MAPPING TO ESTIMATE HEALTH STATE UTILITIES TECHNICAL SUPPORT DOCUMENT 22

[*updates and replaces TSD10*]

REPORT BY THE DECISION SUPPORT UNIT

June 2023

Allan Wailoo, Monica Hernandez Alava, Steve Pudney

School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK.

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA Tel (+44) (0)114 222 0734 E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is based at the University of Sheffield with members at the universities of York, Bristol, Leicester, Warwick and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information: https://nicedsu.sites.sheffield.ac.uk/

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

NICE describes the methods it follows when carrying out health technology evaluations in its process and methods manual. This provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The manual does not provide detailed advice on how to implement and apply the methods it describes. The DSU series of Technical Support Documents (TSDs) is intended to complement the manual by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in selected topic areas. They make recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE technology evaluations, whether companies, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD

lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides. The TSDs will be amended and updated whenever appropriate. Where minor updates or corrections are required, the TSD will retain its numbering with a note to indicate the date and content change of the last update. More substantial updates will be contained in new TSDs that entirely replace existing TSDs.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Prof Allan Wailoo, Director of DSU and TSD series editor.

ACKNOWLEDGEMENTS:

We are grateful to Donna Rowen and Andrea Manca for comments on a previous draft. We also received helpful feedback from colleagues at NICE: Ian Watson, Lewis Ralph and Fatima Chunara.

This report should be referenced as follows:

Wailoo, A., Hernandez Alava, M., Pudney, S. NICE DSU Technical Support Document 22 Mapping to estimate health state utilities. 2023 [Available from http://www.nicedsu.org.uk]

EXECUTIVE SUMMARY

This Technical Support Document (TSD) updates and replaces TSD10 “The use of mapping methods to estimate health state utility values”. Mapping is widely used in NICE technology evaluations as a means of bridging the evidence gap that often exists between the information provided by clinical studies of the effectiveness of the health technology under investigation and the requirements of the cost effectiveness analysis to express benefits in terms of Quality Adjusted life years (QALYs). Mapping is the term used for the process of estimating the relationship between one or more measures of health outcomes (and possibly other contextual explanatory variables) and health state utility values, such as those from the EQ-5D-3L. The mapping is then used to predict the health utility values to be applied for health states and events in cost effectiveness analysis.

It is important for mapping analyses to be performed with care and reported appropriately. Mapping can lead to biased estimates of the value of health benefits from effective health care technologies if good practices are not followed. This document provides guidance, with supporting evidence, to ensure that those that submit evidence to NICE and those that need to critique that evidence, can do so with confidence.

This report provides guidance on when mapping can be considered a valid solution to bridging this evidence gap. It then discusses issues of data selection for mapping. Because health utility data is characterised by several statically challenging features, the appropriateness and performance of different types of statistical models requires detailed consideration. The report covers both direct mapping, where the health utility score is modelled in a single stage, and response mapping, a two-stage method that models the responses to the descriptive system of the utility instrument as a first step and then uses the estimated models to calculate the expected health utility for the relevant value set. The recommendations cover the inclusion of covariates, methods for the assessment of model performance and the subsequent use of mapping models in cost effectiveness analysis. This includes how to reflect parameter uncertainty and, where relevant, individual level heterogeneity. A separate section covers the special case of mapping from the EQ-5D-5L to the value set of the EQ-5D-3L.

CONTENTS

FIGURES

1. INTRODUCTION

Most, though not all, of the National Institute for Health and Care Excellence (NICE) guidance producing programmes require the estimation of health benefits in terms of Quality Adjusted Life Years (QALYs) for the purposes of economic evaluation. QALYs comprise two components: length of life and an adjustment factor for the degree of ill-health. The adjustment factor is calibrated around 1 (full health) and 0 (health states judged equivalent to being dead). Most health states lie in the range 0 to 1, though extremely poor health states may be considered worse than death and therefore would be assigned a negative score. These values are often referred to as health state "utility" values and can be calculated in different ways. Individuals can be asked to directly value the health state they are experiencing, or hypothetical states that are described to them, using valuation tasks such as Time-Trade-Off or Standard Gamble. However, these methods are time consuming and complex to administer. The most typically applied method is to use a generic preference-based measure. These measures have been developed to be applicable across different disease areas and they separate out the measurement or descriptive element of relevant health states from their valuation.

One of the most commonly used generic preference-based measures is EQ-5D-3L[1], currently the preferred method for measuring health related quality of life in NICE health technology evaluations. The EQ-5D-3L instrument is typically administered to patients as a self-completion questionnaire. The instrument describes health in terms of five dimensions: mobility, self-care, usual activities, pain or discomfort, and anxiety or depression; and measures each dimension on three levels (no problems, some problems, extreme problems). With five dimensions, each at a possible three levels, the EQ-5D-3L instrument can describe 243 different health states. Each health state is referred to as a five-digit number – for example, health state 22231 refers to some

problems with mobility, self-care and usual activities, extreme problems with pain or discomfort and no problems with anxiety or depression. A separate valuation study published in 1997 estimated an index or utility score for each of the health states based on the preferences of a sample of the general public, the first UK value set[2]. The health state 11111 represents full health and is assigned a utility score of 1, death is assigned a value of 0, and states with negative utilities are considered worse than death.

There are numerous situations where evidence on EQ-5D-3L is missing or insufficient. The clinical studies of the technology in question may not have included the instrument. Or the evidence from those studies may be insufficient for the needs of the economic analysis, which often needs to extrapolate beyond the time period, or health events observed in the clinical studies. For example, many oncology trials will collect patient quality of life data until the time of cancer progression but not beyond, and, as a consequence, the health state utility for post progression survival is unobserved. It may also be the case that events occur only rarely within clinical studies, such as with many adverse events. Whatever the reason, mapping is one several options that can be considered in the face of this type of evidence gap.

What is mapping?

Mapping is a term used to describe the process of estimating the relationship between one or more measures of health outcomes (and possibly other contextual explanatory variables) and health state utility values, such as EQ-5D-3L. The terms ‘‘mapping,’’ ‘‘cross-walking,’’ and ‘‘transfer to utility’’ have all been used in previous literature[3]. “Mapping” is the most commonly used term. This requires the use of data from one or more studies where the relevant explanatory variables and the health state utility

values have both been recorded – a multi-instrument dataset. This relationship can then be used to predict the unobserved utility values of relevance for a cost-effectiveness analysis based on the observed explanatory variables.

In "NICE Health Technology Evaluations: the manual"[4] (which we refer to as the "Methods Guide"), it is stated:

> *"When EQ-5D data is not available, this data can be estimated by mapping other health-related quality-of-life measures or health-related benefits seen in the relevant clinical trials to EQ-5D. This is considered to be a departure from the reference case. The mapping function chosen should be based on data sets containing both health-related quality-of-life measures and its statistical properties. It should be fully described, its choice justified, and it should be adequately shown how well the function fits the data. Present sensitivity analyses to explore variation in using mapping algorithms on the outputs." (Section 4.3.9)*

At the heart of the concept of mapping is a statistical relationship between one measure and one or more others. The methods used to estimate this relationship should be consistent with accepted best practice in statistical analysis generally. However, there are specific challenges arising from particular features of the data on health utility scores and from the application of mapping to inform cost effectiveness models (such as the needs for accuracy in mean utility estimates and reflection of parameter uncertainty). The purpose of this document is to provide a summary of these issues and make recommendations where it is appropriate to do so, to assist those submitting evidence to NICE and those who review it. It draws on the existing Technical Support Document on mapping[5] (TSD10). It updates that work and

supersedes it. It reflects research evidence published since that time, including a major report published as an HTA monograph[6], ISPOR[7] good practice guidance documents and reporting standards adopted by a number of journals[8]. Compared to TSD10 there are updates in the areas of how to select appropriate datasets for mapping, the appropriateness of different model types, the issue of out-of-sample validation, the assessment of the optimal model type and how to reflect both uncertainty and variability. We also include a new section on the special case of mapping between different variants of EQ-5D.

We will focus our attention on mapping to EQ-5D-3L (and refer to it simply as EQ-5D from here onwards). The document is largely valid for mapping to any preference-based instrument, whether generic or disease specific. Relevant exceptions to this are highlighted. Most published mapping studies and research into methods for mapping are based on case studies using the 3-level EQ-5D[9], but there are significant numbers of studies that consider other generic instruments such as EQ-5D-5L, SF-6D and HUI[6,10]. The circumstances in which there is a need to map to a disease specific preference-based measure should be rare, since the exclusion of these instruments in relevant clinical studies would be difficult to justify. This report also includes a separate section devoted specifically to the issue of mapping between EQ-5D-5L and the 3-level EQ-5D.

Most economic analyses submitted to NICE take the form of a mathematical model that synthesises evidence from various sources on costs and health effects in the form of QALYs. We will therefore focus on issues relating to mapping for model based economic evaluation. In other settings, economic evaluations may be undertaken alongside a clinical trial, using individual patient level data and this can raise some minor differences in the way mapping is used.

Other NICE requirements have implications for the conduct of mapping. Probabilistic Sensitivity Analysis (PSA) is a requirement, where input parameters are assigned probability distributions to reflect parameter uncertainty. Estimates of health state utilities are input parameters whether derived from mapping or any other source, and thus the requirement for uncertainty to be appropriately reflected applies.

NICE recommends the 3-level version of the EQ-5D. Whilst there is a 5-level version of EQ-5D, NICE recommends data collected using the descriptive system of this instrument be mapped to the 3-level EQ-5D value set and a specific set of methods have been proposed for performing this mapping (see section 6).

The aim of this document is to provide more information for those performing and reviewing mapping as part of a NICE submission. This includes those situations where analysts estimate a mapping model specifically for use in a NICE evaluation and where published mapping models are applied within the economic model submitted to NICE. The overarching goal is to ensure that estimates of health benefits that accrue to patients and their carers from the use of health technologies of any type are accurately reflected. Inaccuracy in the estimation of health benefits may mistakenly appear to be quantitatively trivial because of the limited 0 to 1 scale for health utilities. But these apparently small differences can matter greatly for estimates of cost-effectiveness and the decisions that are taken as a result.

2. THE ASSESSMENT OF WHETHER MAPPING IS APPROPRIATE

The motivation for mapping stems from an evidence gap. Whilst cost effectiveness analysis requires health benefits from the technology in question to be quantified in terms of QALYs, there is often an absence or insufficiency of information relating to the health state utilities that permits such an assessment. It is assumed that in all

cases there is evidence on the effect of the technology in some other metric/metrics. The analyst needs to assess the nature of this gap: what are the outcomes on which there is evidence of effect? How are health states to be defined in the economic model? What multi-instrument datasets are available that observe these clinical outcomes and EQ-5D simultaneously?

Clinical studies will typically include a range of clinical measures of disease and disease specific instruments for assessing patient outcomes *inter alia*. It would usually be the case that at least one of these outcomes would align with the way in which health states are defined in the health economic model. The distinction between primary and secondary outcomes is of limited relevance in this context. For mapping to be a relevant tool, or indeed for the use of EQ-5D derived by any method to be considered relevant, there must be a plausible relationship between the clinical measure(s) and EQ-5D. This is often referred to as conceptual overlap. In practice, this usually requires a qualitative judgement of the dimensions being measured by EQ-5D (self-care, anxiety/depression, usual activities, pain, mobility) and the clinical measure in question. In some situations, this might be relatively straightforward. For example, in the case of mapping for rheumatoid arthritis where the Health Assessment Questionnaire Disability index (HAQ-DI) is typically used to define health states in economic models[11] the overlap with EQ-5D dimensions is obvious for several domains. HAQ-DI assesses function disability including usual activities, self-care, mobility and pain (in a separately scored Visual Analogue Scale). Overlap in other disease areas may be less obvious of questionable. For example, in many economic models of treatments for visual disorders, health states are defined by visual acuity[12] which does not directly relate to any dimension on EQ-5D but may be expected to link indirectly to several (mobility, usual activities for example). In this situation, the assessment of

conceptual overlap can be complemented with empirical evidence that considers the psychometric properties of EQ-5D. Mapping to EQ-5D is not appropriate if EQ-5D is considered inappropriate for measuring health benefits in the condition under consideration. The Methods Guide specifies alternative approaches that can be considered in this case (see section 4.3.12)[4].

3. SELECTION OF DATA FOR MAPPING

Mapping requires an external multi-instrument dataset that records – at the individual patient level – both i) the clinical measure(s) of the impact of a health technology in the economic model and ii) EQ-5D, which allows that impact to be translated into health utility terms. These need to be measured at the same time points. There is no requirement for a randomised study for the purposes of mapping. Clinical trials may offer advantages in terms of data quality and consistency of outcome definitions with the evidence of treatment effect. However, they may be conducted in highly selected populations not covering the entire distribution of relevant characteristics, including disease severity. In this respect, observational studies such as disease registries may offer advantages and tend to include larger number of patients and observations. However, this needs to be assessed on a case by case basis.

The need for data to span the relevant spectrum of disease severity is important to minimise the need for out-of-sample extrapolation of modelled results. The relevant spectrum of disease is that which is represented in the economic model where the mapping will be used and, in many situations, will extend beyond the range observed in clinical trials. Models typically have a lifetime horizon tracking patients across the entire natural history course and incorporating patients that have full response to treatments as well as those that have none. The full characteristics of the dataset need

to be reported, including the range and other aspects of the distribution of relevant covariates (such as disease severity) measured both by EQ-5D and the clinical outcomes. This enables the suitability of the dataset used as a basis for the mapping model to be assessed.

There may be situations where more than one relevant dataset is available. In most situations, uncertainty will be minimised from increased sample sizes so pooling of data should be considered. If the datasets were generated under very different conditions, any pooling of the data should be preceded by careful reflection of the differences between the datasets, the potential consequences of merging them and a justification for doing so. However, there can also be benefits from retaining data for cross checking results of estimated statistical models out-of-sample. More information on the issue of validation is provided below.

4. ESTIMATING THE RELATIONSHIP

4.1. Choice of model

The mapping model gives the estimate of the conditional distribution of health utilities $\boldsymbol{f}(\boldsymbol{y}|\boldsymbol{x})$ where y denotes health utility and x the vector of conditioning variables, including the clinical outcome(s).

There are two broad sets of approaches to mapping. The first, “direct” mapping is a one-step approach which models the health utilities directly. This is potentially simple but estimates a relationship that is only relevant for the value set in question. For the purpose of submissions to NICE, this is not a relevant disadvantage. The second approach is referred to as “response mapping” and is a two-stage procedure. Step one models the responses to the descriptive system of the instrument. In the case of EQ-5D-3L a system of five discrete data models, typically assumed independent,

corresponding to the five dimensions of EQ-5D-3L are estimated. Step two uses the estimated models to predict the conditional distribution of responses for each observation (any set of conditioning variables) and then applies the relevant value set to calculate the expected health utility.

4.2. Direct mapping

EQ-5D utility values exhibit several features that need to be accounted for in statistical analysis when using direct mapping. Most of these features are common to a greater or lesser degree to all preference-based instruments, including EQ-5D-5L, and for EQ-5D values for countries other than the UK.

Health utility data are bounded above at full health (1) and below by the worst state that can be described by the instrument. For EQ-5D-3L this is the state 33333 which has a utility value of -0.594 in the UK value set[2]. There tends to be a large proportion of observations at full health. There is a substantial gap between full health and the next feasible value (0.883). The rest of the distribution tends to be multi modal and skewed. Hernandez et al[13] provide examples from a range of different disease areas illustrating these distributional characteristics.

Taken together, this means that many standard statistical models are ill-suited to this setting. Hernandez et al[6] illustrate the reason why the linear regression, widely used in the mapping literature, is considered inappropriate. There are numerous studies that show how such approaches underestimate health benefits for those in good health and overestimate for those in poor health[4]. This in turn leads to misleading cost effectiveness estimates. The Hernandez et al report also provides evidence of this poor performance and concludes that linear regression is not appropriate for mapping.

The Hernandez et al work provides a series of case studies comparing the performance of various types of more flexible approaches. It found that the performance of different mixture model-based approaches was very good for a range of preference-based measures, including 3-level EQ-5D, but the precise specification of those models is important. In particular, the specification includes choices such as the number of components in the mixture and the inclusion of covariates representing disease severity both within components and as predictors of component membership.

Whilst no recommendations can be given as to the precise form of direct mapping model to be used in all situations, it is important that justification is given for model selection. The distribution of the target EQ-5D values in the mapping dataset should be presented and referred to in guiding this choice.

More flexible model types, particularly mixture model approaches, may require a greater degree of statistical experience and judgment than is available to many analysts. Their use in small datasets may often not be feasible. However, used appropriately and judiciously, evidence suggests they provide a very accurate method for direct mapping.

4.3. Response mapping

Response mapping requires sufficient observations in all response categories. The more categories in the descriptive system, the less likely this condition will be met for a given sample size. 3 level EQ-5D is benefitted in this regard since there are only three levels within each domain. Nevertheless, it is sometimes the case that the approach cannot be used because of the small number or even complete lack of responses at severity level three (extreme problems).

4.4. Inclusion of covariates

In addition to the clinical measures used to define health states in the cost effectiveness model which are to be used to map to EQ-5D, consideration should be given to other covariates. We recommend that age is included. Age has been shown to alter the relationship between outcomes of interest in mapping studies, would be a key characteristic defining patients in the cost effectiveness model and provides greater ability to account for differences between the patient sample in the mapping data and the intended population in the cost effectiveness model (which will vary over time as patients are tracked over long periods). Note that the inclusion of age is recommended because it is often the case that individuals in a given health state defined by the clinical outcomes of interest report different EQ-5D states conditional on their age. This is not the same as the adjustments for age recommended to account for the impact of ageing more generally in cost-effectiveness analysis.

Other candidate covariates should follow conventional good practice for statistical modelling, providing a clear rationale prior to inclusion in the model including how they are reflected in the cost effectiveness analysis.

4.5. Assessment of model performance

Predictions from a mapping, regression model provide the expected utility values conditional on covariates. For most economic models, we would want these predictions to be close to the sample means for patients with the same covariates[6]. For example, we want the model to predict a similar value for patients with one category of visual acuity (matching the health state used in the cost effectiveness model) as observed in the data in which the mapping model has been estimated. Model performance should be assessed with this aim in mind and care needs to be

taken to avoid conflating the distribution of the predicted values with the distribution of the actual data. For example, a scatterplot of the data with the predicted fit superimposed is such a conflation and generally not a helpful way of presenting mapping results. A simple account of this distinction by Wailoo et al[14].

By construction, a prediction has to abstract from purely random (i.e. unpredictable) variations in the data. Predictions will therefore always display less variability than actual data. This is not a problem and does not matter for the purposes of cost effectiveness analysis. Cost effectiveness analysis deals with the performance of a technology in the whole user population and therefore inherently averages out the purely random variations within that population.

The mapping model should be used to identify the range in which the predictions lie both within-sample and using combinations of conditioning variables that may be encountered in the cost-effectiveness analysis. The model should not generate predictions outside the feasible range of EQ-5D-3L. Model types that reflect the underlying distribution of EQ-5D-3L data will not predict outside the feasible range by design.

Summary measures of fit provide limited information and alternative measures can conflict with each other. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are widely reported in mapping studies. These are relatively insensitive, particularly given the limited range of utility data (-0.594 to 1 in the case of EQ-5D-3L). They may also conceal systematic differences in the patterns of prediction errors. Mean Error (ME) is not a useful measure in this context. It is common practice to use penalised likelihood criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to choose between models but again, these may

conflict with each other, and cannot be used to compare models with different forms of the dependent variable. AIC and BIC may help to select between model specifications but must be supplemented with graphical representation of fit across the spectrum of disease severity.

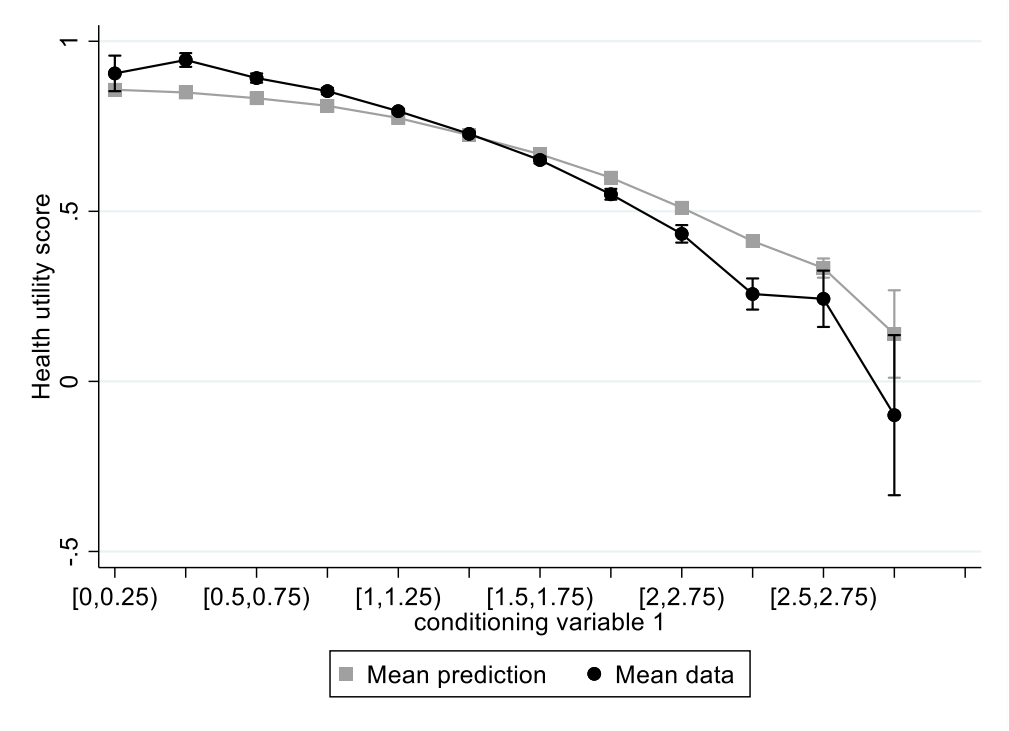
Plots of predicted EQ-5D-3L means versus data group means are very useful to aid understanding of differences between models, and the potential impact of those differences when translated into economic evaluation. These comparisons should be based on groupings of the data according to at least one of the conditioning variables (not the dependent EQ-5D-3L). Where more than one clinical outcome measure is used as a conditioning variable it is helpful to provide these plots by all the clinical outcomes. This can be seen, for example, in a mapping study of outcomes for patients with axial spondyloarthitis [15] where measures of both functional disability (Bath Ankylosing Spondylitis Disease Activity Index – BASDAI) and disease activity (Ankylosing Spondylitis Disease Activity Score - ASDAS) are used in the mapping model as conditioning variables and reported in a series of plots.

These simple plots are very useful in identifying areas of poor fit across the full range of the conditioning variables or any systematic patterns in the conditional means that could signal model misspecification. Discovery of any issues allows the analyst to investigate them fully and solve potential model misspecification problems. A simple example is shown here. Two separate models were estimated on the same dataset. Figure 1 presents the plots corresponding to models 1 and 2. The dark coloured dots linked with a black line are the EQ-5D-3L means of the data in the ranges represented on the x-axis. The grey squares linked with a grey line shows the model predicted conditional means for the same groups of the conditioning variable. Panels a) and b) plot group averages on the first conditioning variable comparing the data group means

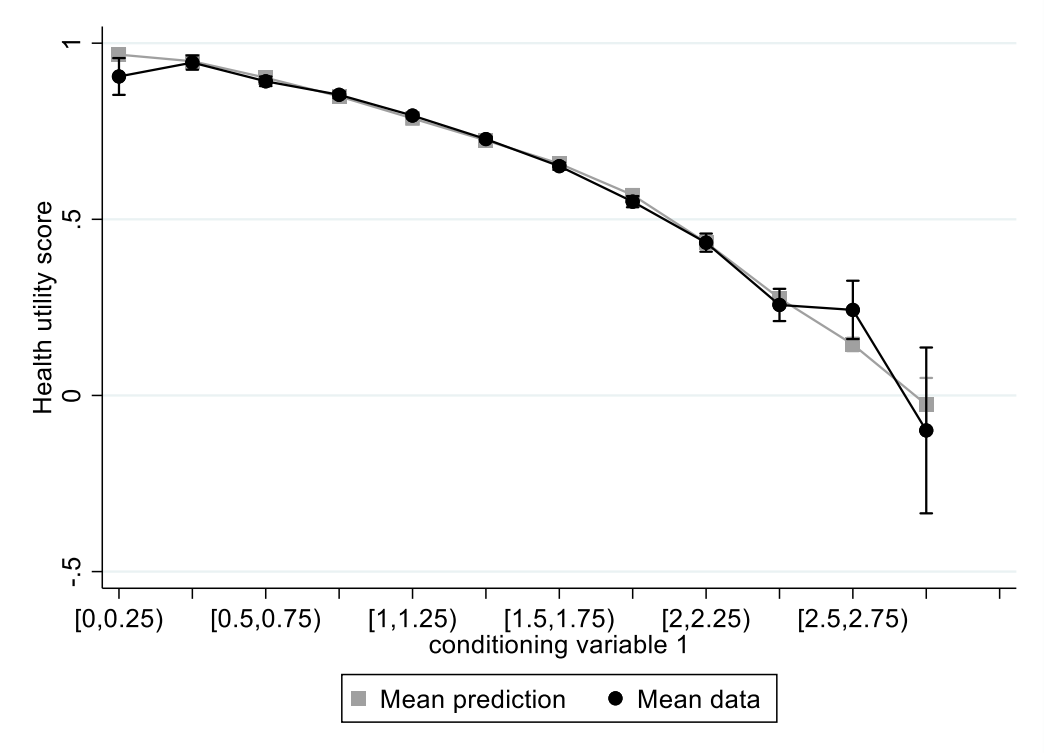
and the models' predicted means. It is clear that model 2 is better able to reproduce the group means in the data. Furthermore, there is a clear pattern in the means of model 1 with systematic underprediction and overprediction at low and high levels of the conditioning variable respectively. This is often a sign of misspecification in the model and the causes of those systematic patterns should be investigated. This is especially important in cases such as this where the 95% confidence intervals are quite small (apart from the extremes of the conditioning variable). Panels c) and d) present the same graphs but for a second conditioning variable of interest. For this second variable, there is evidence that neither of the models can capture adequately the relationship at higher levels of the conditioning variable and this is an issue that needs to be further investigated.

Figure 1: Mean observed and modelled health utility scores by conditioning variable

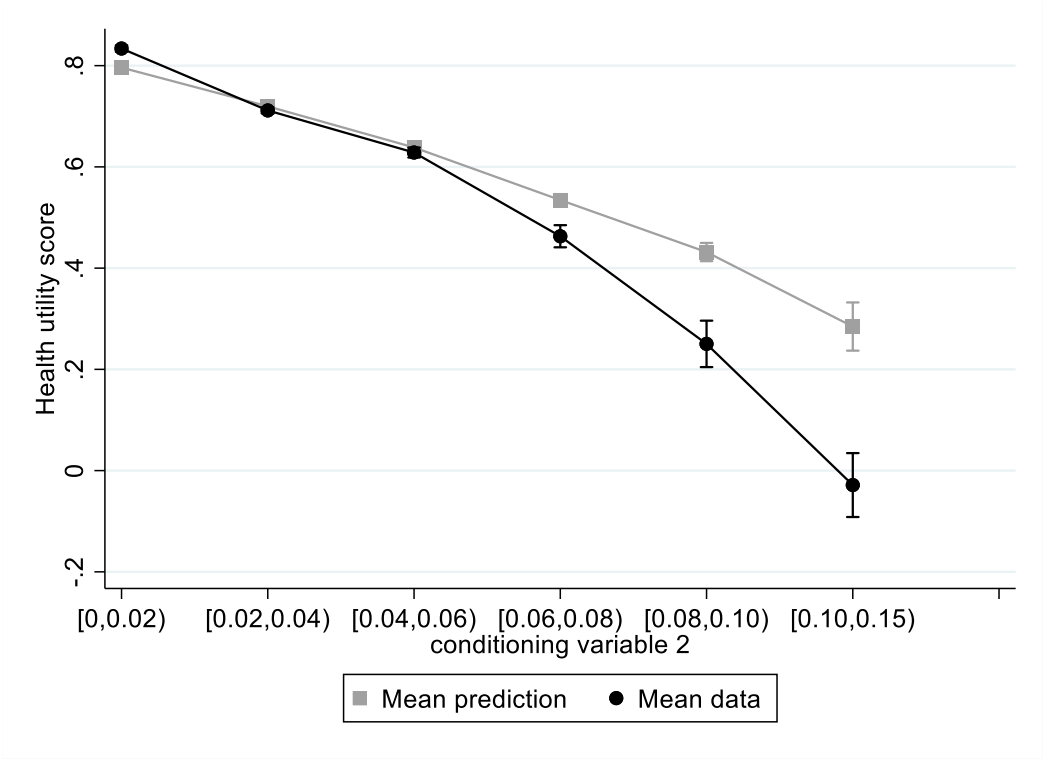
a) Model 1



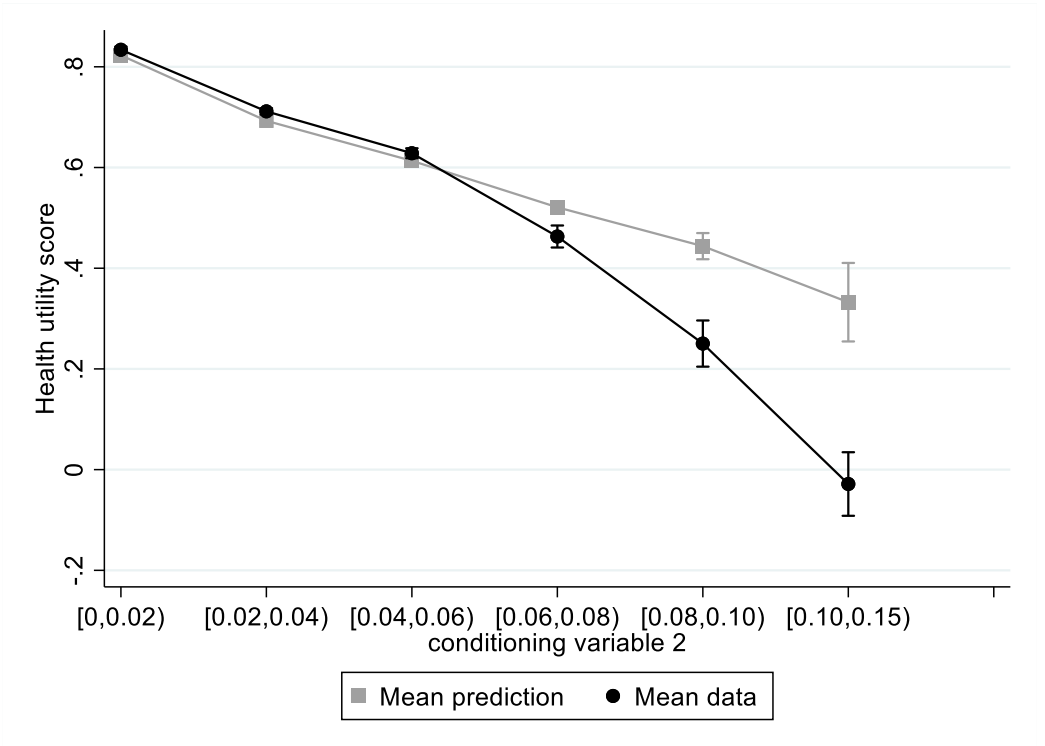
b) Model 2



c) Model 1


Health utility score
-.2
0
.2
.4
.6
.8
[0,0.02)
[0.02,0.04)
[0.04,0.06)
[0.06,0.08)
[0.08,0.10)
[0.10,0.15)
conditioning variable 2
Mean prediction
Mean data

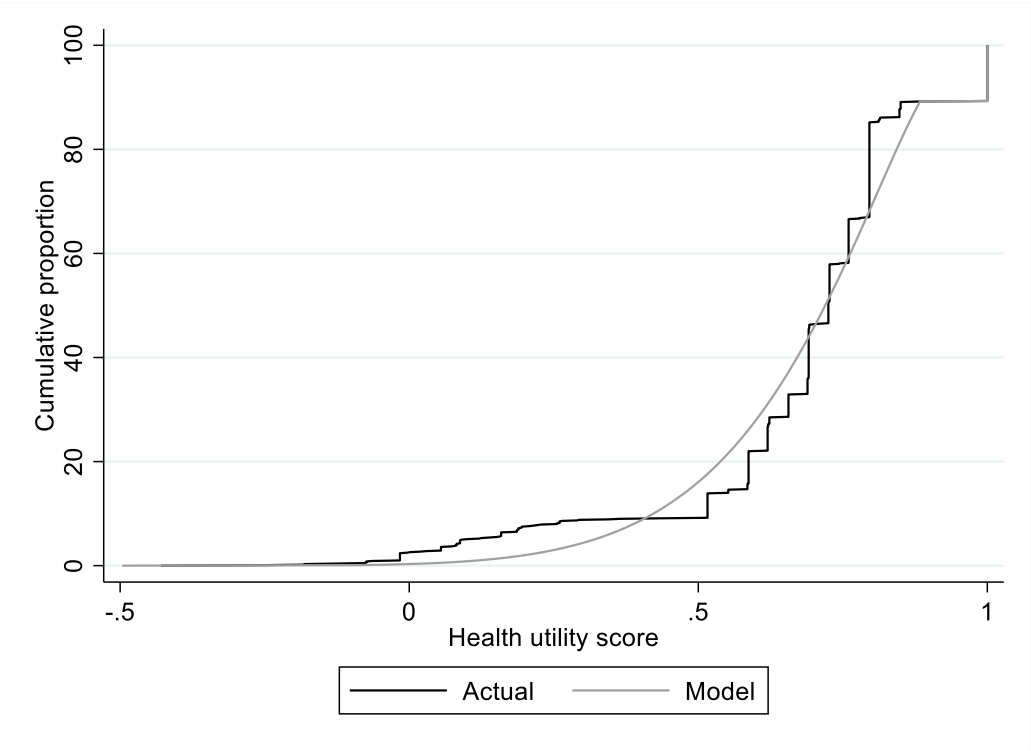
d) Model 2


Health utility score
-.2
0
.2
.4
.6
.8
[0,0.02)
[0.02,0.04)
[0.04,0.06)
[0.06,0.08)
[0.08,0.10)
[0.10,0.15)
conditioning variable 2
Mean prediction
Mean data

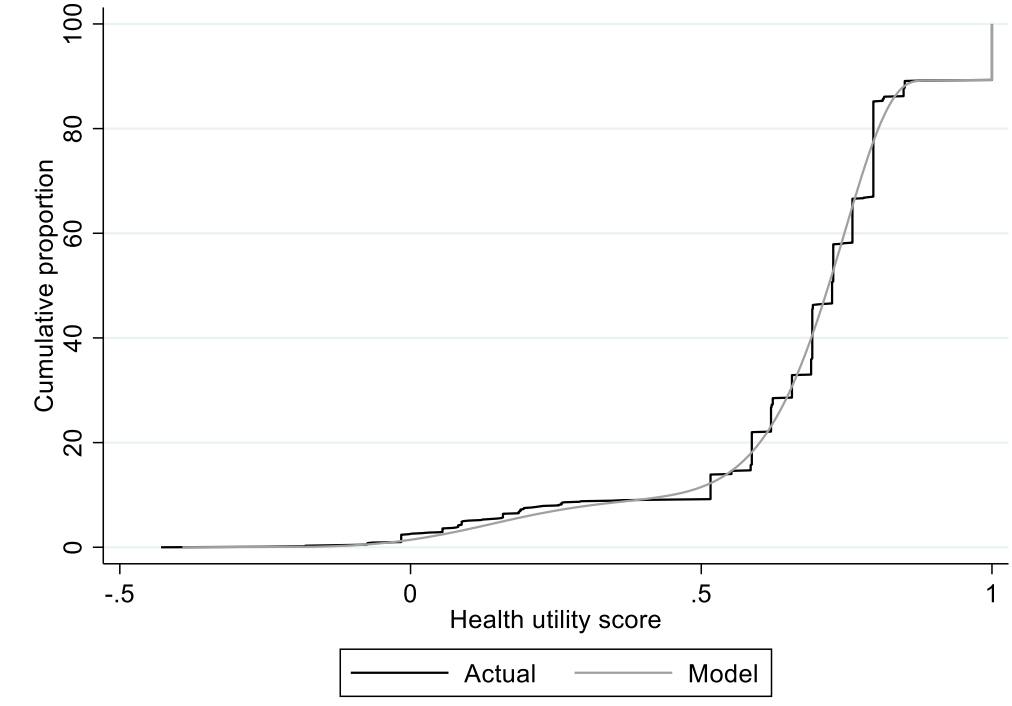
A small number of cost effectiveness models simulate individual patients rather than cohorts of patients[16,17], and attempt to represent heterogeneity between those individuals in order to generate accurate estimates of costs and effects. In these cases, the mapping model needs to be able to reproduce the full distribution of health utilities. In this situation, the relevant comparison for model performance is the distribution of the observed data with the distribution of the data that would have been generated by the mapping model. Even in those situations where the intended use of the mapping model is in a cohort type model, these comparisons provide further insight into model performance and appropriateness generally. Plots of the cumulative proportion of the data versus the data generated from the model are recommended. These types of plots are also useful in signalling any model misspecification. Figure 2 below presents these plots for the two illustrative models underlying Figure 1. The plot for model 1 gives more evidence indicating misspecification of the estimated model and its inability to capture the shape of the EQ-5D-3L distribution in the data. Model 2 shows much better alignment with the data although it does not seem to fully capture the shape of the distribution in the left-hand tail (i.e. very poor health states). This graph, together with the evidence presented on the group means above should prompt the analyst to consider possible reasons for the patterns observed and make a judgement as to whether additional model specification changes might be required.

Figure 2: Cumulative proportions for sample data and predictions

a) Model 1



b) Model 2



4.6. Out of Sample testing

In addition to measuring how well a mapping model performs in the dataset in which it was estimated ("within-sample fit"), it is sometimes feasible to conduct further checks out of sample. Some authors recommend this be performed routinely[8]. In some mapping studies this is achieved by splitting the available dataset and using part of the data for the initial estimation and the remainder for out-of-sample comparisons of fit. Depending on how the sample split is made, these out-of-sample comparisons may provide some additional information, but their limitations also need to be acknowledged. In many situations, analysts do not have the luxury of multiple, large sample datasets. Reserving data for validation has a cost to the sample size of the data available for estimation and this increases parameter uncertainty. Judgements need to be made on a case-by-case basis as to whether these costs are worth incurring for the additional value from out-of-sample testing. Of course, any model will perform less well out-of-sample and any observed differences may be due to the characteristics of the datasets rather than a finding that genuinely undermines the validity of the estimated mapping model.

If there is a single large multi-instrument dataset which is to be split into a within-sample subset for model estimation and an out-of-sample subset for assessment of predictions, the most important consideration is the way that the split is made. A purely random division of the data will have purely random differences in sample characteristics between the within and out of sample subsets, and will therefore have very little power to detect the systematic prediction errors that are caused by model misspecification. On the other hand, if the sample is split by (say) age, there will be

good power to detect misspecifications in the way that age effects are modelled, but possibly not other effects. Moreover, systematically excluding part of the range of an explanatory variable in this way limits the scope for exploring more flexible specifications.

5. USING THE RESULTS IN A COST EFFECTIVENESS MODEL

Methods for assessing the appropriateness of mapping models are designed to reflect their intended use in cost effectiveness models. There is therefore overlap with the analyses performed to demonstrate performance of mapping models described in section 4.

For most cohort models, the need is to generate an estimate with associated uncertainty, of the conditional mean EQ-5D-3L for the health states and patient population, as defined in the cost effectiveness model. At a minimum, the inclusion of age in the mapping model would be expected. In many situations, mapping studies may be undertaken entirely separate from any NICE appraisal, published and then subsequently used by analysts responsible for the cost-effectiveness analysis NICE receives. To aid those analysts, the reporting of mapping models must provide a detailed unambiguous description of the model allowing predictions to be calculated from the reported details, should provide some example predictions for a set of covariates, and should also consider providing pre-programmed software generating predictions for the preferred mapping model.

5.1. Reflecting parameter uncertainty

Mapping provides inputs to cost-effectiveness analysis in the form of a statistical model and is equivalent in this regard to all other statistical models commonly

encountered (for example in cost regressions or survival analysis). Probabilistic Sensitivity Analysis (PSA) is a standard method for reflecting parameter uncertainty by sampling from their joint probability distributions. The reporting of the variance – covariance matrix is essential information that allows analysts to perform PSA for these correlated parameters. Where more flexible models have been used, it can be helpful to cost effectiveness analysts if pre-programmed calculators using standard software are provided which generate both predictions and perform sampling of this type.

5.2. Reflecting patient variability

There are circumstances where there is a need to generate simulated data, rather than predictions, from the mapping model. Most decision models are cohort based, they reflect the course of disease and its impact for a hypothetical group of people. The health utility value to be applied for these groups should be the conditional mean relating to the relevant health state. There are some cases where this is not the case. For example, where the decision model is a genuine patient level simulation model, analysts may wish to impute health utility values for each individual patient via simulation, based on the mapping model and the patient characteristics. This would reflect the individual level differences in patient responses to the EQ-5D descriptive system, that is, reflecting heterogeneity between patients. This may also be the case if the mapping is being used to generate data at the individual patient level, for example as part of an economic evaluation alongside a clinical trial . In this situation, the procedure is equivalent to that used to generate the cumulative distribution plots described in section 4 (see Figure 2). Reporting of the specification of the mapping model, including distributional assumptions about the error terms and the estimated

parameters that shape these distributions, allows these calculations to be made by anyone.

If an appropriate mapping model type has been used then neither sampled values for PSA nor simulated data generated to reflect patient variability will lie outside the feasible range of EQ-5D-3L (between 1 and -0.594). Other model types may generate values outside this range, and have been found to do so in practice. For example, Lamu and Olsen found that linear regression generated predictions in excess of 1 when mapping from the QLQ-C30 using data from a sample of cancer patients[18]. If such models have been used, then the reporting of results and/or the use of the mapping model in a NICE assessment should report the frequency with which observations outside the feasible range occurred (particularly values above 1) and what the analyst did on those occasions (e.g. were they left unaltered, artificially censored, or some other procedure?).

5.3. Extrapolation of mapping results

It is common for there to be an evidence gap when assessing health technologies in one patient population that could be addressed by a mapping that has already been performed in a different patient population. This occurs because the way in which disease severity is described and measured is often common across a broad range of conditions. Examples can be found across cancers, where measures such as the European Organisation for Research and Treatment of Cancer (EORTC QLQ30) patient reported outcome is widely used in clinical studies of patients with any type of tumour. Other examples can be found in the use of measures of visual acuity or vision specific patient reported outcomes (such as the Visual Functioning Questionnaire) for defining health states in decision models for various eye conditions but with mapping

models being available from more common eye conditions such as macular oedema[19] and macular degeneration[20].

It may also be the case that the existing mapping offers some advantages over studies that have been or could be performed in the population of interest. This is particularly the case in rare conditions where limited patient numbers can be a substantial impediment. However, judgements about generalisability also need to consider the nature of differences between populations and conditions, based on patient and clinical insight.

6. SPECIAL CASE OF MAPPING BETWEEN VARIANTS OF EQ5D.

The NICE methods guide[4] calls for the use the EQ-5D-3L value set in the reference case of health technology submissions. If the data was collected using the newer EQ-5D-5L, these data should be mapped to the EQ-5D-3L value set using the mapping function developed by the Decision Support Unit[21] based on the EEPRU dataset collected for this purpose[22].

The specific mapping procedure to follow depends on the type of information available to the analyst. Sometimes, individual patient level data on EQ-5D-5L is available either as a health state description (the responses to the 5 dimensions) or as a utility value. In other cases, the analyst only has access to a mean EQ-5D-5L utility value for the patient population of interest.

If individual patient level data is available, the correct mapping procedure involves mapping to the EQ-5D-3L utility for each individual patient. An unbiased estimate of the average EQ-5D-3L utility value for the patient population can then be calculated as the average across the predictions of all patients in the dataset. Mapping from the health descriptor is always preferred to mapping from the EQ-5D-5L utility value as

the latter involves some loss of information because there is not a one-to-one correspondence between health descriptors and utility values in the EQ-5D-5L value set. A utility value could represent more than one health state and this additional uncertainty needs to be taken into account when mapping. This case is handled in the same way as when mapping from a mean EQ-5D-5L utility value and is described in detail below.

When the analyst only has access to a mean EQ-5D-5L utility value for the patient population of interest, the mapping procedure should take into account that the EQ-5D-5L value is not a valid point in the EQ-5D-5L value set. In this case, mapping should be done using distance-weighted averaging of the utility values within a neighbourhood of the mean EQ-5D-5L utility. In effect, all utility values within a specified distance of the mean EQ-5D-5L utility we want to map from are chosen and individually mapped to EQ-5D-3L. An estimate of the average EQ-5D-3L utility value for the patient population can then be calculated by using a weighted average across all these mapped values. The size of the neighbourhood is controlled by an additional parameter, the bandwidth, which the analyst must choose. The bandwidth also controls the rate at which the weight declines with increasing distance, decreasing the contribution of utility values that are farther away from the mean EQ-5D-5L utility value. Given the structure of the EQ-5D-5L value set for England currently available[23] this procedure can be sensitive to the choice of bandwidth. A small bandwidth for an EQ-5D-5L mean value which lies in an area where there are large gaps between adjacent utility values could result in no or too few values within the neighbourhood. The same bandwidth in an area where utility values are closer together might be perfectly appropriate. Therefore, the current recommendations for the choice of bandwidth when mapping from EQ-5D-5L to EQ-5D-3L are as follows:

- if *x*, the mean EQ-5D-5L, is in the interval $1 \leq x > 0.951$, a small bandwidth, large enough to include the utility value of 1 (perfect health) at the top should be used
- if *x* is in the interval $0.951 \geq x \leq 0.800$, a bandwidth no larger than 0.1 with 0.1 as the default value should be used.
- if $x < 0.800$ a larger bandwidth of 0.2 should be used

Note that the bandwidth recommendations to map in the opposite direction, from EQ-5D-3L to EQ-5D-5L, are similar but the largest bandwidth is larger allowing for the larger range of EQ-5D-3L utilities

For the case where patient level data is available but only as utility values (not health states) the procedure is the same but the appropriate bandwidth differs from the recommendations above in one crucial respect. When mapping from an average utility value, it is important to choose a bandwidth that includes neighbouring utility values. When mapping from an individual patient utility value, it is important that the bandwidth is small enough not to include other neighbouring utility values.

The size of the bandwidth will depend on the precision of the utility value set. For example, a bandwidth of 0.0009 or slightly smaller can be used for value sets using three decimal points of accuracy. In this way, only utility values that are the same but represent different health states will be used by the program to predict EQ-5D-3L. As all points have the same value and the same distance from the EQ-5D-5L utility value, the procedure simplifies to equal weights in this case.

Stata, Excel and R commands implementing this mapping and instructions on how to use them have been developed by the DSU and are available to download from the following website:-

https://www.sheffield.ac.uk/nice-dsu/methods-development/mapping-eq-5d-5l-3l.

The Stata command is also available to download from within Stata as an updated version of the original command. The correct command to download is the updated version published in the Stata Journal[24].

7. SUMMARY OF RECOMMENDATIONS

How can we estimate what the health utility value for a health state relevant to a cost effectiveness model is when existing data may be limited in some way? Mapping offers one solution to this challenge by estimating the relationship between some set of variables, including one or more measures used to define the status of patients in the cost effectiveness model, and health utility. Appropriate data, analysis methods, reporting standards and use of the outputs in the cost-effectiveness model are all required for mapping to provide a suitable option for NICE appraisals. The following recommendations relate to those areas:

1) Consider how health states are to be defined in the economic analysis and how these relate to the dimensions of health contained within the target health utility measure. If there is a fundamental mismatch in the concepts being measured then mapping, like other methods, will not be appropriate.
2) Assess available datasets that include simultaneous assessments of conditioning variables and the target health utility measure. Select datasets for analysis that minimise the need for extrapolation of results outside the sample population. Report the distribution of all variables considered for the analysis and contrast this with the characteristics of patients in the economic model.

3) Statistical modelling methods should be selected with an awareness of the challenging distribution of outcomes like UK EQ-5D health utilities. Model types should be selected consistent with these features. Inappropriate statistical models lead to biased estimates of health gain, the generation of unfeasible values and a misrepresentation of uncertainty. Linear regression is not appropriate in this setting for these reasons.
4) Age should feature as a conditioning variable in mapping models using data for adult patients, and evidence provided to justify its exclusion from the preferred model. This does not negate the need to make separate adjustments in the cost effectiveness model to account for the impact of ageing.
5) Selection of the preferred model specification, and assessment of performance, should not be based only on summary measures of fit. Plots of the predicted EQ-5D means versus data group means should be presented. Groups should be defined by at least one of the clinical conditioning variables (not the dependent EQ-5D). Where more than one clinical outcome measure is used as a conditioning variable it is helpful to provide these plots by all the clinical outcomes.
6) Plots of the cumulative proportion that lies within each EQ-5D value in the data versus the data generated from the model are recommended. These types of plots are useful in signalling any model misspecification.
7) The preferred mapping model needs to be described unambiguously in published papers or reports to allow analysts to generate predictions. Pre-programmed software commands are an efficient method to assist analyst using the results of mapping model, particularly if more complex, flexible methods are used.

8) Standard methods for reflecting uncertainty using PSA are relevant to mapping regression models.
9) There are many instances of confusion between the predictions that arise from mapping models (conditional means) and raw data, which includes individual level variability. Clarity on this issue is required to ensure model performance is assessed appropriately, that the results of mapping models are used correctly in cohort models and that variability is reflected when using mapping models with individual level analyses.
10) If a mapping model type has been used to inform cost-effectiveness analysis that does not restrict output to the feasible range for health utilities by design, then any predictions or PSA samples outside that range need to be reported.

REFERENCES

[1] Brooks R. EuroQol: the current state of play. Health Policy. 1996 Jul;37(1):53-72

[2] Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997 Nov;35(11):1095-108.

[3] Mortimer D, Segal L, Sturm J. Can we derive an "exchange rate" between descriptive and preference-based outcome measures for stroke? Results from the transfer to utility (TTU) technique. Health Qual Life Outcomes 2009;7:33

[4] NICE (2022) "NICE health technology evaluations: the manual". Available at https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741 (last accessed 22nd April 2022)

[5] Longworth, L. and Rowen, D. (2011) "NICE DSU Technical Support Document 10: The Use Of Mapping Methods to Estimate Health State Utility Values", available at https://www.sheffield.ac.uk/sites/default/files/2022-02/TSD10-mapping-FINAL.pdf (last accessed 15th March 2022)

[6] Hernández Alava M, Wailoo A, Pudney S, Gray L & Manca A. (2020) Mapping clinical outcomes to generic preference-based outcome measures: development and comparison of methods. Health Technol Assess Vol.24(34)

[7] Wailoo AJ, Hernandez Alava M, Manca A, Mejia A, Ray J, Crawford B, Botteman M, Busschbach J. (2017) "Use of Mapping to Estimate Utility Values from Non-Preference-Based Outcome Measures for Cost per QALY Economic Analysis : Good Research Practices Task Force", Value in Health, Vol.20:18-27

[8] Petrou S, Rivero-Arias O, Dakin H, Longworth L, Oppe M, Froud R, Gray A. The MAPS Reporting Statement for Studies Mapping onto Generic Preference-Based Outcome Measures: Explanation and Elaboration. Pharmacoeconomics. 2015 Oct;33(10):993-1011. doi: 10.1007/s40273-015-0312-9. PMID: 26232200.

[9] Dakin, H, Abel, L, Burns, R, Yang, Y. 2018. Review and critical appraisal of studies mapping from quality of life or clinical measures to EQ-5D: an online database and application of the MAPS statement. Health and Quality of Life Outcomes. 16:31.

[10] Mukuria C, Rowen D, Harnan S, Rawdin A, Wong R, Ara R & Brazier J (2019) An Updated Systematic Review of Studies Mapping (or Cross-Walking) Measures of

Health-Related Quality of Life to Generic Preference-Based Measures to Generate Utility Values. Applied Health Economics and Health Policy, 17(3), 295-313.

[11] Jansen JP, Incerti D, Mutebi A, Peneva D, MacEwan JP, Stolshek B, Kaur P, Gharaibeh M, Strand V. Cost-effectiveness of sequenced treatment of rheumatoid arthritis with targeted immune modulators. J Med Econ. 2017 Jul;20(7):703-714. doi: 10.1080/13696998.2017.1307205. Epub 2017 Apr 5. PMID: 28294642.

[12] Butt T, Lee A, Lee C on behalf of the UK AMD EMR Study Group, et al The cost-effectiveness of initiating ranibizumab therapy in eyes with neovascular AMD with good vision: an economic model using real-world outcomes BMJ Open 2015;5:e006535. doi: 10.1136/bmjopen-2014-006535

[13] Hernandez Alava, M., Wailoo, AJ, and Ara, R. (2012) "Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Health State Utility Values", Value in Health, doi:10.1016/j.jval.2011.12.014.

[14] HEDS Blog "Why does my mapping not predict health utilities at 1?" Available at http://scharrheds.blogspot.com/2019/01/why-does-my-mapping-not-predict-health.html (last accessed 16th March 2022)

[15] Hernandez Alava M, Wailoo A, Chrysanthou G, et al. (2022) "Measuring Quality of Life of Patients with Axial Spondyloarthritis for Economic Evaluation", BMJ RMD Open RMD Open 2022;8:e001955. doi:10.1136/rmdopen-2021-001955

[16] Stevenson M, Lloyd Jones M, De Nigris E, Brewer N, Davis S and Oakley J. A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis. Health Technol Assess 2005; 9 (22) pp1-160

[17] Youn J, Stevenson MD, Thokala P, Payne K, Goddard M. Modeling the Economic Impact of Interventions for Older Populations with Multimorbidity: A Method of Linking Multiple Single-Disease Models. Med Decis Making. 2019;39(7):842-856

[18] Lamu, A and Olsen, J. (2018) "Testing alternative regression models to predict utilities: mapping the QLQ-C30 onto the EQ-5D-5L and the SF-6D", Quality of Life Research, Vol.27: 2823–2839

[19] Pennington R, Hernandez Alava M, Hykin P, al (2020) "Mapping From Visual Acuity to EQ-5D, EQ-5D With Vision Bolt-On, and VFQ-UI in Patients With Macular Edema in the LEAVO Trial", Value in Health, Vol. 23(7):928–935

[20] Kay, S., & Ferreira, A. (2014). Mapping the 25-item National Eye Institute Visual Functioning Questionnaire (NEI VFQ-25) to EQ-5D utility scores. Ophthalmic Epidemiol, 21(2), 66-78.

[21] Hernández Alava M, Pudney S. (2017) Econometric modelling of multiple self-reports of health states: The switch from EQ-5D-3L to EQ-5D-5L in evaluating drug therapies for rheumatoid arthritis. Journal of Health Economics, 55:139-152

[22] Hernandez Alava, M., Pudney, S., and Wailoo, A. (2020) Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study. Policy Research Unit in Economic Evaluation of Health and Care Interventions. Universities of Sheffield and York. Report 063

[23] Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 2018 Jan;27(1):7-22. doi: 10.1002/hec.3564. Epub 2017 Aug 22. PMID: 28833869; PMCID: PMC6680214.

[24] Hernández-Alava M., and Pudney S. (2022) “eq5dmap: A command for mapping between EQ-5D-3L and EQ-5D-5L.” The Stata Journal, Vol.22, Number 1, pp. 238. Software update st0528_1