameters associated with defined health states and events.

The third research question of the thesis sought to determine the plausibility of the SEM approach serving as a generalised framework for the synthesis of heterogeneous HRQoL evidence in CEA. A series of empirical case studies were conducted to explore the methodological challenges associated with the use of these methods, and to compare parameter estimates obtained via the synthesis of heterogeneous HRQoL evidence with SEM techniques against those obtained via synthesis of homogeneous HRQoL evidence with standard synthesis techniques. The latter method was selected as being representative of current practice based upon an assumption that cost-effectiveness evidence developed for the purposes of policy decision making should strive to incorporate HRQoL effects captured using some reference case measurement.

The SEM approach to evidence synthesis was employed in three separate scenarios that differed in terms of the format of the available evidence. Each of these scenarios revolved around an existing cost-effectiveness model comparing an early surgical intervention to medical management for the treatment of acute coronary syndrome in a patient subgroup with diabetes. This case study was selected for two reasons: (i) there were a range of freely-available HRQoL studies pertaining to this case study and these studies facilitated methodological investigations for each of the scenarios considered; (ii) the availability of an existing model meant that the competing methods could be compared in terms of their impact upon the cost-effectiveness results. The case study stands out as the first application to date involving the use SEM methods to derive HRQoL parameter estimates for a cost-effectiveness model.

Chapter 5 focused upon a scenario where researchers, seeking to estimate HRQoL parameters for a CEA, are faced with an evidence base involving multiple sources of evidence in the AD format and composed of a variety of instruments. HRQoL evidence in the AD format was obtained from an existing meta-analytic study and two approaches were considered with regards to the selection of evidence: (i) a "reference case" approach, that

only used evidence from studies that collected the reference case measurement, and (ii) an "all-inclusive" approach, that incorporated a range of other outcome measures in addition to the reference case measurement. For the "reference case" scenario, a hierarchical meta-regression method was used to synthesise the available EQ-5D evidence. Two methods were proposed for the synthesis of evidence in the "all-inclusive" scenario, both of which utilized SEM techniques to map between the different outcome measures. One of the methods involved a two-step procedure, starting with the mapping of outcomes onto the same scale followed by the evidence synthesis. The other, so-called *integrated approach* conducted these tasks simultaneously.

The parameter estimates derived using the two-step synthesis approach exhibited confidence intervals that were noticeably reduced when compared to those obtained by synthesis of the reference case evidence. This finding was encouraging given that it would seem to validate the rationale for embracing the "all inclusive" approach to evidence synthesis; that is, by incorporating more relevant evidence, we increase the chances of gaining statistical power for the estimation of parameters (Higgins et al., 2009). However, it is important to acknowledge that other forms of uncertainty are introduced into the CEA when these methods are adopted. There is additional methodological uncertainty due to the lack of consensus regarding validity of using indirect estimates of HRQoL (i.e. mapped values). Furthermore, structural uncertainty is also introduced given that the validity of the statistical methods has been questioned in the context of HRQoL research (Alava et al., 2012; Basu and Manca, 2012).

Overall, decision uncertainty in the RITA-3 model did not vary greatly when different HRQoL parameter estimates, derived using the alternative methods considered in this chapter, were employed. A supplementary analysis was conducted to estimate EVPPI and this showed that the impact of HRQoL is modest in the RITA-3 model compared to other parameters. For this reason, Chapter 5 concluded that the choice of method for the estimation of HRQoL parameters would have been unlikely to impact upon a policy reimbursement decision for this specific application. However, it was also recognised that the role of HRQoL parameters within a given decision model is case study dependent. As such, this finding would not constitute grounds for making a recommendation about the implications of the SEM methodology for policy-making more broadly.

In Chapter 6, a scenario involving multiple sources of heterogeneous HRQoL evidence in the IPD format was considered. As well as comparing an "all-inclusive" approach to evidence synthesis to a "reference case" approach, the investigations in this chapter focused on the potential benefits of adopting synthesis techniques capable of exploiting item-level

responses. The latter point was considered to be particularly important given that this pertains to all PRO measures involving likert-type questionnaires.

An empirical case study was conducted using evidence in the IPD format from several freely available studies, in addition to data from a clinical trial associated with the cost-effectiveness model. Two methods for synthesising evidence were proposed using SEM methods, both of which specified the HRQoL measures in terms of item-responses using polychoric methods. Once again, a distinction was made between the "reference case" approach to evidence synthesis and an "all-inclusive" approach. As with Chapter 5, two methods were proposed for the synthesis of evidence in the "all-inclusive" scenario, both of which utilised SEM techniques to map between the different outcome measures.

The first of the proposed SEM methods implemented mapping and synthesis procedures in separate stages, while the second method combined these procedures into a single modelling approach using an FIML estimation method. Despite offering a more computationally efficient approach, the latter method could only be implemented in one software package, Mplus (Muthén and Muthén, 2010). It was felt that the lack of alternative software options represents a significant obstacle for the uptake of this method. As such, it is hoped that this capability will be extended to other software packages in the future.

The HRQoL parameter estimates derived using the SEM methods were noticeably lower than those obtained using the reference case data. The discrepancies appear to have been caused by differences in the estimated impact of age upon HRQoL. One possible explanation for this finding might be that the additional evidence incorporated in the "all-inclusive" scenario was able to offset the spurious covariate associations between age and HRQoL in the "reference case" scenario. The case study in Chapter 6 did not show any clear impact of incorporating a broader range of HRQoL measures in terms of either parameter or decision uncertainty. As with the scenario considered in Chapter 5, parameter estimation with SEM methods introduces both methodological and structural uncertainty.

Chapter 6 concluded that there are substantial benefits to be gained from synthesising heterogeneous HRQoL measures using item-level responses rather than summary scores and/or index values. Unlike the LISREL methods considered in the Chapter 5, item-level SEM methods ensure the prediction of plausible HRQoL responses. The empirical case study conducted in Chapter 6 found that the item-level approach had a superior predictive performance when compared to the LISREL approach. IPD evidence brings additional benefits including the ability to produce HRQoL parameter estimates stratified according

to patient characteristics, which is needed to explore heterogeneity in cost-effectiveness estimates. Furthermore, IPD allows researchers to control for covariate imbalances in the data that would otherwise bias the parameter estimates of interest.

Chapter 7 examined the application of SEM methods in a scenario involving a combination of evidence in the IPD and AD formats. This chapter sought to develop an SEM-based approach to evidence synthesis capable of both exploiting the benefits of having access item-level responses in the IPD, whilst also ensuring that the AD evidence is used to inform parameter estimates. Of all of the scenarios considered in the thesis, this one was considered to be particularly important given that it probably reflects the most plausible situation encountered in practice. This is partly due to the fact that implementing SEM methods for the synthesis of HRQoL evidence is likely to rely upon the availability of IPD in order to estimate factor loadings (see Chapter 5 for further details). However, researchers are unlikely to be able to obtain all relevant evidence in IPD format and, thus, will have to make do with results in AD format for many studies.

A method was proposed that involved combining evidence in multiple stages via a Bayesian updating process. Heterogeneous IPD evidence would be synthesised first using either of one the item-level SEM techniques proposed in Chapter 6. The resulting parameter estimates from this first step would then be used as informed priors in a subsequent synthesis of the heterogeneous AD evidence. Assuming that there is IPD linking all of the relevant outcome measures in the AD format, then the synthesis of evidence in the AD format is no longer restricted to the use of the LISREL methods employed in Chapter 5. Instead, Chapter 7 proposed that the IPD evidence could be used to predict item-level responses on the reference case measurement using evidence from alternative measures in the AD format. This method is preferable to the LISREL approach given that it guarantees the prediction of plausible HRQoL responses (unlike the LISREL approach).

The proposed method was employed to estimate HRQoL parameters in an empirical case study using the AD evidence from Chapter 5 and the IPD evidence from Chapter 6. These estimates were then compared to parameter estimates obtained via a "reference case" approach to evidence synthesis. The case study showed this multi-stage method to be successful in making optimal use of the available evidence and found that parameter uncertainty was dramatically reduced when a combination of IPD and AD evidence was used compared to IPD alone. However, there were doubts raised over the comparability of the IPD and AD evidence given that there were large differences observed in the HRQoL estimates. Chapter 7 suggested that these differences might have been explained by differences in the way that the patient populations were defined. This finding illustrates the

importance of considering the consistency of the evidence when combining the results from multiple studies.

8.1.1 *Original Contributions*

Chapter 2 of the thesis provides three important contributions. Firstly, there is no published article, to the author's knowledge, to have explored the consistency with which statistical methods are employed in applied CEA research to incorporate HRQoL evidence. To this end, Chapter 2 can be considered as being a contribution to the field. Secondly, the findings in Chapter 2 indicate that there are fundamental inconsistencies regarding the utilisation of HRQoL evidence for the purposes of informing technology adoption decisions. This has important implications for policy-making given that these inconsistencies undermine the comparability of evidence used to inform different reimbursement decisions. Finally, the methodological inconsistencies indicate that there may be deficiencies in the published literature and policy guidance pertaining to the statistical methods employed in applied CEA research to synthesise HRQoL evidence.

An important contribution of the research in this thesis is the fact that the investigations reflect a range of scenarios that might be encountered in practice. Overall, the empirical research conducted in this thesis has shown that the "all-inclusive" approach to evidence selection, via SEM methods, can potentially improve precision in the estimation of HRQoL parameters for CEA compared to an approach relying on evidence collected using some prespecified reference case measurement. Relevant code has been provided with instructions on how to implement the methods in each scenario in open source software (Team, 2014). It is hoped that this will facilitate the use of these methods and encourage future research efforts in this area.

Chapter 7 represents the first known study to have considered the synthesis of heterogeneous HRQoL evidence in both IPD and AD formats. This chapter presents a novel multi-stage method for incorporating HRQoL evidence in a way that exploits the benefits of having access to item-level data, whilst also incorporating AD evidence. Furthermore, it illustrates how researchers can exploit available IPD evidence to predict item-level responses for a reference case measurement using evidence from alternative measures in the AD format. Previous research exploring the use of SEM methods for the analysis of HRQoL evidence has relied on LISREL methods (Lu et al., 2013, 2014), despite the fact that linear modeling techniques such as these have been shown to be inappropriate for the analysis of HRQoL preference values (Alava et al., 2012; Basu and Manca, 2012). The item-level approach is preferable for mapping from AD given that it ensures the prediction

of plausible item responses.

Although a previous study was found to have employed SEM techniques for the synthesis of HRQoL evidence in the AD format, this did not involve HRQoL preference values (Lu et al., 2013). Chapter 5 is the first study, to the author's knowledge, to have employed SEM techniques for the synthesis for the synthesis of HRQoL preference values. There is only one existing study known to have implemented item-level SEM methods for the analysis of HRQoL evidence (Gibbons et al., 2014). However, this study did not consider some of the challenges likely to be encountered by researchers implementing these methods in conjunction with a cost-effectiveness model. Another original aspect of the research in this thesis is that it considers the implementation these methods in the context of an associated cost-effectiveness model, principally of how to account for the uncertainty surrounding the parameter estimates.

8.1.2 *Limitations*

Although the investigations in this thesis provide a number of valuable contributions to the field of health economic evaluation, it is important to acknowledge the limitations of the research. The limitations chiefly pertain to the difficulties in forming methodological recommendations on the basis of empirical findings that are case study specific. The benefits of adopting a more comprehensive approach to evidence synthesis, via SEM methods, are likely to differ depending upon the condition under evaluation. To this end, a key priority for future research will be to undertake simulation exercises to test the implications of using competing methods across a variety of scenarios. The benefit of simulation exercises is that the researcher has full control over all of the underlying parameters feeding into the model to understand the mechanisms at work.

Another limitation of the case studies in this thesis is that they are solely concerned with a limited number of preference-based measures of HRQoL. In theory, the SEM methods for evidence synthesis would allow researchers to draw upon the wealth of non-preference-based HRQoL measures available. As well as synthesising a broader range of outcome measures, the SEM methodology is capable of estimating parameters with greater precision by incorporating disease-specific outcome measures. The latter point has important implications for the identification of evidence given that the range of relevant outcomes would vary according to the condition under evaluation. Further research is needed to establish how researchers would go about identifying the range of outcomes that might potentially be considered within a search strategy.

By only focusing upon applications involving generic measures of HRQoL, this thesis also overlooks the potential complexities involved in the specification of models that also incorporate disease-specific measures. The case studies in this thesis dealt purely with models assuming that HRQoL effects could be captured by a unitary construct. For certain disease areas, such as psoriatic arthritis (Kavanaugh et al., 2016), patients may experience several distinctive morbidity effects that need to be captured in a cost-effectiveness model. To this end, additional exemplars are required to explicitly demonstrate the modeling techniques that might be performed.

The cost-effectiveness study selected in this thesis was flawed for the purposes of exploring the implications of employing alternative methods for synthesising HRQoL evidence. Taken at face value, the findings would suggest that the methodological issues are unimportant from a policy-maker's perspective, given that there was no observable impact upon decision uncertainty. However, a supplementary analysis to estimate EVPPI results in the RITA-3 model showed that the impact of HRQoL is modest compared to other parameters. As such, it is unlikely that the choice of method for the estimation of HRQoL parameters would have had a substantial impact upon the decision uncertainty.

8.2 Recommendations

8.2.1 *Recommendations for researchers and decision makers*

Researchers are advised to consider using SEM techniques for the synthesis of evidence involving multiple heterogeneous outcome measures. Empirical research conducted in this thesis has shown that this "all-inclusive" approach to evidence selection can potentially improve precision in the estimation of HRQoL parameters for CEA compared to an approach relying on evidence collected using some pre-specified reference case measurement. Relevant code has been provided with instructions on how to implement the SEM-based synthesis techniques in open source software (Team, 2014). It is hoped that this will facilitate the use of these methods and encourage future research efforts in this area. It is vital that researchers take account of the uncertainty surrounding the factor loadings and threshold values. Otherwise, they might risk underestimating the uncertainty that is introduced by using indirect estimates of HRQoL.

Researchers should be aware that access to IPD containing all of the measurements of interest is likely to be a pre-requisite for the implementation of SEM-based methods. While these methods can also be employed using sample covariance data (as in Chapter 5), it is doubtful that this information can be obtained from published studies. Researchers are

advised to take advantage of having access to IPD wherever possible. The availability of IPD enables investigations into patient heterogeneity as well as allowing analyses to control for missing data or covariate imbalances. Significantly, HRQoL evidence in the IPD format also permits the implementation of SEM techniques that exploit item-level responses. The analysis of item responses, rather than index or summary scores, is preferable on the grounds that it ensures the prediction of plausible HRQoL values. Moreover, analyses involving item responses have been shown to exhibit lower levels of bias than those involving summary scores.

Decision makers are recommended to consider providing more prescriptive methodological guidance in relation to the methods employed to synthesise HRQoL evidence. Chapter 3 found the recommendations currently on offer to be lacking in this regard. In particular, decision makers are advised to consider recommending a more inclusive approach with regards to the synthesis of HRQoL evidence, i.e. consider the use of non-reference case HRQoL evidence, even where directly relevant, reference case evidence is available. Although concerns may exist in relation to the use of indirect estimates of HRQoL (i.e. mapped estimates), there are other examples in the field of health economic evaluation where indirect evidence is employed in evidence synthesis for parameter estimation, i.e. treatment comparisons (Sutton et al., 2012).

8.2.2 *Recommendations for future research*

Further research is needed to establish the precise set of circumstances in which the methods for synthesising heterogeneous HRQoL evidence via SEM techniques are likely to be most important. The most appropriate approach for handling investigations on this matter would be to undertake simulation exercises to test the implications of using competing methods across a variety of scenarios. This approach would permit the analyst to both specify and vary some of the key assumptions underlying the data including the population distribution of HRQoL responses/values for the parameters of interest and the measurement error associated with alternative HRQoL measures.

Although the case studies in this thesis dealt exclusively with generic, preference-based measures of HRQoL, it is anticipated that the SEM framework is sufficiently flexible for use in other scenarios. However, additional empirical research is needed to demonstrate the use of these methods in a wider range of scenarios, particularly those involving disease-specific measures of HRQoL. As previously mentioned, disease-specific measures sometimes capture distinctive morbidity effects with separate domains and this would bring additional complexities into the model specification process. Existing research from the field of

psychometrics would be well placed to inform these investigations given that there have been significant methodological advances to contend with the complex relationships often hypothesized in this field.

If policy makers were to endorse an "all-inclusive" approach to evidence synthesis, then this would have important implications for the methods used to identify HRQoL evidence. Further research is needed to establish how these considerations would be integrated within a search strategy. For instance, this could encompass a scoping process prior to the commencement of the literature search to identify all relevant outcome measures. For certain conditions, research may have even been conducted already to determine which patient-reported outcomes are appropriate for the measurement of HRQoL (Gibbons et al., 2014; Hadi et al., 2010; Mackintosh et al., 2009).

Chapter 6 identified a number of deficiencies in the statistical software considered regarding the implementation of SEM models for evidence synthesis involving IPD. Unfortunately, the methods employed could not incorporate hierarchical structures to reflect heterogeneity existing between different studies. This was considered to be a simplifying assumption and further research was recommended, particularly in relation to the use of a Bayesian multilevel approach (Lee, 2007). Another issue was the fact that the FIML estimation method proposed in Chapter 6 could only be implemented in a single software package (Mplus). As such, this constitutes a significant obstacle in the implementation of this methodology. Further developments are recommended to facilitate the implementation of this method in other software packages.

8.3 Conclusions

This thesis set out to explore the plausibility of the SEM approach serving as a generalised framework for the synthesis of heterogeneous HRQoL evidence in CEA. The research has demonstrated that this methodology can, theoretically, be implemented to synthesise evidence in a range of formats and has the potential to deliver more precise estimates of HRQoL. However, the research has also recognised that the fact that a scenario solely involving evidence in the AD format is unlikely to occur in practice given that covariance data are rarely available in the published literature. As such, a pre-requisite for the implementation of the SEM approach is that researchers have access to at least one source of IPD evidence capturing all of the HRQoL measures of interest. It is hoped that the findings of the thesis will encourage health economists to consider implementing the SEM approach in order to synthesise HRQoL evidence in a more comprehensive and transparent

manner. It is also hoped that this research will prompt further discussions amongst policy makers in regards to the methods for synthesising HRQoL evidence.

Appendix A

List of Studies Identified in the Published Literature

Table A.1: List of Studies Identified in the Published Literature

Year	Authors	Disease Area	Outcome Measures	Respondent Type	Statistical Methodology	Independent Variables
1999	Cheng	Deafness	HUI VAS QWB	Patients Values Public Values	Fixed-effects meta-analysis	N/A
2002	Tengs	HIV/AIDS	TTO SG Rating Scale QWB Judgement	Patient Values Non-Patient Values	Meta-regression using a hierarchical linear model	Disease Stage Outcome Measure Respondent Type Scale Boundaries
2003	Tengs	Stroke	TTO SG Rating Scale Judgement Other	Patient Values Public Values Expert Values	Meta-regression using a hierarchical linear model	Stroke Severity Outcome Measure Respondent Type Scale Boundaries
2007	Bremner	Prostate cancer	TTO SG Rating Scale QWB Judgement HUI	Vignettes Patient Values Public Values	Meta-regression using a linear mixed-effects model	Disease Stage Symptom Severity Respondent Type Outcome Measure
2008	Liem	End-stage renal disease	TTO SG EQ-5D EQ-VAS	Patient Values Public Values	Random-effects meta-analysis Separate analyses for different outcome measures Separate analyses for different treatment comparisons	N/A
2008	McLernon	Liver disease	TTO SG VAS EQ-5D TVAS HUI-2 HUI-3	Public Values Vignettes Patient Values	Meta-regression using a hierarchical model	Disease State Outcome Measure
2009	Peasgood	Osteoporosis-related conditions	EQ-5D	Public Values	Two alternative pooling approaches: One weighting values by the inverse of the variance Another weighting values by sample size	N/A
2010	Doth	Neuropathic pain	EQ-5D	Public Values	Random effects meta-analysis Separate analyses for different conditions	N/A

Year	Authors	Disease Area	Outcome Measures	Respondent Type	Statistical Methodology	Independent Variables
2010	Peasgood	Breast cancer	SG TTO VAS EQ-5D Other	Clinical Staff Public Values Vignettes	Metaregression with clustering Separate analyses for cancer stage	Condition-Specific States Age Outcome Measure Respondent Type Scale Boundaries
2010	Sturza	Lung Cancer	SG Judgement Direct Rating HALex AQOL EQ-5D TTO	Patient Values Expert Values Public Values	Meta-regression using a hierarchical model	Cancer Stage Cancer Type Scale Boundaries Respondent Type Outcome Measure
2011	Lung	Diabetes	EQ-5D TTO SG HUI-2 HUI-3 SF-6D	Not reported	Random effects meta-regression	Sample Size Age Males (%) Outcome Measures
2012	Wyld	Chronic kidney disease	TTO SG EQ-5D EQ-5D (mapped) 15D SF-6D	Public Values Patient Values	Random-effects meta-regression	Treatment Outcome Measure
2014	Djalalov	Colorectal Cancer	TTO EQ-5D HUI-3 SG VAS	Patient Values Public Values Vignettes	Meta-regression using a linear mixed-effects (LME) model	Disease Site Stage Time to/from Initial Care Outcome Measure Respondent Type Mode of Administration
2014	Mohiuddin	Unipolar depression	SG EQ-5D	Patients Values Public Values	Random-effects meta-analysis Separate analyses for different outcome measures Separate analyses for different severities	N/A
2014	Si	Osteoporosis-related conditions	EQ-5D VAS Others	Not reported	Meta-regression Separate analyses for different fracture types	Time after Fracture Age Outcome Measure Gender Fracture History

Appendix B

Chapter 5 Code

```
########################
# R Code ##############
# Model 4.1 ##########
########################
library(R2jags)
library(coda)
library(lattice)
library(R2WinBUGS)
library(rjags)
# Data Inputs ########
EQ5D #mean values
EQ5D.SE #standard errors
precision <- 1/(EQ5D.SE^2)
MI #dummy variable for previous MI
N #the number of values being synthesized
Nstudy #the number of studies involved
study #study reference number
data.inputs <- list("EQ5D","precision",
                    "MI","N","Nstudy","study")
# Model specification
model <- function(){
  for(i in 1:N){
    EQ5D[i] ~ dnorm(THETA[i],precision[i])
    THETA[i] <- theta[study[i]] + beta*MI[i]
```

```
  }
  for(z in 1:Nstudy){
    theta[z] ~ dnorm(mu, tau.z)
  }
  mu ~ dunif(-0.594,1)
  tau.z <- 1/(sig.z^2)
  sig.z ~ dunif(0.001, 0.2)
  beta ~ dnorm(0, 10)
}
# Parameters of interest
combined.params <- c("mu", "beta", "sig.z")
# Initial values
inits<-list(
  list(mu=c(0.8),beta=c(-0.05),sig.z=c(0.001)),
  list(mu=c(0.65),beta=c(-0.1),sig.z=c(0.01)),
  list(mu=c(0.5),beta=c(-0.2),sig.z=c(0.1)))
# Run Model
model.fit <- jags(data = data.inputs, inits = inits,
                  n.chains = 3, n.iter = 9000,
                  parameters.to.save = combined.params,
                  n.burnin = 1000, model.file = model)
```

```
######################
# R Code #############
# Model 4.2 ##########
# Step 1 #############
######################
library(lavaan) #for more info: http://lavaan.ugent.be/
# Data Inputs #######
N #Number of patients in the mapping study
EQ5D.mean #EQ-5D mean from the mapping study
SF6D.mean #SF-6D mean from the mapping study
HUI3.mean #HUI-3 mean from the mapping study
means <- c(EQ5D.mean,SF6D.mean,HUI3.mean)
COV.EQ5D.SF6D #Covariance term #1 from the mapping study
COV.EQ5D.HUI3 #Covariance term #2 from the mapping study
COV.SF6D.HUI3 #Covariance term #3 from the mapping study
EQ5D.var #EQ-5D variance from the mapping study
SF6D.var #SF-6D variance from the mapping study
HUI3.var #HUI-3 variance from the mapping study
# Formulate covariance matrix
lower <- '
EQ5D.var
COV.EQ5D.SF6D    SF6D.var
COV.EQ5D.HUI3    COV.EQ5D.SF6D   HUI3.var '
covariance.matrix <-
  getCov(lower, names = c("EQ5D","SF6D","HUI3"))
# Model Specification
model <- 'CF      =~ EQ5D + SF6D + HUI3'
# Run Model
fit <- sem(model,
           sample.cov = covariance.matrix,
           sample.mean = means,
           sample.nobs = N,
           meanstructure=TRUE)
param.estimates <- parameterEstimates(fit)
# Note step 2 uses the same model specification as Model 4.1
```

```
#######################
# R Code ##############
# Model 4.3 ##########
#######################
library(R2jags)
library(coda)
library(lattice)
library(R2WinBUGS)
library(rjags)
# Data Inputs ########
EQ5D.int #EQ5D intercept - derived using Model 4.2 - Step 1
SF6D.int #SF6D intercept - derived using Model 4.2 - Step 1
HUI3.int #HUI3 intercept - derived using Model 4.2 - Step 1
lambda2 #This is derived using Model 4.2 - Step 1
lambda2.SE #This is derived using Model 4.2 - Step 1
lambda2.precision <- 1/(lambda2.SE^2)
lambda3 #This is derived using Model 4.2 - Step 1
lambda3.SE #This is derived using Model 4.2 - Step 1
lambda3.precision <- 1/(lambda3.SE^2)
EQ5D.mean[i] #mean values; NA where missing
EQ5D.SE[i] #standard errors; NA where missing
SF6D.mean[i] #mean values; NA where missing
SF6D.SE[i] #standard errors; NA where missing
HUI3.mean[i] #mean values; NA where missing
HUI3.SE[i] #standard errors; NA where missing
MI #dummy variable for previous MI
N #the number of values being synthesized
Nstudy #the number of studies involved
study #study reference number
data.inputs <- list("lambda2","lambda2.SE",
                    "lambda3","lambda2.SE",
                    "EQ5D.int","SF6D.int","HUI3.int",
                    "EQ5D.mean[i]","EQ5D.SE[i]",
                    "SF6D.mean[i]","SF6D.SE[i]",
                    "HUI3.mean[i]","HUI3.SE[i]",
```

```
                    "MI","N","Nstudy","study")
# Model specification
model <- function(){
  FL2 ~ dnorm(lambda2,lambda2.precision)
  FL3 ~ dnorm(lambda3,lambda3.precision)
  for (i in 1:N){
    EQ5D.se[i] ~ dnorm(theta1.se[i],phi1.prec)
    SF6D.se[i] ~ dnorm(theta2.se[i],phi2.prec)
    HUI3.se[i] ~ dnorm(theta3.se[i],phi3.prec)
    var1[i] <- EQ5D.se[i]^2
    var2[i] <- SF6D.se[i]^2
    var3[i] <- HUI3.se[i]^2
    prec1[i] <- 1/(var1[i])
    prec2[i] <- 1/(var2[i])
    prec3[i] <- 1/(var3[i])
    EQ5D.mean[i] ~ dnorm(THETA1[i],prec1[i])
    SF6D.mean[i] ~ dnorm(THETA2[i],prec2[i])
    HUI3.mean[i] ~ dnorm(THETA3[i],prec3[i])

    THETA1[i] <- Int1 + LF[i]
    THETA2[i] <- Int2 + lambda2*LF[i]
    THETA3[i] <- Int3 + lambda3*LF[i]

    theta1.se[i] <- LF.se[i]
    theta2.se[i] <- FL2*LF.se[i]
    theta3.se[i] <- FL3*LF.se[i]

    LF[i] ~ dnorm(delta[i], phi.prec)
    delta[i] <- SLE[study[i]] + beta*HA[i]

    LF.se[i] ~ dunif(0.001,0.2)

  }

  for(z in 1:Nstudy){
```

```
    SLE[z] ~ dnorm(mu, tau.z)
  }

  phi1.se~dunif(0.001,0.2)
  phi2.se~dunif(0.001,0.2)
  phi3.se~dunif(0.001,0.2)

  phi1.prec <- 1/(phi1.se^2)
  phi2.prec <- 1/(phi2.se^2)
  phi3.prec <- 1/(phi3.se^2)

  phi.se ~ dunif(0.001,0.2)
  phi.prec <- 1/(phi.se^2)

  mu ~ dnorm(0,10)
  beta ~ dnorm(0,10)

  tau.z <- 1/(sig.z)
  sig.z ~ dunif(0.001,0.2)

}

# Parameters of interest
combined.params <- c("mu", "beta", "sig.z")

# Initial values
inits<-list(
  list(phi1.se=c(0.1),phi2.se=c(0.1),phi3.se=c(0.1),
       mu=c(-0.05),beta=c(-0.05),phi.se=c(0.1),
       sig.z=c(0.001)),
  list(phi1.se=c(0.01),phi2.se=c(0.01),phi3.se=c(0.01),
       mu=c(-0.1),beta=c(-0.1),phi.se=c(0.01),
       sig.z=c(0.01)),
  list(phi1.se=c(0.001),phi2.se=c(0.001),phi3.se=c(0.001),
       mu=c(-0.2),beta=c(-0.2),phi.se=c(0.001),
```

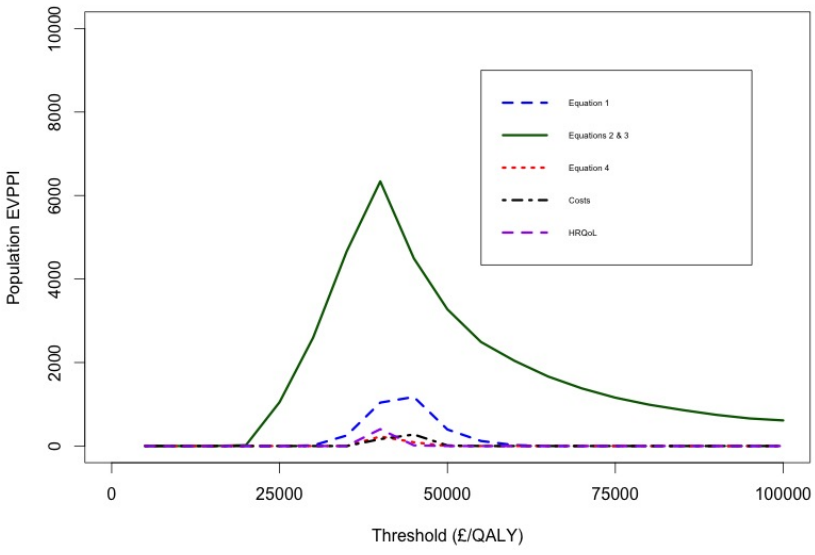
```
        sig.z=c(0.1)))

# Run Model
model.fit <- jags(data = data.inputs, inits = inits,
                  n.chains = 3, n.iter = 9000,
                  parameters.to.save = combined.params,
                  n.burnin = 1000, model.file = model)
```

Appendix C

Expected Value of Perfect Information for Parameters

The expected value of perfect information for parameters (EVPPI) is conceptually similar to the EVPI and represents the value of reducing uncertainty around specific parameters (Briggs et al., 2006). Traditionally, the rationale for estimating the EVPPI has been to identify the priority areas for further research. However, the EVPPI is also a useful method for exploring the extent to which different parameters contribute to the final outputs and impact upon the decision uncertainty (i.e. expected costs and QALYs). As an additional exercise, the EVPPI was estimated for each of the parameters in the RITA-3 model in order to evaluate the relative *contribution* of the HRQoL parameters. Figure C.1 illustrates the relationship between EVPPI for the model parameters and the cost-effectiveness threshold.

Figure C.1: Expected Value of Perfect Information for Parameters



These results show that additional HRQoL evidence would not be valuable given that it would be unlikely to change the decision. This provides important context for the findings in Chapter 5 (as well as the subsequent chapters) because it illustrates the fact that there is little scope for the methodological differences to have an impact upon the decision uncertainty for this particular case study. In conclusion, the impact of differing HRQoL synthesis techniques upon the expected cost-effectiveness is likely to be case study dependent (i.e. it will also depend upon other factors such as the model structure and evidence on the other parameters).

Appendix D

Chapter 6 Code

```
#######################
# R Code ##############
# Model 5.2 ###########
# Step 1 ##############
#######################
library(lavaan) #package required for SEM analysis
Model <-  #This is the bi-factor model specification
          #explaining SF-12 and EQ-5D responses
  "CF =~  SF36.Q1 + SF36.Q2 + SF36.Q3 +
          SF36.Q4 + SF36.Q5 + SF36.Q6 +
          SF36.Q7 + SF36.Q8 + SF36.Q9 +
          SF36.Q10 + SF36.Q11 + SF36.Q12 +
          EQ5D.Q1 + EQ5D.Q2 + EQ5D.Q3 +
          EQ5D.Q4 + EQ5D.Q5
  EQ =~   EQ5D.Q1 + EQ5D.Q2 + EQ5D.Q3 +
          EQ5D.Q4 + EQ5D.Q5
  SF =~   SF36.Q1 + SF36.Q2 + SF36.Q3 +
          SF36.Q4 + SF36.Q5 + SF36.Q6 +
          SF36.Q7 + SF36.Q8 + SF36.Q9 +
          SF36.Q10 + SF36.Q11 + SF36.Q12
CF ~ 0*EQ
CF ~ 0*SF
EQ ~ 0*SF
SF36.Q1 | t1 + t2 + t3 + t4
```

```
SF36.Q2 | t1 + t2
SF36.Q3 | t1 + t2
SF36.Q4 | t1 + t2 + t3 + t4
SF36.Q5 | t1 + t2 + t3 + t4
SF36.Q6 | t1 + t2 + t3 + t4
SF36.Q7 | t1 + t2 + t3 + t4
SF36.Q8 | t1 + t2 + t3 + t4
SF36.Q9 | t1 + t2 + t3 + t4
SF36.Q10 | t1 + t2 + t3 + t4
SF36.Q11 | t1 + t2 + t3 + t4
SF36.Q12 | t1 + t2 + t3 + t4
EQ5D.Q1 | t1 + t2
EQ5D.Q2 | t1 + t2
EQ5D.Q3 | t1 + t2
EQ5D.Q4 | t1 + t2
EQ5D.Q5 | t1 + t2 "
Dataset[, # define the dataset as being ordinal
      c("SF36.Q1","SF36.Q2","SF36.Q3",
        "SF36.Q4","SF36.Q5","SF36.Q6",
        "SF36.Q7","SF36.Q8","SF36.Q9",
        "SF36.Q10","SF36.Q11","SF36.Q12",
        "EQ5D.Q1","EQ5D.Q2","EQ5D.Q3",
        "EQ5D.Q4","EQ5D.Q5")]
       <- lapply(Dataset[,
      c("SF36.Q1","SF36.Q2","SF36.Q3",
        "SF36.Q4","SF36.Q5","SF36.Q6",
        "SF36.Q7","SF36.Q8","SF36.Q9",
        "SF36.Q10","SF36.Q11","SF36.Q12",
        "EQ5D.Q1","EQ5D.Q2","EQ5D.Q3",
        "EQ5D.Q4","EQ5D.Q5")],ordered)
fit <- #now fit the model
  sem(Model, data = NHMS, std.lv=TRUE,
       ordered=c("SF36.Q1","SF36.Q2",
                 "SF36.Q3","SF36.Q4",
                 "SF36.Q5","SF36.Q6",
```

```
                "SF36.Q7","SF36.Q8",
                "SF36.Q9","SF36.Q10",
                "SF36.Q11","SF36.Q12",
                "EQ5D.Q1","EQ5D.Q2",
                "EQ5D.Q3","EQ5D.Q4",
                "EQ5D.Q5"))

# Same results as that in Mplus
model.5.2.output <-
  parameterEstimates(fit) #capture the model output
parameter.covariance <-
  lavaan::vcov(fit) #covariance matrix for parameters
vars <- c(1:17,35:88) # select variables required
model.5.2.vcov <- parameter.covariance[vars,vars]
model.5.2.cholesky.decomposition <-
  chol(model.5.2.vcov) #cholesky decomposition
model.5.2.cholesky.decomposition <- #transpose cholesky decomp
  t(model.5.2.cholesky.decomposition)
model.5.2.specification <- parTable(fit)
```

```
#######################
# R Code ##############
# Model 5.2 ###########
# Step 2 ##############
#######################
library(lavaan)
# Analyse MEPS & WLS
# Need model with fixed parameters
# We set up a model template using
# parameters obtained in previous step
model.5.2.specification$ustart[1:17] <-
  model.5.2.output$est[1:17]
model.5.2.specification$ustart[38:91] <-
  model.5.2.output$est[38:91]
# Only interested in SF-12 rows
vars <- c(1:12,38:81,109,129)
model.5.2.specification.SF12 <-
  model.5.2.specification[vars,]
model.5.2.specification.SF12[57,8] <- NA
model.5.2.specification.SF12[58,8] <- NA
model.5.2.specification.SF12[58,3]<- "~"
model.5.2.specification.SF12[58,4]<- 1
# The next step transforms this into an
# specification that Lavaan can interpret
model.5.2.specification.SF12 <-
  paste0(model.5.2.specification.SF12$lhs,
         model.5.2.specification.SF12$op,
         model.5.2.specification.SF12$ustart,
         "*", model.5.1.specification.SF12$rhs,
         collapse = "\n")
# Now take the MEPS/WLS [Dataset2]
Dataset2[, # define the dataset as being ordinal
        c("SF36.Q1","SF36.Q2","SF36.Q3",
          "SF36.Q4","SF36.Q5","SF36.Q6",
          "SF36.Q7","SF36.Q8","SF36.Q9",
```

```
            "SF36.Q10","SF36.Q11","SF36.Q12")] <-
  lapply(Dataset2[,
         c("SF36.Q1","SF36.Q2","SF36.Q3",
           "SF36.Q4","SF36.Q5","SF36.Q6",
           "SF36.Q7","SF36.Q8","SF36.Q9",
           "SF36.Q10","SF36.Q11","SF36.Q12")],ordered)
fit <- #now fit the model
  sem(model.5.2.specification.SF12,
      data = Dataset2, std.lv=TRUE,
      ordered=c("SF36.Q1","SF36.Q2","SF36.Q3",
                "SF36.Q4","SF36.Q5","SF36.Q6",
                "SF36.Q7","SF36.Q8","SF36.Q9",
                "SF36.Q10","SF36.Q11","SF36.Q12"))
# This will give us a model with parameter
# estimates the same as those from the NHMS
# analysis. The next step involves predicting
# latent factor scores
Dataset2$latent <- predict(fit)
# With these latent factor scores we
# now must transform these into response
# probabilities for the EQ-5D items.
Dataset2$Pred.EQ5D.Q1.L1 <-
  pnorm(model.5.2.specification$ustart[82]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q1.L2 <-
  pnorm(model.5.2.specification$ustart[83]-Dataset2$latent) -
  pnorm(model.5.2.specification$ustart[82]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q1.L3 <- 1 -
  pnorm(model.5.2.specification$ustart[83]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q2.L1 <-
  pnorm(model.5.2.specification$ustart[84]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q2.L2 <-
  pnorm(model.5.2.specification$ustart[85]-Dataset2$latent) -
  pnorm(model.5.2.specification$ustart[84]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q2.L3 <- 1 -
  pnorm(model.5.2.specification$ustart[85]-Dataset2$latent)
```

```
Dataset2$Pred.EQ5D.Q3.L1 <-
  pnorm(model.5.2.specification$ustart[86]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q3.L2 <-
  pnorm(model.5.2.specification$ustart[87]-Dataset2$latent) -
  pnorm(model.5.2.specification$ustart[86]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q3.L3 <- 1 -
  pnorm(model.5.2.specification$ustart[87]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q4.L1 <-
  pnorm(model.5.2.specification$ustart[88]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q4.L2 <-
  pnorm(model.5.2.specification$ustart[89]-Dataset2$latent) -
  pnorm(model.5.2.specification$ustart[88]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q4.L3 <- 1 -
  pnorm(model.5.2.specification$ustart[89]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q5.L1 <-
  pnorm(model.5.2.specification$ustart[90]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q5.L2 <-
  pnorm(model.5.2.specification$ustart[91]-Dataset2$latent) -
  pnorm(model.5.2.specification$ustart[90]-Dataset2$latent)
Dataset2$Pred.EQ5D.Q5.L3 <- 1 -
  pnorm(model.5.2.specification$ustart[91]-Dataset2$latent)

# Note that we will estimate expected HRQoL values so need
# probability of being in full health and N3 term
Dataset2$Prob.Full.Health <- Dataset2$Pred.EQ5D.Q1.L1*
  Dataset2$Pred.EQ5D.Q2.L1*Dataset2$Pred.EQ5D.Q3.L1*
  Dataset2$Pred.EQ5D.Q4.L1*Dataset2$Pred.EQ5D.Q5.L1
Dataset2$Prob.No.N3 <- Dataset2$Prob.Full.Health +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
  Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
  Dataset2$Pred.EQ5D.Q5.L1 +
  Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
  Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
  Dataset2$Pred.EQ5D.Q5.L1 +
  Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
```

```
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L2 +
```

```
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
Dataset2$Pred.EQ5D.Q5.L1 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L1*
Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
Dataset2$Pred.EQ5D.Q5.L2 +
Dataset2$Pred.EQ5D.Q1.L1*Dataset2$Pred.EQ5D.Q2.L2*
Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
```

```
  Dataset2$Pred.EQ5D.Q5.L2 +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L1*
  Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
  Dataset2$Pred.EQ5D.Q5.L2 +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
  Dataset2$Pred.EQ5D.Q3.L1*Dataset2$Pred.EQ5D.Q4.L2*
  Dataset2$Pred.EQ5D.Q5.L2 +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
  Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L1*
  Dataset2$Pred.EQ5D.Q5.L2 +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
  Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
  Dataset2$Pred.EQ5D.Q5.L1 +
  Dataset2$Pred.EQ5D.Q1.L2*Dataset2$Pred.EQ5D.Q2.L2*
  Dataset2$Pred.EQ5D.Q3.L2*Dataset2$Pred.EQ5D.Q4.L2*
  Dataset2$Pred.EQ5D.Q5.L2

# And finally apply EQ-5D weights
Dataset2$constant <- (1-Dataset2$Prob.Full.Health)*0.081
Dataset2$N3 <- (1-Dataset2$Prob.No.N3)*0.269
Dataset2$decrement_mob <- (Dataset2$Pred.EQ5D.Q1.L2*0.069) +
  (Dataset2$Pred.EQ5D.Q1.L3*0.314)
Dataset2$decrement_sc <- (Dataset2$Pred.EQ5D.Q2.L2*0.104) +
  (Dataset2$Pred.EQ5D.Q2.L3*0.214)
Dataset2$decrement_usual <- (Dataset2$Pred.EQ5D.Q3.L2*0.036) +
  (Dataset2$Pred.EQ5D.Q3.L3*0.094)
Dataset2$decrement_pain <- (Dataset2$Pred.EQ5D.Q4.L2*0.123) +
  (Dataset2$Pred.EQ5D.Q4.L3*0.386)
Dataset2$decrement_anx <- (Dataset2$Pred.EQ5D.Q5.L2*0.071) +
  (Dataset2$Pred.EQ5D.Q5.L3*0.236)
Dataset2$eq5d <- 1
Dataset2$eq5d <- 1 - Dataset2$constant - Dataset2$decrement_mob -
  Dataset2$decrement_sc -
  Dataset2$decrement_usual - Dataset2$decrement_pain -
  Dataset2$decrement_anx - Dataset2$N3
```

```
######################
# Mplus Code ########
# Model 5.3 #########
# Step 1 ############
######################

TITLE:  this is an example of a threshold
structure CFA for categorical factor indicators
DATA:   FILE IS Data.dat;
VARIABLE:       NAMES ARE HA sex SF36_Q1 SF36_Q4 SF36_Q6
SF36_Q14 SF36_Q15 SF36_Q18 SF36_Q19 SF36_Q22
SF36_Q26 SF36_Q27 SF36_Q28 SF36_Q32 EQ5D_Q1
EQ5D_Q2 EQ5D_Q3 EQ5D_Q4 EQ5D_Q5 age1 age2;
CATEGORICAL ARE SF36_Q1 SF36_Q4 SF36_Q6
SF36_Q14 SF36_Q15 SF36_Q18 SF36_Q19 SF36_Q22
SF36_Q26 SF36_Q27 SF36_Q28 SF36_Q32 EQ5D_Q1
EQ5D_Q2 EQ5D_Q3 EQ5D_Q4 EQ5D_Q5;
MISSING = .;
ANALYSIS:
  TYPE = general missing h1 ;
MODEL:  f1 BY SF36_Q1* SF36_Q4 SF36_Q6
SF36_Q14 SF36_Q15 SF36_Q18 SF36_Q19 SF36_Q22
SF36_Q26 SF36_Q27 SF36_Q28 SF36_Q32 EQ5D_Q1
EQ5D_Q2 EQ5D_Q3 EQ5D_Q4 EQ5D_Q5;
f2 BY SF36_Q1* SF36_Q4 SF36_Q6
SF36_Q14 SF36_Q15 SF36_Q18 SF36_Q19 SF36_Q22
SF36_Q26 SF36_Q27 SF36_Q28 SF36_Q32;
f3 BY EQ5D_Q1* EQ5D_Q2 EQ5D_Q3 EQ5D_Q4 EQ5D_Q5;
f1@1;
f2@1;
f3@1;
f1 WITH f2@0;
f1 WITH f3@0;
f2 WITH f3@0;
f1 ON HA sex age1 age2;
```

```
[SF36_Q1$1 SF36_Q4$1 SF36_Q6$1
 SF36_Q14$1 SF36_Q15$1
 SF36_Q18$1 SF36_Q19$1 SF36_Q22$1
 SF36_Q26$1 SF36_Q27$1 SF36_Q28$1
 SF36_Q32$1
 EQ5D_Q1$1 EQ5D_Q2$1 EQ5D_Q3$1 EQ5D_Q4$1 EQ5D_Q5$1
 SF36_Q1$2 SF36_Q4$2 SF36_Q6$2 SF36_Q14$2 SF36_Q15$2
 SF36_Q18$2 SF36_Q19$2 SF36_Q22$2 SF36_Q26$2 SF36_Q27$2
 SF36_Q28$2 SF36_Q32$2
 EQ5D_Q1$2 EQ5D_Q2$2 EQ5D_Q3$2 EQ5D_Q4$2 EQ5D_Q5$2
 SF36_Q1$3 SF36_Q14$3 SF36_Q15$3
 SF36_Q18$3 SF36_Q19$3 SF36_Q22$3
 SF36_Q26$3 SF36_Q27$3 SF36_Q28$3 SF36_Q32$3
 SF36_Q1$4 SF36_Q14$4 SF36_Q15$4
 SF36_Q18$4 SF36_Q19$4 SF36_Q22$4
 SF36_Q26$4 SF36_Q27$4 SF36_Q28$4
 SF36_Q32$4];
OUTPUT:
  Standardized TECH3;
SAVEDATA:
  Difftest is difftestfile.dat;
```

```
######################
# R Code #############
# Model 5.3 ##########
# Step 2 #############
######################

library(MplusAutomation) #use this command to run models
# from R and then extract the model output
library(simsem) # required for simulations

# In order to make conduct simulations we need a
# "template" model to which we can add parameter
# estimates from the mplus output
# We run analyses on RITA data to obtain this template

RITA[,c("EQ5D.Q1",
        "EQ5D.Q2",
        "EQ5D.Q3",
        "EQ5D.Q4",
        "EQ5D.Q5")] <-
  lapply(RITA[,c("EQ5D.Q1",
                 "EQ5D.Q2",
                 "EQ5D.Q3",
                 "EQ5D.Q4",
                 "EQ5D.Q5")], ordered)
model <- 'latent =~ EQ5D.Q1 + EQ5D.Q2 + EQ5D.Q3 +
EQ5D.Q4 + EQ5D.Q5
EQ5D.Q1 | t1 + t2
EQ5D.Q2 | t1 + t2
EQ5D.Q3 | t1 + t2
EQ5D.Q4 | t1 + t2
EQ5D.Q5 | t1 + t2 '
library(lavaan)
fit <- sem(model,RITA,
           ordered=c("EQ5D.Q1",
```

```
                    "EQ5D.Q2",
                    "EQ5D.Q3",
                    "EQ5D.Q4",
                    "EQ5D.Q5"))
model.5.3.specification <- parTable(fit)

# Next the appropriate parameter estimates
# from the mplus model are assigned to this
# template. These will vary depending upon
# the patient profile. Now simulate dataset

simulation <- data.frame(sim(1,100000,
                    model.5.3.specification,
                    model = model, dataOnly = TRUE))

# EQ5D values then assigned to the simulations
```

Appendix E

Chapter 6 Results

Table E.1: CFA Approach - Threshold Values

Indicator	Threshold 1	Threshold 2	Threshold 3	Threshold 4
SF36 Q1	-1.59	-0.70	0.23	1.21
SF36 Q2	-0.89	-0.40		
SF36 Q3	-0.60	0.11		
SF36 Q4	-1.37	-0.72	-0.04	0.51
SF36 Q5	-1.35	-0.83	-0.18	0.28
SF36 Q6	-1.59	-1.15	-0.60	-0.26
SF36 Q7	-1.84	-1.28	-0.72	-0.40
SF36 Q8	-0.05	0.48	0.91	1.54
SF36 Q9	-0.65	0.56	1.10	1.51
SF36 Q10	-1.21	-0.12	0.64	1.15
SF36 Q11	-1.90	-1.29	-0.72	-0.25
SF36 Q12	-1.57	-1.09	-0.50	0.00
EQ5D Q1	0.20	2.27		
EQ5D Q2	1.23	2.34		
EQ5D Q3	0.32	1.57		
EQ5D Q4	-0.35	1.30		
EQ5D Q5	0.42	1.56		

Table E.2: SEM Approach - Threshold Values

Indicator	Threshold 1	Threshold 2	Threshold 3	Threshold 4
SF36 Q1	-1.84	-0.82	0.21	1.23
SF36 Q2	-0.96	-0.22		
SF36 Q3	-0.74	0.06		
SF36 Q4	-1.34	-0.78	-0.13	0.33
SF36 Q5	-1.37	-0.87	-0.27	0.15
SF36 Q6	-1.30	-0.84	-0.31	0.05
SF36 Q7	-1.39	-0.93	-0.39	-0.02
SF36 Q8	-0.36	0.24	0.69	1.41
SF36 Q9	-1.36	-0.12	0.61	1.25
SF36 Q10	-1.52	-0.41	0.40	1.01
SF36 Q11	-1.39	-0.83	-0.05	0.62
SF36 Q12	-1.35	-0.81	-0.20	0.24
EQ5D Q1	0.29	2.04		
EQ5D Q2	0.90	2.13		
EQ5D Q3	0.14	1.36		
EQ5D Q4	-0.33	1.33		
EQ5D Q5	-0.10	1.12		

Appendix F

Chapter 7 Code

```
#######################
# R Code ##############
# Model 6.2 ###########
# Mixed outcomes ####
# model ###############
#######################

Model.6.2 <-
  "CF =~ sf6d + hui3score +
  EQ5D.Q1 + EQ5D.Q2 + EQ5D.Q3 + EQ5D.Q4 + EQ5D.Q5
EQ5D.Q1 | t1 + t2
EQ5D.Q2 | t1 + t2
EQ5D.Q3 | t1 + t2
EQ5D.Q4 | t1 + t2
EQ5D.Q5 | t1 + t2 "
NHMS[,  # we need to define the dataset as being ordinal
      c("EQ5D.Q1","EQ5D.Q2",
        "EQ5D.Q3","EQ5D.Q4",
        "EQ5D.Q5")] <-
  lapply(NHMS[,c("EQ5D.Q1",
         "EQ5D.Q2","EQ5D.Q3",
         "EQ5D.Q4","EQ5D.Q5")],
         ordered)
fit <- #now fit the model
```

```
  sem(Model, data = NHMS, std.lv=TRUE,
      ordered=c("EQ5D.Q1",
                "EQ5D.Q2","EQ5D.Q3","EQ5D.Q4",
                "EQ5D.Q5"))

param.estimates <- parameterEstimates(fit)
```

References

Ades, A. E. and Sutton, A. J. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(1):5–35, 2006.

Ades, AE; Lu, Guobing, and Madan, Jason 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*, 16(1):185–194, 2013.

Alava, Mónica Hernández; Wailoo, Allan J, and Ara, Roberta. Tails from the peak district: adjusted limited dependent variable mixture models of eq-5d questionnaire health state utility values. *Value in Health*, 15(3):550–561, 2012.

Alava, Mónica Hernández; Wailoo, Allan; Wolfe, Fred, and Michaud, Kaleb. A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. *Medical Decision Making*, 34(7):919–930, 2014.

Ara, Roberta and Brazier, John E. Populating an economic model with health state utility values: moving toward better practice. *Value in Health*, 13(5):509–518, 2010.

Ara, Roberta and Wailoo, Allan. Nice dsu technical support document 12: the use of health state utility values in decision models. *Sheffield, UK: School of Health and Related Research, University of Sheffield*, 2011.

Ara, Roberta and Wailoo, Allan. Using health state utility values in models exploring the cost-effectiveness of health technologies. *Value in Health*, 15(6):971–974, 2012.

Askew, Robert L; Swartz, Richard J; Xing, Yan; Cantor, Scott B; Ross, Merrick I; Gershenwald, Jeffrey E; Palmer, J Lynn; Lee, Jeffrey E, and Cormier, Janice N. Mapping fact-melanoma quality-of-life scores to eq-5d health utility weights. *Value in Health*, 14 (6):900–906, 2011.

Baio, Gianluca. *Bayesian methods in health economics*. CRC Press, 2012.

Baker, R.; Donaldson, C.; Mason, H., and Jones-Lee, M. Willingness to pay for health. In Culyer, Anthony J., editor, *Encyclopedia of Health Economics*, pages 495 – 501. Elsevier, San Diego, 2014. ISBN 978-0-12-375679-4. doi: http://dx.doi.org/10.1016/B978-0-12-375678-7.00503-4. URL `http://www.sciencedirect.com/science/article/pii/B9780123756787005034`.

Bansback, Nick; Sun, Huiying; Guh, Daphne P; Li, Xin; Nosyk, Bohdan; Griffin, Susan; Barnett, Paul G, and Anis, Aslam H. Impact of the recall period on measuring health utilities for acute events. *Health economics*, 17(12):1413–1419, 2008.

Basu, Anirban and Manca, Andrea. Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, 32(1):56–69, 2012.

Basu, Anirban and Meltzer, David. Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, 27(2):112–127, 2007.

Bleichrodt, Han. A new explanation for the difference between time trade-o¡ utilities and standard gamble utilities. *Health Econ*, 11:447–456, 2002.

Bojke, Laura; Claxton, Karl; Sculpher, Mark, and Palmer, Stephen. Characterizing structural uncertainty in decision analytic models: a review and application of methods. *Value in Health*, 12(5):739–749, 2009.

Bovaird, JA and Koziol, NA. Measurement models for ordered-categorical indicators. *Handbook of structural equation modeling*, pages 495–511, 2012.

Brazier, John. *Measuring and valuing health benefits for economic evaluation*. Oxford University Press, 2007.

Brazier, John; Roberts, Jennifer, and Deverill, Mark. The estimation of a preference-based measure of health from the sf-36. *Journal of health economics*, 21(2):271–292, 2002.

Brazier, John E; Yang, Yaling; Tsuchiya, Aki, and Rowen, Donna Louise. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European journal of health economics*, 11(2):215–225, 2010.

Briggs, A. and Gray, A. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technology Assessment*, 3(2):134, 1999. doi: 10.3310/hta3020. URL `http//journalslibrary.nihr.ac.uk/hta/hta3020`.

Briggs, AH and Gray, AM. Handling uncertainty when performing economic evaluation of healthcare interventions. *Health technology assessment (Winchester, England)*, 3(2): 1–134, 1998.

Briggs, Andrew; Sculpher, Mark, and Claxton, Karl. *Decision modelling for health economic evaluation*. Oxford university press, 2006.

Brouwer, Werner BF; Culyer, Anthony J; van Exel, N Job A, and Rutten, Frans FH. Welfarism vs. extra-welfarism. *Journal of health economics*, 27(2):325–338, 2008.

Brown, Timothy A and Moore, Michael T. Confirmatory factor analysis. *Handbook of structural equation modeling*, pages 361–379, 2012.

Bujkiewicz, Sylwia; Thompson, John R; Sutton, Alex J; Cooper, Nicola J; Harrison, Mark J; Symmons, Deborah PM, and Abrams, Keith R. Multivariate meta-analysis of mixed outcomes: a bayesian approach. *Statistics in medicine*, 32(22):3926–3943, 2013.

CADTH, . Cadth guidelines for the evaluation of health technologies. Technical report, Canadian Agency for Drugs and Technologies in Health (CADTH), 2006.

Cheng, André K and Niparko, John K. Cost-utility of the cochlear implant in adults: a meta-analysis. *Archives of Otolaryngology–Head & Neck Surgery*, 125(11):1214–1218, 1999.

Cheung, Mike W-L. Multivariate meta-analysis as structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3):429–454, 2013.

Cheung, Mike W-L. Fixed-and random-effects meta-analytic structural equation modeling: Examples and analyses in r. *Behavior research methods*, 46(1):29–40, 2014a.

Cheung, Mike W-L. metasem: an r package for meta-analysis using structural equation modeling. *Frontiers in psychology*, 5, 2014b.

Chuang, Ling-Hsiang and Kind, Paul. Converting the sf-12 into the eq-5d. *Pharmacoeconomics*, 27(6):491–505, 2009.

Claxton, Karl; Martin, Steve; Soares, Marta; Rice, Nigel; Spackman, Eldon; Hinde, Sebastian; Devlin, Nancy; Smith, Peter C, and Sculpher, Mark. Methods for the estimation of the nice cost effectiveness threshold. Technical report, 2013.

Cleemput, Irina; Neyt, M; Thiry, N; De Laet, C, and Leys, M. Threshold values for cost-effectiveness in health care. *Health Technology Assessment (HTA). Brussels: Belgian Health Care Knowledge Centre (KCE)*, 2008.

Conigliani, Caterina; Manca, Andrea, and Tancredi, Andrea. Prediction of patient-reported outcome measures via multivariate ordered probit models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):567–591, 2015.

Conner-Spady, Barbara and Suarez-Almazor, Maria E. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Medical care*, 41 (7):791–801, 2003.

Cooper, Nicola; Coyle, Douglas; Abrams, Keith; Mugford, Miranda, and Sutton, Alexander. Use of evidence in decision models: an appraisal of health technology assessments in the uk since 1997. *Journal of health services research & policy*, 10(4):245–250, 2005.

Crott, Ralph and Briggs, Andrew. Mapping the qlq-c30 quality of life cancer questionnaire to eq-5d patient preferences. *The European journal of health economics*, 11(4):427–434, 2010.

Dakin, Helen; Gray, Alastair, and Murray, David. Mapping analyses to estimate eq-5d utilities and responses based on oxford knee score. *Quality of Life Research*, 22(3): 683–694, 2013.

Djalalov, Sandjar; Rabeneck, Linda; Tomlinson, George; Bremner, Karen E; Hilsden, Robert, and Hoch, Jeffrey S. A review and meta-analysis of colorectal cancer utilities. *Medical Decision Making*, 34(6):809–818, 2014.

Dolan, Paul. Modeling valuations for euroqol health states. *Medical care*, 35(11):1095–1108, 1997.

Dolan, Paul and Tsuchiya, Aki. The elicitation of distributional judgements in the context of economic evaluation. *The elgar companion to health economics*, page 382, 2006.

Dolan, Paul; Shaw, Rebecca; Tsuchiya, Aki, and Williams, Alan. Qaly maximisation and people's preferences: a methodological review of the literature. *Health economics*, 14 (2):197–208, 2005.

Doth, Alissa H.; Hansson, Per T.; Jensen, Mark P., and Taylor, Rod S. The burden of neuropathic pain: a systematic review and meta-analysis of health utilities. *PAIN®*, 149(2):338–344, 2010.

Drummond, Michael. Future prospects for pharmacoeconomics and outcomes research in the emerging regions. *Value in Health Regional Issues*, 1(2):3–4, 2013a.

Drummond, Michael. Twenty years of using economic evaluations for drug reimbursement decisions: what has been achieved? *Journal of health politics, policy and law*, 38(6): 1081–1102, 2013b.

Drummond, Michael F. *Methods for the economic evaluation of health care programmes*. Oxford university press, 2005.

Edwards, Michael C; Wirth, RJ; Houts, Carrie R; Xi, Nuo, and Hoyle, RH. Categorical data in the structural equation modeling framework. *Handbook of structural equation modeling*, pages 195–208, 2012.

Egger, Matthias and Smith, George Davey. Meta-analysis: potentials and promise. *Bmj*, 315(7119):1371–1374, 1997.

Espinoza, Manuel A; Manca, Andrea; Claxton, Karl, and Sculpher, Mark J. The value of heterogeneity for cost-effectiveness subgroup analysis conceptual framework and application. *Medical Decision Making*, page 0272989X14538705, 2014.

Fayers, Peter M; Aaronson, Niel K; Bjordal, Kristin; Curran, D, and Grønvold, Mogens. Eortc qlq-c30 scoring manual. Technical report, EORTC, 1999.

Feeny, David; Furlong, William; Torrance, George W.; Goldsmith, Charles H.; Zhu, Zenglong; DePauw, Sonja; Denton, Margaret, and Boyle, Michael. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical care*, 40(2):113–128, 2002.

Fryback, Dennis G; Dunham, Nancy Cross; Palta, Mari; Hanmer, Janel; Buechner, Jennifer; Cherepanov, Dasha; Herrington, Shani; Hays, Ron D; Kaplan, Robert M; Ganiats, Theodore G, and others, . Us norms for six generic health-related quality-of-life indexes from the national health measurement study. *Medical care*, 45(12):1162, 2007.

Fuguitt, Diana and Wilcox, Shanton J. *Cost-benefit analysis for public sector decision makers*. Greenwood Publishing Group, 1999.

Gafni, Amiram. The standard gamble method: what is being measured and how it is interpreted. *Health Services Research*, 29(2):207, 1994.

Gibbons, Robert D; Perraillon, Marcelo Coca, and Kim, Jong Bae. Item response theory approaches to harmonization and research synthesis. *Health Services and Outcomes Research Methodology*, 14(4):213–231, 2014.

Gold, Marthe and Siegel, . Panel on cost-effectiveness in health and medicine. *Medical care*, pages DS197–DS199, 1996.

Graham, John W and Coffman, Donna L. Structural equation modeling with missing data. *Handbook of structural equation modeling*, pages 277–94, 2012.

Gray, Alastair M; Rivero-Arias, Oliver, and Clarke, Philip M. Estimating the association between sf-12 responses and eq-5d utility values by response mapping. *Medical Decision Making*, 26(1):18–29, 2006.

Griffin, Susan C; Claxton, Karl P; Palmer, Stephen J, and Sculpher, Mark J. Dangerous omissions: the consequences of ignoring decision uncertainty. *Health economics*, 20(2): 212–224, 2011.

Gudex, Claire and others, . Time trade-off user manual: props and self-completion methods. Technical report, 1994.

Hadi, M; Gibbons, E, and Fitzpatrick, R. A structured review of patient reported outcome measures (proms) for colorectal cancer, report to the department of health, 2010.

Harris, John. It's not nice to discriminate. *Journal of Medical Ethics*, 31(7):373–375, 2005.

Henriksson, Martin; Epstein, David M; Palmer, Stephen J; Sculpher, Mark J; Clayton, Tim C; Pocock, Stuart J; Henderson, Robert A; Buxton, Martin J, and Fox, Keith AA. The cost-effectiveness of an early interventional strategy in non-st-elevation acute coronary syndrome based on the rita 3 trial. *Heart*, 94(6):717–723, 2008.

Higgins, Julian; Thompson, Simon G, and Spiegelhalter, David J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):137–159, 2009.

Hinde, Sebastian and Spackman, Eldon. Bidirectional citation searching to completion: An exploration of literature searching methods. *PharmacoEconomics*, 33(1):5–11, 2015.

Hunger, Matthias; Baumert, Jens, and Holle, Rolf. Analysis of sf-6d index data: is beta regression appropriate? *Value in Health*, 14(5):759–767, 2011.

Hurley, J. Welfarism and extra-welfarism. *Encyclopedia of Health Economics*, pages 483–489, 2014.

Hutton, John; McGrath, Clare; Frybourg, Jean-Marc; Tremblay, Mike; Bramley-Harker, Edward, and Henshall, Christopher. Framework for describing and classifying decision-making systems using technology assessment to determine the reimbursement of health technologies (fourth hurdle systems). *International journal of technology assessment in health care*, 22(01):10–18, 2006.

Kavanaugh, Arthur; Helliwell, Philip, and Ritchlin, Christopher T. Psoriatic arthritis and burden of disease: Patient perspectives from the population-based multinational assessment of psoriasis and psoriatic arthritis (mapp) survey. *Rheumatology and Therapy*, pages 1–12, 2016.

Kenny, David A and Milan, Stephanie. Identification: A non-technical discussion of a technical issue. *Handbook of Structural Equation Modeling (R. Hoyle, ed.). Guilford Press, New York, This volume*, 2011.

Kharroubi, Samer A; O'Hagan, Anthony, and Brazier, John E. A comparison of united states and united kingdom eq-5d health states valuations using a nonparametric bayesian method. *Statistics in medicine*, 29(15):1622–1634, 2010.

Kharroubi, Samer A; Brazier, John E, and McGhee, Sarah. A comparison of hong kong and united kingdom sf-6d health states valuations using a nonparametric bayesian method. *Value in Health*, 17(4):397–405, 2014.

Kim, Joseph; Henderson, Robert A; Pocock, Stuart J; Clayton, Tim; Sculpher, Mark J, and Fox, Keith AA. Health-related quality of life after interventional or conservative strategy in patients with unstable angina or non–st-segment elevation myocardial infarction: one-year results of the third randomized intervention trial of unstable angina (rita-3). *Journal of the American College of Cardiology*, 45(2):221–228, 2005.

Kind, Paul; Hardman, Geoffrey; Macran, Susan, and others, . *UK population norms for EQ-5D*, volume 172. Centre for Health Economics, University of York York, 1999.

Kinney, Marguerite R; Burfitt, Sandra N; Stullenbarger, Elizabeth; Rees, Barbara, and DeBolt, Marie Read. Quality of life in cardiac patient research: a meta-analysis. *Nursing research*, 45(3):173–180, 1996.

Kline, Rex B. Structural equation modeling, 2006.

Le, Quang A and Doctor, Jason N. Probabilistic mapping of descriptive health status responses onto health state utilities using bayesian networks: an empirical analysis

converting sf-12 into eq-5d utility index in a national us sample. *Medical care*, 49(5): 451–460, 2011.

Lee, Sik-Yum. *Structural equation modeling: A Bayesian approach*, volume 711. John Wiley & Sons, 2007.

Lei, Pui-Wa and Wu, Qiong. Estimation in structural equation modeling. *Handbook of structural equation modeling*, pages 164–179, 2012.

Liem, Ylian S; Bosch, Johanna L, and Myriam Hunink, MG. Preference-based quality of life of patients on renal replacement therapy: A systematic review and meta-analysis. *Value in Health*, 11(4):733–741, 2008.

Longworth, Louise and Rowen, Donna. Nice dsu technical support document 10: the use of mapping methods to estimate health state utility values. *London: NICE*, page b3, 2011.

Lu, Guobing; Brazier, JE, and Ades, AE. Mapping from disease-specific to generic health-related quality-of-life scales: a common factor model. *Value in Health*, 16(1):177–184, 2013.

Lu, Guobing; Kounali, Daphne, and Ades, AE. Simultaneous multioutcome synthesis and mapping of treatment effects to a common scale. *Value in Health*, 17(2):280–287, 2014.

Lung, Tom WC; Hayes, Alison J; Hayen, Andrew; Farmer, Andrew, and Clarke, Philip M. A meta-analysis of health state valuations for people with diabetes: explaining the variation across methods and implications for economic evaluation. *Quality of Life Research*, 20(10):1669–1678, 2011.

Lunn, David; Jackson, Chris; Best, Nicky; Thomas, Andrew, and Spiegelhalter, David. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2012.

Mackintosh, A; Gibbons, E, and Fitzpatrick, R. A structured review of patient-reported outcome measures (proms) for heart failure, 2009.

Mahon, Ronan. *TEMPORAL UNCERTAINTY IN COST-EFFECTIVENESS DECISION MODELS: METHODS TO ADDRESS THE UNCERTAINTIES THAT ARISE WHEN THE APPROPRIATE ANALYSIS TIME HORIZON EXCEEDS THE EVIDENCE TIME HORIZON IN COST-EFFECTIVENESS DECISION MODELS AS APPLIED TO HEALTHCARE INTERVENTIONS*. PhD thesis, University of York, 2014.

McKenzie, Lynda and Van Der Pol, Marjon. Mapping the eortc qlq c-30 onto the eq-5d instrument: The potential to estimate qalys without generic preference data. *Value in Health*, 12(1):167–171, 2009.

McLernon, David J; Dillon, John, and Donnan, Peter T. Health-state utilities in liver disease: a systematic review. *Medical Decision Making*, 2008.

Menzel, P. T. Utilities for health states: Whom to ask. In *Encyclopedia of Health Economics*. Elsevier, 2014.

Mohiuddin, Syed and Payne, Katherine. Utility values for adults with unipolar depression systematic review and meta-analysis. *Medical Decision Making*, page 0272989X14524990, 2014.

Moon-Ho, RH; Stark, Stephen, and Chernyshenko, OS. Graphical representation of structural equation models using path diagrams. *Handbook of Structural Equation Modeling; NY: Guilford Press.*, page 43, 2012.

Muthén, Linda K and Muthén, Bengt O. *Mplus User's Guide: Statistical Analysis with Latent Variables: User'ss Guide*. Muthén & Muthén, 2010.

NatCen Social Research, . Welsh health survey, 2013. URL `http://dx.doi.org/10.5255/UKDA-SN-7632-1`.

NICE, . Guide to the methods of technology appraisal. Technical report, National Institute for Health and Care Excellence (NICE), 2008.

NICE, . Guide to the methods of technology appraisal. Technical report, National Institute for Health and Care Excellence (NICE), 2013.

Papaioannou, Diana; Brazier, John, and Paisley, Suzy. Nice dsu technical support document 9: the identification, review and synthesis of health state utility values from the literature. Technical report, 2011.

Papaioannou, Diana; Brazier, John, and Paisley, Suzy. Systematic searching and selection of health state utility values from the literature. *Value in Health*, 16(4):686–695, 2013.

Parkin, David; Rice, Nigel, and Devlin, Nancy. Statistical analysis of eq-5d profiles: does the use of value sets bias inference? *Medical Decision Making*, 30(5):556–565, 2010.

Paulden, Mike; Culyer, Anthony J, and others, . *Does Cost-effectiveness Analysis Discriminate Against Patients with Short Life Expectancy?: Matters of Logic and Matters of Context*. Toronto Health Economics and Technology Assessment Collaborative, 2010.

PBAC, . Guidelines for preparing submissions to the pharmaceutical benefits advisory committee. Technical report, Pharmaceutical Benefits Advisory Committee (PBAC), 2008.

Peasgood, T.; Herrmann, K.; Kanis, J. A., and Brazier, J. E. An updated systematic review of health state utility values for osteoporosis related conditions. *Osteoporosis International*, 20(6):853–868, 2009.

Peasgood, Tessa and Brazier, John. Is meta-analysis for utility values appropriate given the potential impact different elicitation methods have on values? *PharmacoEconomics*, pages 1–5, 2015.

Peasgood, Tessa; Ward, Sue E, and Brazier, John. Health-state utility values in breast cancer. 2010.

Piantadosi, Steven; Byar, David P, and Green, Sylvan B. The ecological fallacy. *American Journal of Epidemiology*, 127(5):893–904, 1988.

Pornprasertmanit, Sunthud; Miller, Patrick, and Schoemann, Alexander. simsem: simulated structural equation modeling. r package version 0.5-3, 2013.

Rawlins, Michael D and Culyer, Anthony J. National institute for clinical excellence and its value judgments. *BMJ: British Medical Journal*, 329(7459):224, 2004.

Reise, Steven P. The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5):667–696, 2012.

Riley, RD; Price, MJ; Jackson, D; Wardle, M; Gueyffier, F; Wang, J; Staessen, JA, and White, IR. Multivariate meta-analysis using individual participant data. *Research Synthesis Methods*, 2014.

Rivero-Arias, Oliver; Ouellet, Melissa; Gray, Alastair; Wolstenholme, Jane; Rothwell, Peter M, and Luengo-Fernandez, Ramon. Mapping the modified rankin scale (mrs) measurement into the generic euroqol (eq-5d) health outcome. *Medical decision making*, 30(3):341–354, 2010.

Rosseel, Yves. lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

Rowen, Donna; Brazier, John; Young, Tracey; Gaugris, Sabine; Craig, Benjamin M; King, Madeleine T, and Velikova, Galina. Deriving a preference-based measure for cancer using the eortc qlq-c30. *Value in Health*, 14(5):721–731, 2011.

Sakthong, Phantipa. Measurement of clinical-effect: utility. *Journal of the Medical Association of Thailand*, 91:S43–S52, 2008.

Sampson, Christopher J; Tosh, Jonathan C; Cheyne, Christopher P; Broadbent, Deborah, and James, Marilyn. Health state utility values for diabetic retinopathy: protocol for a systematic review and meta-analysis. *Systematic reviews*, 4(1):1, 2015.

Saramago, Pedro; Manca, Andrea, and Sutton, Alex J. Deriving input parameters for cost-effectiveness modeling: taxonomy of data types and approaches to their statistical synthesis. *Value in Health*, 15(5):639–649, 2012.

Sculpher, Mark J; Claxton, Karl; Drummond, Mike, and McCabe, Chris. Whither trial-based economic evaluation for health care decision making? *Health economics*, 15(7): 677–688, 2006.

Shah, Koonal K. Severity of illness and priority setting in healthcare: a review of the literature. *Health policy*, 93(2):77–84, 2009.

Shah, Koonal K; Tsuchiya, Aki, and Wailoo, Allan J. Valuing health at the end of life: an empirical study of public preferences. *The European Journal of Health Economics*, 15(4):389–399, 2014.

Skrondal, Anders and Rabe-Hesketh, Sophia. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press, 2004.

StataCorp, LP. Stata 13, 2013.

Sturza, Julie. A review and meta-analysis of utility values for lung cancer. *Medical Decision Making*, 30(6):685–693, 2010.

Su, Yu-Sung and Yajima, Masanao. R2jags: A package for running jags from r. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*, 2012.

Sullivan, Patrick W; Slejko, Julia F; Sculpher, Mark J, and Ghushchyan, Vahram. Catalogue of eq-5d scores for the united kingdom. *Medical Decision Making*, 31(6):800–804, 2011.

Sutton, Alex J and Abrams, Keith R. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, 10(4):277–303, 2001.

Sutton, Alex J; Abrams, Keith R; Jones, David R; Jones, David R; Sheldon, Trevor A, and Song, Fujian. *Methods for meta-analysis in medical research*. J. Wiley Chichester; New York, 2000.

Sutton, Alexander J; Welton, Nicky J; Cooper, Nicola; Abrams, Keith R, and Ades, AE. *Evidence synthesis for decision making in healthcare*, volume 132. John Wiley & Sons, 2012.

Team, R Core. R: A language and environment for statistical computing. r foundation for statistical computing, vienna, austria, 2012, 2014.

Tengs, Tammy O and Lin, Ting H. A meta-analysis of utility estimates for hiv/aids. *Medical Decision Making*, 22(6):475–481, 2002.

Tengs, Tammy O and Lin, Ting H. A meta-analysis of quality-of-life estimates for stroke. *Pharmacoeconomics*, 21(3):191–200, 2003.

Thompson, Simon G and Barber, Julie A. How should cost data in pragmatic randomised trials be analysed? *BMJ: British Medical Journal*, 320(7243):1197, 2000.

Thompson, Simon G and Higgins, Julian PT. How should meta-regression analyses be undertaken and interpreted? *Statistics in medicine*, 21(11):1559–1573, 2002.

Tosh, Jonathan C; Longworth, Louise J, and George, Elisabeth. Utility values in national institute for health and clinical excellence (nice) technology appraisals. *Value in Health*, 14(1):102–109, 2011.

van Hout, Ben; Janssen, MF; Feng, You-Shan; Kohlmann, Thomas; Busschbach, Jan; Golicki, Dominik; Lloyd, Andrew; Scalone, Luciana; Kind, Paul, and Pickard, A Simon. Interim scoring for the eq-5d-5l: mapping the eq-5d-5l to eq-5d-3l value sets. *Value in Health*, 15(5):708–715, 2012.

Wei, Yinghui and Higgins, Julian. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in medicine*, 32(7):1191–1205, 2013.

Weinstein, Milton C; Siegel, Joanna E; Gold, Marthe R; Kamlet, Mark S, and Russell, Louise B. Recommendations of the panel on cost-effectiveness in health and medicine. *Jama*, 276(15):1253–1258, 1996.

Wyld, Melanie; Lisa Morton, Rachael; Hayen, Andrew; Howard, Kirsten, and Claire Webster, Angela. A systematic review and meta-analysis of utility-based quality of life in chronic kidney disease treatments. *PLoS medicine*, 9(9):1435, 2012.