Capturing information loss in estimates of uncertainty that arise from mapping algorithms

Jeff Round

Leeds Institute of Health Sciences
Academic Unit of Health Economics
University of Leeds
j.a.round@leeds.ac.uk

Background

1 Mapping between health-related quality of life instruments is an increasingly common technique for estimating utility scores for use in economic evaluations. Mapping refers to the process of estimating utility values for use in an economic evaluation through predicting results for a target preference based measure from data collected using a non-preference based measure. This process, if possible and practical, is assumed to allow economists to include information on preferences for health states within economic evaluation where direct evaluation is not possible, thereby improving resource allocation decisions. The use of mapping algorithms is increasing in the literature and is set to grow further as the National Institute for Health and Clinical Excellence (NICE) has proposed the use of mapped estimates as a legitimate approach to estimating utility values in the recently published methodology guidelines for technology appraisals[1]. Despite increasing levels of interest in this area, the use of such approaches has not yet been fully theoretically justified

2 A large number of non-preference based health-related quality of life instruments have been and continue to be developed for use in health research, as have a much smaller number of preference based instruments. Those non-preference based instruments represent a rich source of information about the quality of life of the studied populations. However because no set of preference weights exists for these measures they have limited use in allocating scarce healthcare resources, both between treatments for the same or similar conditions but particularly across different areas of healthcare. The development of a statistical model that would allow researchers to translate the information gained from a non-preference based instrument into utility scores for use in economic evaluations could be very useful in a variety of scenarios. For instance such mapped estimates might allow historical data to be used to evaluate technologies that are already in use and for which it is unlikely to collect new data.

3 Current methods of mapping have focused on the accuracy with which the results of one instrument can be predicted based on responses to a second instrument. A number of studies have been conducted to estimate utility scores from non-preference based health related quality of life instruments. These studies have shown that under a given set of circumstances it is sometimes, though not always, possible to estimate a function that allows for a consistent, or reliable, estimation of utility. The typical approach is to use one of a range of regression approaches where the emphasis is placed on finding the model that provides the greatest explanation of the variance in the results found on each measure. By finding a model that best explains the variance in the results of different instruments, or that best predicts the actual results obtained from mapping, it is often assumed that this is therefore the model that leads to the greatest reduction in uncertainty in the final mapped estimates.

4 However, current methods ignore an important conceptual issue that arises from mapping. At the crux of the mapping problem is the fact that just because one instrument can be used to predict the scores of another this does not mean that each instrument has measured the same preference for health. In this paper I refer to this problem as the uncertainty arising from information loss.

5 To understand how uncertainty is generated by information loss it is first necessary to understand the principle reasons for developing health related quality of life instruments. Instruments that measure health related quality of life provide information about how some healthcare intervention impacts on an individual's life beyond the standard assessments of clinical outcomes. Some instruments such as the EQ-5D or SF-6D have been developed to give a broad picture of overall health but as a result are less sensitive to smaller changes in specific aspects of health. Other condition specific instruments such as the Asthma Quality of Life Questionnaire (AQoL) measure different domains of health than generic instruments or measure the same domains at a different level of detail. These measures are often designed to be more sensitive to changes in health that are specifically a result of that condition, though they may not be a good indicator of overall health or preferences for health. It is these differing objectives in instrument design that leads to the information loss in mapping.

6 This information loss happens at two levels. The first is at the level of the domains of health measured and the second is at the level of individual item responses. The first source of information loss can demonstrated using the EQ-5D, commonly used as the preference based target instrument in mapping studies (see Table 1), which measures

five domains of health - mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Alternative instruments, which may be either generic or condition specific, may not measure these same five domains, and yet are still mapped to the EQ-5D preference set. For example the AQoL measures just four domains of health considered to be relevant to patients with asthma - symptoms, activity limitation, emotional function and environmental stimuli. While on the face of it there is some overlap with the EQ-5D it is clear that the two instruments do not measure exactly the same aspects of health as each other. This difference means that when mapping from the AQoL to the EQ-5D some information will not be included in the estimates, leading to potentially increased uncertainty in the estimate.

7 In a similar fashion, there is a further potential for the loss of information even when the same domains are being measured. There may be clear overlap between many different instruments in the case of some domains. For example, the SF-12 and EQ-5D both contain domains on mental health (in the SF-12 this is the mental health domain, in the EQ-5D it is the anxiety/depression domain) but they give slightly different amounts of information. The SF-12 includes two questions on mental health compared to just the one in the EQ-5D, potentially providing more information. Mapping directly between these two domains will lead to a loss of some of this information.

8 Where researchers have attempted to establish a predictive statistical relationship between the results of two different measures of quality of life, I argue that they have not yet considered in sufficient detail whether or not the results can be relied upon to tell us anything meaningful. It is not enough to say, for instance, that the scores of Measure A are highly correlated with the scores of Measure B, if we cannot say with confidence that A and B are measuring the same thing. As shown, the loss of information that may occur when mapping needs also to be considered. Before any mapping exercise is undertaken it is necessary to consider in some fashion how to determine if mapping is an appropriate approach to the problem. If it is considered appropriate, it must also be established to what extent the results can be relied upon to give an accurate answer to the original resource allocation decision. This is particularly true when the objective of the mapping exercise is to estimate utility scores to be used in economic analysis to inform resource allocation decisions. It is now largely the case that the key theoretical difficulties with mapping lie not in generating reliable predictive relationships; a number of methods for this have been suggested and in many cases have been argued to be a reliable alternative to indirect valuation using preference

based instruments or other direct methods (Table 1). Whether or not one of these methods is the 'best' method remains unclear but is not the focus of this review. Nevertheless, estimating a reliable mapping function using an appropriate method is a necessary condition for the process to be considered valid, but it is not sufficient.

9 The remainder of this study considers the ways in which previous researchers have considered the implications of the loss of information that arise from mapping, as well as ways in which to assess whether or not estimates of cost-effectiveness based on mapped estimates can be considered valid. The first item is addressed through a systematic review of previous studies that have generated mapped estimates using non-preference based instruments to predict results using preference based measures. Using the results of the review I then consider the implications of information loss in mapping and the consequences that mapping has for the validity of decision making in the context of healthcare resource allocation problems.

Methods

10 A systematic review has been undertaken to examine the extent to which authors of previous mapping studies have considered the problem of uncertainty arising from the loss of information that occurs when moving between instruments. Search terms used in the review were selected to obtain the greatest possible sensitivity given the expected relatively small number of published estimates of mapping algorithms. Searches have been run using the OvidSP interface for MEDLINE and EMBASE interface. A full list of search terms used is available from the author on request.

11 In order for a study to be included in the full review it had to satisfy the following criteria:

a. The study must include data from a non-preference based measure of health that is either:

   i. A subjective measure of quality of life as defined by the authors, such as the SF-12 or the Child Health Questionnaire, or

   ii. An objective measure of health or other clinical indicator such as blood pressure or lung capacity

b. The study must have also included data from a preference based measure of quality of life such as the EQ-5D

c. The study must have attempted to predict the results of the preference based measure of quality of life using the information from the non-preference based measure

12 By extension, the following common types of studies, examples of which were identified in the literature search, are excluded:

d. Valuation studies designed to derive preference sets for non-preference based instruments using standard gamble or time-trade off methods.

e. Comparisons of two or more preference based measures where there is no attempt to derive a mapping function for a non-preference based measure.

13 Searches of MEDLINE and EMBASE identified 160 articles after duplicates were removed. An initial sift of articles by title and abstract led to 34 articles being retrieved for more detailed appraisal. Twenty-four of these articles have been judged to have met the inclusion criteria and have been reviewed. Two recent reviews of the mapping literature have also been identified as part of this work, one journal article by Mortimer and Segal[2] and one discussion paper by Brazier et al[3]. In addition to the articles identified as part of the search strategy, all articles included in these two reviews were evaluated for possible inclusion in this review.

14 Both of the previous reviews addressed a different question than the review undertaken here, focusing on the different methods that have been used by researchers to develop the algorithms used for mapping. The review by Mortimer and Segal[2] identified 46 examples of studies that attempt to derive utility scores from non-preference based measures of health, including those that attempt to map directly between non-preference and preference based instruments. The authors of the review concluded that the choice of algorithm can have an impact on whether or nota treatment receives funding, but that this will also be influenced by the initial choice of instruments and the characteristics of the patient population being studied. In their own words 'variation between individual algorithms appears to be more important in explaining variation in predicted QALY-weights than the choice of derivation technique'. The review did not address how to deal with information lost when mapping between instruments.

15 The review by Brazier et al[3] was more focused in scope, focusing on those studies that map from a non-preference based to a preference based instrument. Twenty-eight studies were included in the review and 27 of these mapped from a non-preference to preference based instrument, with one mapping from a preference based to a

preference based instrument. Amongst other conclusions the review found that the performance of a mapping function was often dependent on the degree of overlap between the domains being measured by each instrument. It was noted that this is more likely to be the case when mapping from condition specific measures to generic measures. Although alternative approaches to mapping are suggested there was no discussion of how to capture the uncertainty that arises from this loss of information in those cases where the development of a preference based index is not possible.

Results

16 Of the 17 studies included in the full review to date none have explicitly considered the uncertainty that arises in mapping as a result of losing information when moving between instruments. All studies did however consider in detail the best way to minimise uncertainty through finding the best predictive mapping model. The approaches used in these studies are given in Table 1. However, as the review was not designed to establish the best approach to mapping no further comment or analysis is undertaken here.

17 The most reliable conclusion that may be drawn from this review is that the problem of losing information during the mapping process has not yet been fully addressed in the literature, despite its importance in understanding the uncertainty involved in mapping. Only a very small number of the studies explicitly recognise the problems that arise specifically relating to the fact that the two instruments being mapped may not be measuring the same domains of health. And within this small number of studies, none suggests methods for how to address this problem when generating mapping algorithms. Two other general conclusions may also be drawn based on the evidence gathered as part of this review.

18 First, many studies use a condition specific measure as the source instrument which is then mapped to a preference based generic measures. Condition specific measures are often narrower in scope by design than generic preference based instruments, meaning that more information may be collected but about a narrower set of domains of health. This additional information will be lost when mapping between instruments. In addition, potentially relevant information will be missed relating to those domains of the preference based instrument that are not captured in the non-preference based instrument. As the preference based instrument will have been valued based on all

included domains this will lead to information loss and uncertainty about how well the mapped estimate captures the genuine preferences of the population from which the utility scores were derived. Current mapped estimates do not adequately capture all of the uncertainty that is generated as a result.

19 Second, three studies attempt to use clinical indicators to predict health related quality of life[4,5,6]. However, these studies cannot overcome the flaws associated with using clinical indicators directly in economic evaluations. Clinical indicators capture only a very narrow aspect of health related quality of life and do not necessarily capture all of the important changes to health related quality of life that may occur as a result of treatment. Using clinical indicators as the source data to map to a preference based measure therefore limits the amount of information that can be translated into utility scores. In effect, this fails to overcome the principle objection to using clinical indicators which is that comparisons cannot be made between interventions for different conditions because we cannot know from clinical indicators alone the different preferences for health states. A simple mapped estimate of preferences does not address this problem. Such mapped estimates will reflect only those domains directly relevant to the indicator selected, without capturing broader information about quality of life. Because mapped estimates based solely on clinical outcome indicators cannot distinguish which aspects of health related quality of life are being captured in the final estimate it is unlikely that using clinical indicators can ever be a valid source for mapped estimates.

Table 1: Studies included in systematic review

Study	Instruments		Mapping approach used and author's conclusions	Discussion of uncertainty
	Predictor	Predicted	Note: Comments is italics are taken directly from the relevant paper.	
Bansback et al[7]	Health Assessment Questionnaire Disability Index (HAQ-DI)	EQ-5D, SF-36 (transformed to SF-6D utility values using a published algorithm).	Five different linear regression models were evaluated. The focus was on the predictive ability of the algorithm, and not the explanatory power. The models were assessed against their ability to predict actual scores on the preference based measures. The authors conclude that it is possible to derive reliable models for estimating scores on both the EQ-5D and SF-6D from response to the HAQ-DI, although different models performed 'best' for the different preference based instruments.	There is a discussion of the fact that not all relevant domains may be captured by mapping from HAQ-DI to the EQ-5D or SF-6D. The authors suggest that information on missing domains may be captured in the responses obtained, but this hypothesis is not tested in this study (it is clearly beyond the scope of the study as published). It is also suggested that the domains included of the EQ-5D and SF-6D do not capture all of those aspects of health related quality of life that are relevant to the study population.
Barton et al[8]	Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)	EQ-5D	Linear regression was used to predict EQ-5D scores based on WOMAC responses. Five different models using different predictor variables were developed. The preferred model was chosen based on how closely actual EQ-5D scores were predicted by the models (using mean absolute error). Cost-effectiveness of the four interventions in the study was then estimated using both the mapped estimates of utility and the observed EQ-5D scores. The authors conclude that the mapped estimates consistently underestimated the mean QALY gain from the intervention and that while mapped estimates may be useful in some circumstances, direct measurement using a preference based measure is preferable.	The authors recognise that in using a condition specific instrument as the source data for mapping they may not be detecting all changes in health recorded by the generic instrument used in the study. They do not link this idea to the concept of information loss or identify strategies within the mapping process to address it other than to recommend the collection of preference based data where possible.
Brennan and Spencer[9]	Oral Health Impact Profile (OHIP)	EQ-5D	Two samples (one of clinicians, one of patients) were used to obtain data to map between the OHIP and the EQ-5D. Tobit regression models were used with the EQ-5D preference weight as the dependent variable and the OHIp scores as the independent variable alongside age and sex. Models with	There is no discussion of the implications of information being lost when mapping between measures. Discussion of uncertainty focuses only on the degree of variance explained using different mapping models.

			different combinations independent variables were tested but did not perform better. The authors conclude that the mapped estimates can be used where preferences have not or cannot be measured otherwise.	
Brazier et al[10]	Impact of Weight on Quality of Life (IWQOL)	SF-6D	Five OLS regression models were considered, using an additive specification and through the use of a stepwise procedure. The authors found a model that they believe adequately generates utility scores from the non-preference based instrument.	It is highlighted that the two instruments may each measure aspects of quality of life that the other does not. It is argued that the method is only as valid as the degree of overlap between the domains measured by the two instruments. It is acknowledged that for the mapping to be valid the descriptive system of the generic instrument must be a valid measure of the condition. There is no discussion of methods by which this might be assessed.
Buxton et al[11]	1) Inflammatory Bowel Disease Questionnaire 2) Crohn's Disease Activity Index (CDAI)	1) EQ–5D 2) SF-6D	The authors compared paired contemporaneous observations from the study population. Comparisons were made between each non-preference based measure and each preference based measure for a total of four different estimates. Correlations were measured between each paired group using summary scores from each instrument using Spearman correlation coefficients. Regression relationships were described using restricted (or residual) maximum likelihood estimation. Estimates were made in SAS using the MIXED procedure, and in each case the utility scores were the dependent variable. R2 and mean absolute percentage error were used to establish goodness of fit (ie how accurately the model describes the relationship). Covariates considered were age, gender, baseline concomitant medication, previous surgery for IBDQ and allocated treatment (active or placebo). The authors conclude that the IBDQ can be used to predict EQ-5D and .	The authors include a discussion of why they believe the scores of the IBDQ predict utility scores from the preference based measures; the IBDQ overlaps with three of five dimensions of the EQ-5D and five out of six dimensions of the SF-6D. However, there is no exploration of the manner in which this overlap might impact on the results, particularly with regards to estimates of uncertainty in the predicted scores.
Chancellor et al[12]	Combination of the Rotterdam Symptom Checklist (RSC), the Hospital Anxiety and	EQ-5D and Health Utilities Index Mark III (HUI3)	Two approaches to mapping to each preference based instrument were used. In the first, all 55 responses to the three non-preference based measures were used to predict scores on the EQ-5D and HUI3. The authors concluded that this was not a feasible approach. The second approach used	The conclusion reached by the authors in relation to direct mapping was based on the premise that the full range of information collected by the three measures could not be reflected in the summary scores on the preference based measures with sufficient certainty. The three non-preference

	Depression Scale (HADS) and a daily activities index (unspecified)		was factor analysis, where factors common to groups of variables are used to explain observed correlations between variables. The authors concluded however that the factor-analysed questions were no more suitable for mapping to the preference based measures than the full 55 response data set.	measures also did not include measurements of domains included in the preference based measures. A similar conclusion was reached for the factor analysed data. In each case the authors felt that the loss of information was too great to permit mapped estimates being made. There was no discussion of whether or not methods could be developed to overcome this difficulty or whether under different circumstances a mapped estimate could be produced using either method.
Franks et al[13]	SF-12	EQ-5D	OLS regression was used to regress EQ-5D index scores onto the SF-12 physical and mental summary scores. The conclusion of the authors is that the performance of the predicted EQ-5D scores seems to be adequate for group comparisons. Mapping was not attempted on individual item scores.	The authors focus on minimising the uncertainty of the predicted scores through variations in the specification of the regression model. There is no discussion of the uncertainty that arises from the loss of information when mapping between instruments.
Franks et al[14]	SF-12	EQ-5D; Health Utilities Index Mark 3	Least square regression analysis was used to estimate models for EQ-5D and HUI3 scores using the SF-12 physical and mental health subscale scores. The authors conclude that reliable estimates can be made by mapping the SF-12 to the EQ-5D and HUI3.	Discussion of uncertainty focuses on the reliability and accuracy of the predicted scores, not the validity. Although the study identifies weaknesses, there is no discussion of the uncertainty that arises from the loss of information when mapping between instruments.
Fryback et al[15]	SF-36	Quality of Well-being index (QWB)	Stepwise and best-subset regressions based on the eight scales of the SF-36, their squares and on pair-wise cross products. Best-subset analysis was undertaken where for each set of candidate variables 1 to n the software finds the two best expressions for explaining variance (ie those with the highest R2 value) and then lists these from the fewest number of variables to the most. Cp statistics were also calculated – these explain the precision of each individual model for predicting the dependent variable. Larger Cp scores are said to indicate lower precision. The model chosen as the best had the lowest number of variables from among those with a Cp of less than the number of variables in the model.	The authors emphasise that the algorithm developed is to be used to predict scores of one instrument using another and not to either explain or explore any relationships between the two instruments. The potential loss of information is not explicitly addressed.

Gray et al[16]	SF-12	EQ-5D	"*Multinomial logit regressions are used to estimate the probability that a respondent selects a particular response level for each question in the other generic instrument*". The authors argue that their method provides a second-best (to direct measurement using a preference based measure) method of estimating utility values.	There is no discussion of the uncertainty that arises from the loss of information when mapping between instruments.
Livingston and Ko[4]	BMI (by category)	Health and Activities Limitation Index (HALex)	Results are presented as class of obesity (from underweight to superobese) by utility score derived from the HALex instrument. Significance of results was determined using Student's t test. The authors conclude that the HALex instrument is sensitive to changes in obesity level and could be used to generate utility scores in obese patients. Covariance with other medical conditions was tested and statistical significant relationships were found between BMI and all tested co-morbidities except for 'current pregnancy'.	There is no discussion of the information that is lost when using a single clinical indicator to predict utility scores.
Longworth et al[5]	Clinical indicators for patients with stable angina	EQ-5D	OLS was used to predict EQ-5D scores from the clinical and demographic data. A forward stepwise approach was used in the regression analysis. Observed and predicted scores showed greater consensus at the higher end of the EQ-5D scale (ie, those with higher utility scores) than at the lower end, though this may be because of the lower number of observations at that end. The authors conclude that the clinical outcome measures can be used to predict utility scores using their preferred method of ordinary least squares regression.	The only estimates of uncertainty given in the paper relate to the accuracy of the predicted scores compared with the observed scores. There is no discussion of the uncertainty generated through information loss. However, the authors do highlight that "the uncertainty surrounding the estimates should also be considered within in the mapping process", recognising the implications of uncertainty for the results of mapping studies.
Mortimer et al[17]	SF-36	Assessment of Quality of Life (AQoL)	Various linear and OLS regression models were used to derive alternative item-based, subscale-based and scale-based algorithms. These were assessed with regards to their parsimony, identifiability, goodness of fit, theoretical consistency and predictive power. The authors found that none of the mapping algorithms were useful for predicting individual scores but may be useful for group scores and between group comparisons.	The authors describe several weaknesses with mapping but do not discuss the issue of measuring different domains of health with each instrument and the implications that this has for estimates of uncertainty.

Shih et al[3]	MD Anderson Symptom Inventory (MDASI) for cancer patients	SF-36 (then converted using a published algorithm for deriving utilities from the SF-36)	Using a summary score from the MDASI instrument along with demographic data, the authors developed three regression models to try and explain the variance in SF-36 scores among the study population. They conclude that MDASI scores are moderately, but significantly correlated with SF-36 scores and that further research is required to map the MDASI to utility values for use in economic evaluations.	The authors address the uncertainty relating to the predictive nature of the mapping process, but do not address the problem of the loss of information that may occur in mapping.
Sullivan and Ghushcyan[18]	SF-12	EQ-5D	A common method, ordinary least squares regression was rejected by arguing that because the EQ-5D exhibits a ceiling effect and that this was likely to bias the results. A Tobit model was also considered though this was rejected on the grounds that it too could be biased based on the potential for heteroskedasticity or a skewed (non-normal) distribution of responses. The authors found that the censored least absolute deviations approach generated the best estimate and had the greatest potential for reducing bias.	The authors discuss the uncertainty that arises from the predictive relationship, but no mention is made of the additional uncertainty arising from the loss of information when translating between two measures that are estimating HRQoL based on different domains of health.
Tsuchiya et al[19]	Asthma Quality of Life Questionnaire (AQLQ)	EQ-5D	Six different regression models used. Five models used an ordinary least squares (OLS) approach and one used a multinomial logistic approach. The authors conclude that use of mapping is a poor alternative to including preference based instruments as part of data collection. The mapping function cannot be used to estimate individual quality of life scores.	All discussion of uncertainty is focused on finding the model with the best fit to the data – ie the one most likely to predict EQ-5D scores using AQLQ scores. That the instruments were not designed to measure the same domains of health is highlighted but there is no discussion of this with regards to the uncertainty it creates.
Wodchis et al[20]	Minimum Data Set (for elderly care)	Health Utilities Index Mark II (HUI2)	Responses to the MDS were matched with responses to the HUI2. Not all items were included in the mapping, and of those that were a minority did not match exactly, being either of the same indicator but with different scales, or appearing in only one of two different versions of the MDS used. The MDS responses are then mapped to the relevant HUI2 score to develop the scoring system.	The authors address the issue of the potential loss of relevant information by selecting a preference based measure (the HUI2) that most closely mimics the domains of health being measured by the non-preference based measure. There appears to be an implicit assumption that in mapping each individual score in this way that there is no need to consider further the generation of uncertainty from mapping a non-preference based measure to the preferences generated using a different measure.

Discussion

20 The treatment of information loss in mapping is at an embryonic stage. Whilst some authors have considered whether mapping is appropriate, this is not reflected in the literature. Rather, the approaches to dealing with uncertainty in mapping have focused entirely on the predictive nature of the mapping function. This leaves a theoretical gap in the mapping process as the predictive accuracy can only tell us how closely scores are related. What cannot be said with any degree of certainty is whether or not the mapped estimates tell us anything useful about the preferences of the population that completed the non-preference based measure. If we do not understand how closely the mapped estimates relate to the actual preferences of the study population this may render any resource allocation decision based on mapping invalid.

21 Without including information loss in mapped estimates it is not possible to determine whether allocation decisions based on such estimates are valid. The validity of an individual instrument has traditionally been measured according to a set of established criteria and in particular four key criteria: face, content, criterion and construct validity (Streiner and Norman[21]). Increasingly these measures of validity are applied not just within the context of the psychometric properties of the instrument but also with relation to the people being assessed using the instrument[20]. This differs from how validity needs to be considered as part of the development of mapping functions, where it will be necessary to establish a method for assessing the validity of mapping from one instrument to another. This will use the established definitions of validity where appropriate but must also consider if there are other criteria by which the validity of a mapping function between two instruments can be judged.

22 An additional though less commonly applied approach to judging validity is the concept of decision validity[22]. Decision validity refers to the degree of certainty that the decision made is the correct one, given the available information. Although not explicitly stated by Streiner and Norman[20] this decision validity is what they refer to when questioning the process of establishing whether a measure exhibits the accepted individual concepts of validity. They consider these individual concepts to be part of a broader question of validity: 'Does the hypothesis of this validation study make sense in light of what the scale[s] [are] designed to measure?' and 'Do the results of this study allow us to draw

the inferences we wish to make?' (p174) [20]. Decision validity then is the ultimate arbiter for any mapping function; there is a direct relationship between the robustness of mapped utility estimates and the validity of any decision informed by those estimates. And so, if the uncertainty surrounding a decision falls within an acceptable range given the currently available information then it can be concluded that the decision is valid and therefore that the mapping function is valid.

23 For establishing decision validity two conditions need to be met: first, there must be a *prima facie* case that a function that maps between any two given instruments is valid. The *prima facie* case for decision validity is established by testing the validity of the source instrument against the target instrument. To satisfy this condition the source instrument should demonstrate the following aspects of validity to some degree:

a. face/content validity: do the two instruments appear to be measuring the same or similar concepts? If face validity cannot be established then it is highly unlikely that a prima facie case for decision validity can be established

b. construct validity: to what degree do the two instruments measure the same hypothetical constructs? Measuring the correlation between domain scores on separate tests and factor analysis are methods that can be used to measure construct validity for establishing the prima facie case

c. criterion validity: to what extent do the overall scores from the source instrument correlate with those from the target (or gold standard) instrument?

24 If it is established that there is a prima facie case that a mapping function may potentially lead to a valid decision then the second condition that must be met is whether or not it is possible to quantify the amount of information lost as part of the mapping. Any loss of information in moving between instruments affects the level of certainty that it is possible to have in the results. This information loss needs to be quantified, therefore, within the estimates of uncertainty that arise from mapping between instruments. (Developing methods for quantifying the loss of information is an obvious next step in this research but is beyond the remit of this paper, and may be a useful point of discussion at HESG.)

25 Where there is a *prima facie* case and there exists a method for quantifying the uncertainty arising from information loss, we could finally test the appropriateness of retaining current evidence (without explicit preference based measures of QoL, mapping, or an additional utility elicitation method). Here, a value of information analysis aids

consistency in the decision making process[23] by simplifying a problem to whether or not it is expected to improve net benefit. As the value of improved accuracy can be estimated for any given parameter it becomes possible to determine the expected benefit under the available options for resolving uncertainty.

26 An expected value of perfect information analysis could be undertaken in each case where an economic evaluation will be conducted using any particular mapping algorithm. Each economic evaluation, even those concerning the same or similar interventions or programmes will be subject to a specific set of uncertainties that may influence the outcome. Therefore it is unlikely that it would be possible to generalise the results of one analysis to another.

27 Whilst it is probably impractical to expect that an analysis should be undertaken in all cases, it should be noted that in many cases providing new utility values is likely to be inexpensive relative to the cost and consequences of treatment. If the results of value of information analysis are used, we might expect the gap between 'current evidence' and perfect information narrow quickly.

28 The approach suggested here of using value of information to judge whether additional utility information is necessary (and its type) is in contrast to NICE's guidance permitting the use of mapped estimates in technology appraisals[1]. Introducing and promoting this technique before the methodology is established may lead to a series of unintended and unwelcome consequences - treatments may erroneously be approved or rejected or large scale, irreversible investment may be made on the basis of poor information about the quality of life of the intended population. As is shown here, current methods used for mapping are inadequate to the task. Any allocation decision taken based on current mapped estimates, even if by chance they are the correct decisions, can only be described as valid in a post-hoc manner; yet waiting until the decision has been implemented to determine whether it is a valid decision does not address the original problem. Until the methodology exists for determining the validity of decisions based on mapped estimates mapping should not be used as a routine approach to estimating preferences and the best option remains valuation through the use of an established preference-based instrument or valuation method.

References

[1] National Institute for Health and Clinical Excellence (2008). *Guide to the methods of technology appraisal*.

[2] Mortimer, D and L Segal (2008). *Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights*. Medical Decision Making. Jan-Feb: 66 – 89.

[3] Brazier, J, Y Yang, A Tsuchiya (2008). *A review of studies mapping (or cross walking) from non-preference based measures of health to generic preference based measures of health*. Sheffield Health Economics Group Discussion Paper Series. 08 / 02.

[4] Shih, Ya-Chen T, XS Wang, SB Cantor and CS Cleeland (2006). *The association between symptom burdens and utility in Chinese cancer patients*. Quality of Life Research 15: 1427 – 1438.

[5] Longworth, L, M Buxton, M Sculpher and D Smith (2005). *Estimating utility data from clinical indicators for patients with stable angina*. European Journal of Health Economics. 6:347-353.

[6] Longworth, L, M Buxton, M Sculpher and D Smith (2005). *Estimating utility data from clinical indicators for patients with stable angina*. European Journal of Health Economics. 6:347-353.

[7] Bansback, N, c Marra, A Tsuchiya, A Anis, D Guh, T Hammond and J Brazier (2007). *Using the Health Assessment Questionnaire to estimate preference-based single indices in patients with rheumatoid arthritis*. Arthritis and Rheumatism. 57(6): 963-971.

[8] Barton, G, T Sach, C Jenkinson, A Avery, M Doherty, K Muir (2008). *Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores?* Health and Quality of Life Outcomes, 6:51.

[9] Brennan, D and J Spencer (2006). *Mapping oral health related quality of life to generic health state values*. BMC Health Services Research. 6:96

[10] Brazier, John E, Ronette L Kolotkin, Ross D Crosby, G Rhys Williams (2004). *Estimating a preference-based single index for the impact of weight on quality of life (IWQOL-Lite) instrument from the SF-6D*. Value in Health 7(4):490- 498.

[11] Buxton, MJ; LA Lacey; BG Feagan; T Niecko; DW Miller; RJ Townsend (2007). *Mapping from Disease-Specific Measures to Utility: An analysis of the Relationships between the Inflammatory Bowel Disease Questionnaire and Crohn's Disease Activity Index in Crohn's Disease and Measures of Utility*. Value in Health. 10(3): 214-220.

[12] Chancellor, J, D Coyle and M Drummond (1997). *Constructing health state preference values from descriptive quality of life outcomes: mission impossible?* Quality of Life Research 6: 159 – 168.

[13] Franks, P, E Lubetkin, M Gold, D Tancredi and H Jia (2004). *Mapping the SF-12 to the EuroQol EQ-5D index in a national US sample*. Medical Decision Making. 24: 247 – 254.

[14] Franks, P, E Lubetkin, M Gold, D Tancredi (2003). *Mapping the SF-12 to preference-based instruments: Convergent validity in a low-income, minority population*. Medical Care 41: 1277 – 1283.

[15] Fryback, DG, WF Lawrence, PA Martin, R Klein, B Klein (1997). *Predicting Quality of Well-Being Scores from the SF-36: Results from the Beaver Dam Health Outcomes Study*. Medical Decision Making. 17:1-9.

[16] Gray, A, O Rivero-Arias, P Clarke (2006). *Estimating the association between SF-12 responses and EQ-5D utility values by response mapping*. Medical Decision Making 26: 18 - 29.

[17] Mortimer, D, L Segal, G Hawthorne and A Harris (2007). *Item-based versus subscale-based mappings from the SF-36 to a preference-based quality of life measure*. Value in Health 10(5): 398 – 407.

[18] Sullivan, PW and V Ghusshchyan (2006). *Mapping the EQ-5D Index from the SF-12: us General Population Preferences in a Nationally Representative Sample*. Medical Decision Making 26: 401-409.

[19] Tsuckiya, Aki, J Brazier, E McColl and D Parkin (2002). *Deriving preference based single indices from non-preference based condition specific instruments: Converting AQLQ into EQ5D indices.* Sheffield Health Economics Group Discussion Paper Series. 02 / 01, May 2002.

[20] Wodchis, W, J Hirdes and D Feeny (2003). *Health related quality of life measure based on the Minimum Data Set*. International Journal of Technology Assessment in Health Care. 19(3): 490 – 506.

[21] Streiner, D and G Norman (2003). *Health Measurement Scales*. 3rd Edition. Oxford University Press. Oxford.

[22] Dowie, J. (2004). *Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions* Health Economics. 11(1):1-8

[23] Briggs A, K Claxton, M Sculpher (2006). *Decision Modelling for Health Economic Evaluation*. Oxford University Press. Oxford.