# DISEASE SPECIFIC VERSUS GENERIC MAPPING METHODS: HOW TO LINK OUTCOMES TO EQ-5D

REPORT BY THE NICE DECISION SUPPORT UNIT

19th November 2019

Monica Hernandez, Georgios Chrysanthou, Allan Wailoo

NICE 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
Website www.nicedsu.org.uk
Twitter @NICE_DSU

# ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester 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 www.nicedsu.org.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.

**Acknowledgements**

We would like to thank Kaleb Michaud from FORWARD, the National Databank for Rheumatic Diseases, for continuing kind permission to use the data for analyses reported here.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. NICE provided helpful comments on a previous draft. 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.

# EXECUTIVE SUMMARY

NICE currently recommends that EQ-5D-5L data be valued using a method for mapping to the EQ-5D-3L value set. There are a set of models available to do this which we refer to as “generic” mapping approaches. NICE recommends a particular approach for mapping patient level responses to the EQ-5D5L to 3L published by van Hout et al. However, in some circumstances, analysts may also have the option of mapping to EQ-5D-3L using some disease specific mapping model. We refer to these as “disease specific” mapping approaches. It is not known which of these approaches is preferable in terms of the ability of the models to accurately predict EQ-5D-3L values (or utilities).

In this report we present evidence that compares generic and disease specific mapping models. We use a large dataset (n=5192) provided by FORWARD, the National Databank for Rheumatic Diseases (NDB) for the analysis. Patients with rheumatoid arthritis (RA) completed EQ-5D-3L, EQ-5D-5L and the Health Assessment Questionnaire (HAQ) which comprises a measure of functional disability (the disability index) and also includes a measure of pain severity. There is also an existing, published mapping study that links patient HAQ scores to EQ-5D-3L using mixture model methods and that also reports the results of a simple linear model.

Results differ according to the precise measure of fit selected. In broad terms, the generic mapping approaches perform better than the RA disease specific methods using some summary measures of fit (Mean Absolute Error [MAE] and Root Mean Squared Error [RMSE]). The opposite is the case when Mean Error is used to define performance. The disease-specific linear regression is always the worst performing method when assessed by RMSE and MAE.

Model fit assessed over the range of disease severity assessed by a) HAQ and b) pain shows better performance of the disease specific mixture model approach compared to the generic approaches, with very close alignment to the observed values across the range except where data are very sparse. The differences between estimates of health gain for patients moving from severe impairment/pain to no impairment/pain are

substantial and are apparent for all generic methods. Improvements in pain of 8 to 0 (where 96% of the data lie) would result in an underestimation of health gain by 0.126 using the generic van Hout method and a small overestimation of benefit by 0.003 using the disease-specific mixture model approach. These measures of fit are of most relevance for the performance of mapping models in informing economic evaluation.

On balance, findings show that it is not possible to draw conclusions about which approach is likely to be most suitable for mapping solely on the basis of whether a mapping is “generic” or “disease specific”. In the RA case study presented here, very poor performance is evident in relation to the use of the linear regression. This aligns with many existing studies from other disease areas. For analysts that face the choice between using a disease specific mapping approach that has been previously published, or the generic van Hout mapping approach, full justification and demonstration of acceptable performance of the disease specific mapping is essential. In the absence of convincing evidence we recommend the generic method be used.

The evidence presented here suggests that disease specific mapping approaches are likely to be more appropriate than generic methods that map from EQ-5D-5L to 3L if they a) use appropriate methods, b) are based on outcome measures that reflect the dimensions of health considered relevant to patients in that disease area and c) have demonstrated good performance empirically. Differences between methods can be substantial, particularly when considered on the limited scale of health utilities and within the marginal health gains observed for most health technologies evaluated by NICE.

# 1. CONTENTS



## TABLES



## FIGURES

# 2. INTRODUCTION

As with many public sector decision making bodies, the National Institute for Health and Care Excellence (NICE) is required to make its recommendations based on procedural fairness. Consistency is a critical aspect of this and manifests itself both in processes and methods that determine the way in which evidence of health benefits from different health technologies are assessed. It is for this reason that the NICE Guide to the Methods of Technology Appraisal[1] ("The Methods Guide"), outlines a reference case: a set of methods to be used across its technology appraisals. One of the recommendations contained there is that the EQ-5D-3L (henceforth '3L') instrument be used as the basis for calculation of health state utilities and subsequent calculation of Quality Adjusted Life Years (QALYs), other than in those situations where it is demonstrably inappropriate.

In many situations, sufficient 3L data have not been collected that permit the required cost-effectiveness analysis to be performed. There may have been no such data included in the relevant clinical studies, or data from a different health utility instrument may have been collected, or the data may not be sufficient to provide estimates of rarer complications, long term events, or other aspects of the course of disease and its treatment that the clinical studies are too short to observe.

In these situations, the use of "mapping" can be used to bridge the evidence gap. This is a method recognised by NICE and referred to in the Methods Guide. Mapping entails the use of an external dataset that contains observations from patients that have simultaneously completed or had recorded the clinical measures that have been in the clinical studies and the preference based outcome measure (3L in this case) that decision makers require. Using a statistical model to link the two sets of outcomes then allows the treatment effect to be quantified in terms of QALYs. For the purpose of this report, we refer to this type of mapping as "disease specific" because it is based on data from patients with the same (or at least similar) health condition and uses outcome measures as explanatory variables that are also designed to be relevant to that same health condition. Whilst there is much consideration of the performance of alternative methods for mapping, NICE Technology Appraisals (TAs) are well used to receiving such mapping models as part of the evidence submission. Kearns et al[2]

identified mapping to be used in nearly a quarter of TAs from a review of NICE appraisals spanning from 2008 to 2011.

In recent years, a new version of the EQ-5D, the EQ-5D-5L (henceforth '5L'), has been developed. The descriptive system has been expanded to allow respondents to indicate their degree of impairment from 3 levels to 5. In addition, a new value set for England has been estimated[3]. We know that moving from 3L to 5L will cause substantial differences in estimates of cost effectiveness[4]. Concerns have been raised about the validity of the 5L value set[5]. However, since many clinical studies have included the 5L descriptive system such evidence is now beginning to form part of the submissions NICE receives. Currently, NICE recommends[6] that individual responses to the 5L descriptive system be transformed to estimated 3L values using a mapping method published by van Hout et al[7]. This is an approach that is based on data from a EuroQoL group coordinated series of studies that collected data on both 3L and 5L. The studies were carried out in 6 countries: Denmark, England, Italy, the Netherlands, Poland and Scotland and included eight broad patient groups (cardiovascular disease, respiratory disease, depression, diabetes, liver disease, personality disorders, arthritis, and stroke) and a student cohort (healthy population). 3,691 responses were obtained but the way in which the data were then used in the van Hout mapping approach is not straightforward. A full description of this process is provided in Hernandez Alava et al[8] which also compares the van Hout mapping method with methods developed by the DSU. We refer here to these mapping approaches as "generic" since they map to 3L using another outcome measure that is intended to be applicable across a large scope of disease types, and the statistical models are expected to be applied to patients of all types.

In some situations, both generic and disease specific mapping options are feasible. Where 5L has been administered in the relevant clinical studies the generic approach is clearly an option. But it may also be the case that relevant clinical outcomes have been collected that would enable the use of a previously published mapping model to be used to estimate 3L, or the analysts may be able to access data that would permit them to estimate a suitable mapping model themselves. NICE guidance currently provides no instruction as to whether the generic or disease specific approaches should be preferred and there is currently no evidence on which to base any

recommendation. This report provides evidence that compares the performance of generic and disease specific mapping methods that may be taken into account when formulating more helpful guidance to those submitting evidence to NICE and to those responsible for interpreting that evidence.

# 3. METHODS

## 3.1. DATA

To compare the performance of generic and disease specific mapping methods for estimating 3L, one requires individual respondent level data with a number of different features. First, conducting the generic mapping requires responses to the 5 domains of the 5L instrument. Second, there must be a validated, existing mapping study that predicts 3L from clinical outcomes and those outcomes must also be recorded for the same patients.

The only substantial dataset that we are aware of that meets these requirements is FORWARD, the National Databank for Rheumatic Diseases (NDB). We have previously described the key features of this data[8]. The NDB is a register of patients with rheumatoid disease, primarily recruited by referral from US and Canadian rheumatologists. Information supplied by participants is validated by direct reference to records held by hospitals and physicians (a minority of cases come by self-referral, with medical details obtained by NDB in the same way). Full details of the recruitment process are given by Wolfe and Michaud (2011)[9]. The EQ-5D responses and other patient-supplied data are collected by various means, primarily postal and web-based questionnaires completed directly by patients. Data collection began in 1998, and continues to the present, in waves administered in January and July of each year. In 2011, there was a switch from 3-level to the 5-level version of EQ-5D and both versions were collected during the January 2011 wave. The NDB questionnaire is 27 pages long and it includes many general as well as rheumatoid disease specific questions. In particular, the questionnaire contains the Health Assessment Questionnaire (HAQ). The HAQ is a widely used outcome measure in clinical studies of rheumatoid arthritis, is the key explanatory variable used for mapping studies (see 3.2 below) and is also

used in numerous cost effectiveness models to define health states. 5L and 3L are on pages 11 and 22 of the questionnaire respectively.

### 3.2. MAPPINGS METHODS

The primary focus of the analysis is the generic van Hout et al method that is recommended by current NICE guidance. We also include two variants of the generic DSU copula based mapping method which differ according to the dataset that was used for estimation[10]. One is based on the same EuroQoL group (EQG) data as the van Hout et al method. The other is based on the FORWARD NDB data. A previous DSU report conducted extensive comparisons and validations of these three generic methods for predicting 3L from 5L so, in this report, we limit reporting of results pertaining to the DSU methods to those situations where important differences are illustrated[8].

Two disease specific mapping models are used to link outcomes in rheumatoid arthritis to 3L. Our primary focus is on results obtained from applying the Adjusted Limited Dependent Variable Mixture Model (ALDVMM) reported in full detail in Hernandez et al[11]. This method has been developed specifically to reflect the key characteristics of health utility data, and 3L in particular, and has been demonstrated in numerous applications to perform well, avoiding problems of poor fit associated with more standard methods. We would therefore expect that, if generic methods outperform the ALDVMM approach, then we would have more confidence in the generalisability of any conclusions because often disease specific mappings use methods that perform poorly compared to the ALDVMM.

For the sake of comparison, we also undertook analyses using the linear model also reported by Hernandez et al.

In both cases, the explanatory variables were the summary score from the HAQ instrument, pain described on the visual analogue scale, which also forms part of the HAQ instrument, age and sex.

All utility values for 3L are based on the UK tariff[12].

### 3.3. ASSESSMENT OF MODELS

There were 5,192 observations. We used the responses to the dimensions of the 5L to predict 3L utility values (UK tariff) for all patients. We then used responses to HAQ, pain, age and sex to predict 3L utility values for the same patients. We then assessed the differences between the observed 3L values and those predicted by the different mapping models using standard measures of summary fit: Mean Error (ME), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for the entire patient sample and for subgroups of observations. Box 1 describes each of these measures. Visual inspection of plots of model fit by severity of disease was also undertaken.

**Box 1: Summary measures of fit**

$$Mean\ error = \frac{\sum_{i=1}^{n} y_i - \widehat{y_i}}{n}$$

$$Mean\ absolute\ error = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$$

$$Root\ mean\ squared\ error = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n}}$$

*Where $y_i$ is the observed value and $\widehat{y_i}$ the prediction. Mean error is the average distance between observed and predicted values. Overprediction and underprediction will cancel out. Mean absolute error is the average of the absolute errors. All are measured on the same scale as the variable being measured. RMSE gives a greater weight to large errors than MAE.*

# 4. RESULTS

Results relating to summary measures of fit are displayed in Table 1, for the entire sample and for subsamples selected to represent different classes of disease severity defined by HAQ and pain.

The comparisons between the generic methods have previously been reported for this same sample in our previous DSU report[8]. We focus here on the differences between generic and disease specific mapping models.

Results differ according to the precise measure of fit selected. In broad terms, the generic mapping approaches perform better than the RA disease specific methods when performance is measured by MAE and RMSE. The opposite is the case when ME is used to define performance.

The linear regression is always the worst performing method when assessed by RMSE and MAE. Within sample, linear regression seeks to minimise the mean error. In this out-of-sample assessment it is noticeable that mean error for the overall sample is also relatively high (within sample we would expect the linear regression to be close to zero). Performance assessed by ME is variable when considering different disease severity subgroups. In several parts of the distribution it performs relatively well (pain >0 and ≤3, or >6.5 and ≤9.5) whilst for other parts of the distribution it fits poorly (e.g. pain = 0, HAQ >1 and ≤2).

The ALDVMM routinely outperforms the linear regression. It is almost always one of the best two performing methods in terms of ME both for the entire sample (where it is only inferior to the DSU NDB copula based model for which this is a within sample comparison) and for most parts of the disease severity distribution. The mean from the ALDVMM is very close to the sample mean. However, the performance of the ALDVMM is generally worse than the generic mapping methods when using the MAE and RMSE measures.

The generic mapping methods all display varied performance. They varied both according to the precise measure of fit being used, the degree of disease severity and whether that was measured by HAQ or pain. The DSU NDB model has the best performance using model fit statistics for the overall sample except for MAE. The importance and rationale for the difference between MAE and RMSE is discussed in this context in more detail in our previous DSU report[8].RMSE penalises large errors more heavily than MAE. **Error! Reference source not found.** displays mean EQ-5D-3L by HAQ, for the observed versus modelled values. We only present results from

the van Hout model (part a) for the generic mapping approaches, and the ALDVMM for the disease specific mapping approaches (see part b). The van Hout approach shows underestimation of EQ-5D-3L where HAQ is low (1 or below), which represents mild functional impairment. These differences are substantial when considered on the health utility scale where 1 is equal to full health and 0 to states equivalent to dead. The difference between observed 3L scores, and 3L scores predicted from 5L using the generic van Hout approach, was 0.024 when HAQ is zero, 0.037 when HAQ is 0.125, 0.025 when HAQ is 0.25.

The ALDVMM model shows much closer alignment to the conditional means of the observed data until HAQ is greater than 2.75, the most severe part of functional disability. As noted above, the data here are both sparse and produce unusual spikes in the mean observed EQ-5D-3L score owing to the small numbers of observations.

Considering an improvement in HAQ from 2.25 to 0 (where over 97% of the data lie) the van Hout method would underestimate health gain by 0.057 compared to the observed data. The ALDVMM model underestimates gain by 0.026.

It is also notable that the linear regression (results not shown), underestimates health utility when HAQ is zero by 0.031.

It is worth noting that the DSU NDB based mapping showed very close fit to the data across all parts of the distribution, with the only noticeable diversion occurring where HAQ is equal to 2.75 or 3. Data are very sparse at this extreme degree of functional disability. Only 30 observations report HAQ at 2.75 or greater (0.58% of the overall data) and there is a counter intuitive, large increase in mean EQ-5D-3L at a HAQ of 2.875 (which has just 5 observations). This is also the only part of the HAQ distribution where this model resulted in a greater distance between observed and predicted values than the van Hout approach. However, this is the only model for which this is an in-sample comparison.

The DSU EQG model demonstrated very similar performance to the van Hout method, with marginally closer fit to the observed data where HAQ exceeds 1.5 and marginally worse fit between 0 and 1. These differences appear negligible.

**Table 1**

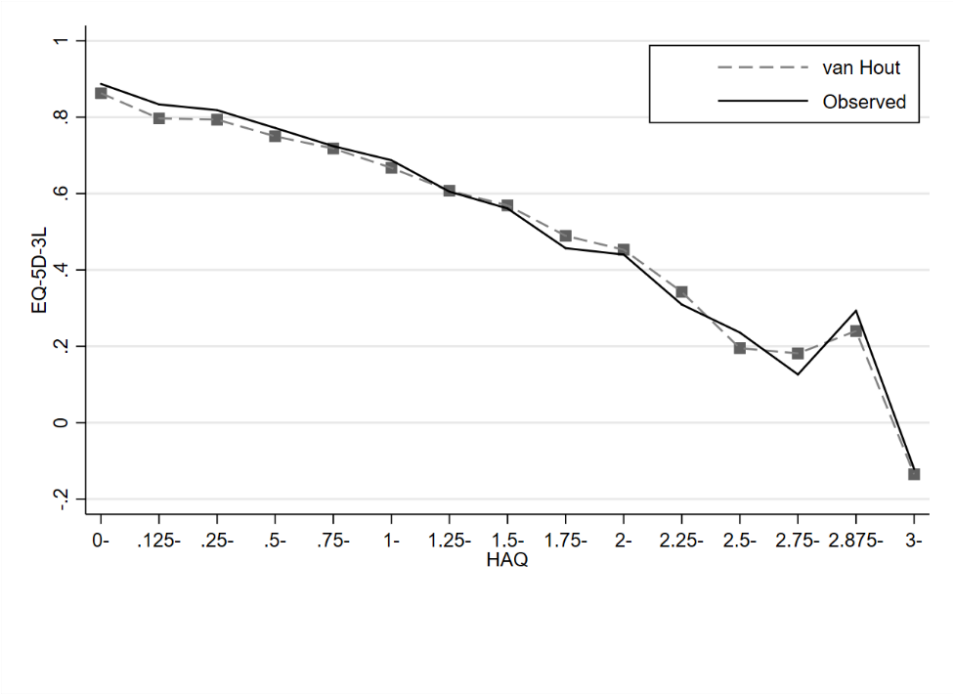
| | Generic mapping | | | Disease specific mapping | |
|---|---|---|---|---|---|
| | **Van Hout** | **DSU NDB** | **DSU EQG** | **ALDVMM** | **Linear** |
| *Overall sample (n=5,192)* | | | | | |
| Mean (sample mean = 0.6808) | 0.6721 | 0.6802 | 0.6610 | 0.6789 | 0.6702 |
| ME | 0.0087 | 0.0006 | 0.0198 | 0.0019 | 0.0106 |
| MAE | 0.0941 | 0.1004 | 0.0996 | 0.1165 | 0.1270 |
| RMSE | 0.1491 | 0.1472 | 0.1485 | 0.1648 | 0.1693 |
| *HAQ<=1 (n=2984)* | | | | | |
| ME | 0.0208 | 0.0098 | 0.0330 | 0.0043 | 0.0097 |
| MAE | 0.0729 | 0.0788 | 0.0794 | 0.0859 | 0.0960 |
| RMSE | 0.1123 | 0.1093 | 0.1146 | 0.1167 | 0.1235 |
| *HAQ>1&HAQ<=2 (n=1,845)* | | | | | |
| ME | -0.0070 | -0.0108 | 0.0050 | -0.0020 | 0.0121 |
| MAE | 0.1128 | 0.1204 | 0.1184 | 0.1466 | 0.1560 |
| RMSE | 0.1748 | 0.1751 | 0.1740 | 0.2020 | 0.2053 |
| *HAQ>2&HAQ<=3 (n=363)* | | | | | |
| ME | -0.0111 | -0.0166 | -0.0131 | 0.0027 | 0.0106 |
| MAE | 0.1736 | 0.1761 | 0.1697 | 0.2148 | 0.2347 |
| RMSE | 0.2430 | 0.2364 | 0.2316 | 0.2629 | 0.2655 |
| *Pain =0 (n=341)* | | | | | |
| ME | 0.0315 | 0.0348 | 0.0464 | 0.0120 | 0.0880 |
| MAE | 0.0601 | 0.0715 | 0.0748 | 0.0785 | 0.1205 |
| RMSE | 0.1076 | 0.1011 | 0.1129 | 0.1004 | 0.1360 |
| P10>0&P10<=0.3 (n=2618) | | | | | |
| ME | 0.0246 | 0.0083 | 0.0345 | 0.0047 | 0.0026 |
| MAE | 0.0775 | 0.0812 | 0.0811 | 0.0886 | 0.0940 |
| RMSE | 0.1165 | 0.1106 | 0.1154 | 0.1201 | 0.1239 |
| P10>0.3&P10<=0.65 (n=1411) | | | | | |
| ME | -0.0015 | 0.0005 | 0.0179 | -0.0116 | 0.0154 |
| MAE | 0.0903 | 0.1000 | 0.0968 | 0.1126 | 0.1236 |
| RMSE | 0.1505 | 0.1530 | 0.1500 | 0.1718 | 0.1723 |
| P10>0.65&P10<=0.95 (n=786) | | | | | |
| ME | -0.0321 | -0.0340 | -0.0306 | 0.0170 | 0.0043 |
| MAE | 0.1682 | 0.1740 | 0.1732 | 0.2290 | 0.2396 |
| RMSE | 0.2315 | 0.2310 | 0.2295 | 0.2664 | 0.2699 |
| P10==1 (n=36) | | | | | |
| ME | -0.0726 | -0.1176 | -0.1205 | -0.0920 | -0.1872 |
| MAE | 0.1595 | 0.1778 | 0.1838 | 0.2029 | 0.2694 |
| RMSE | 0.2260 | 0.2374 | 0.2437 | 0.2547 | 0.3003 |

= best performing | = 2nd best performing | = worst performing

Abbreviations: DSU – Decision Support Unit, NDB – FORWARD The National Databank for Rheumatic Diseases, EQG – EuroQoL Group, ALDVMM – Adjusted Limited Dependent Variable Mixture Model, ME – Mean Error, MAE – Mean Absolute Error, RMSE – Root Mean Squared Error

**Figure 1: Mean observed and predicted EQ-5D-3L by HAQ**

**a) Observed vs van Hout generic mapping**




**b) Observed vs ALDVMM disease specific mapping**

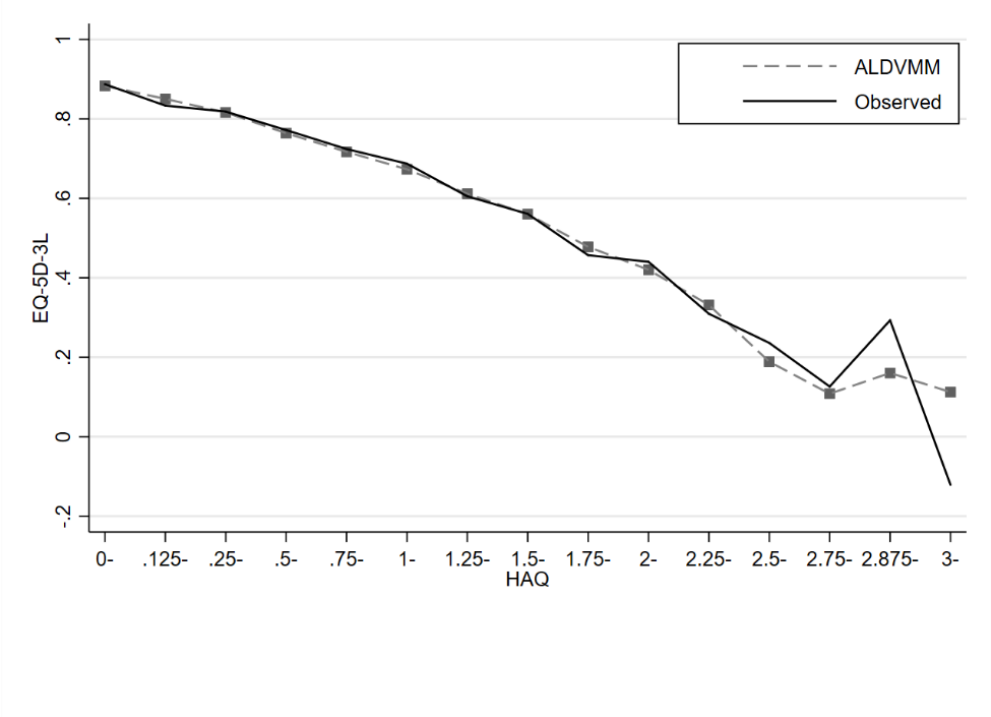
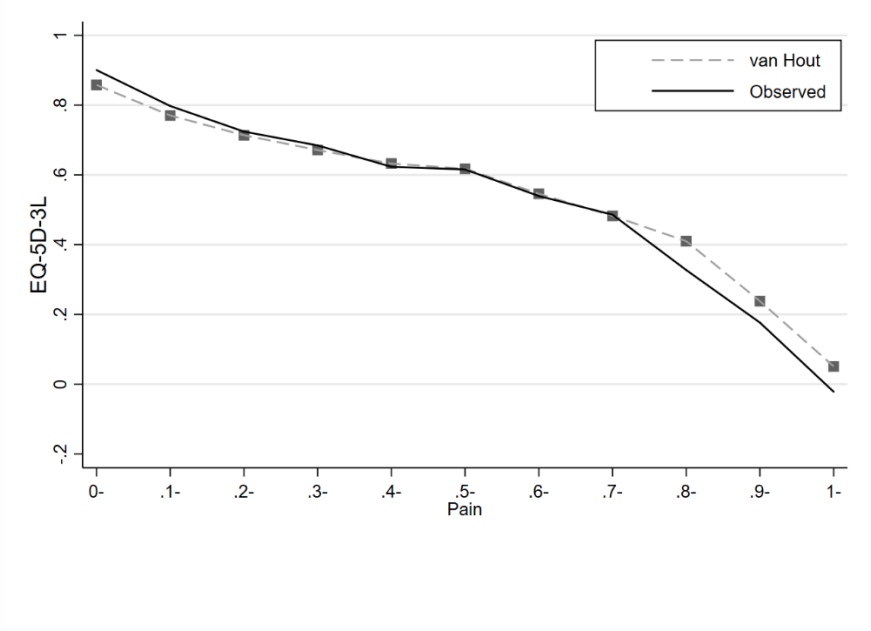
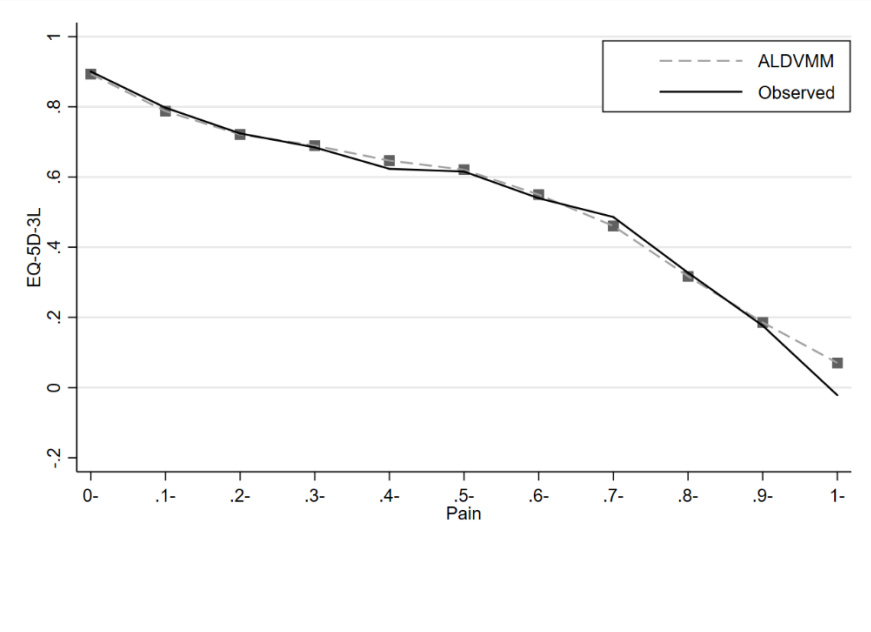


**Figure 2: Mean observed and predicted EQ-5D-3L by Pain**

**a) Observed vs van Hout generic mapping**




**b) Observed vs ALDVMM disease specific mapping**



**Error! Reference source not found.** displays mean EQ-5D-3L by pain score for the observed data and predicted values from a) the van Hout et al mapping and b) the ALDVMM model. Similar findings to those relating to EQ-5D-3L scores by HAQ are illustrated but are more pronounced in relation to pain.

The van Hout model underpredicts health utility at low levels of pain and overpredicts health utility at high levels of pain. For pain scores of 0, 1, 2 and 3 the van Hout approach underestimates health utility compared to the observed data by 0.043, 0.027, 0.011 and 0.013 respectively. The differences between predicted values and the observed data are largest where pain is 8 or greater. At pain levels of 8, 9 and 10 the van Hout approach overestimates health utility by 0.083, 0.062 and 0.073. There are relatively few observations where pain is high, though these numbers are far greater than where HAQ is at its highest. 7.3% (n=379) of observations record pain at 8 or greater, 2.2% (n=114) record pain at 9 or greater.

Alternative generic mapping methods are not shown in the figures. However, the DSU EQG model also shows underestimation of health utilities, persisting for levels of pain up to 6, and overestimation of utilities at levels 8, 9 ad 10. Better performance is observed for the DSU NDB based model for which this is an in-sample validation.

The ALDVMM mirrors the mean observed data conditional on pain more closely over most of the pain distribution. There is a notable overestimation of health utility of 0.092 where pain is equal to 10, a greater difference than reported for the van Hout method. 0.7% of observations (n=36) report pain at this level.

Overall, an improvement in pain from 9 to 0 (where 99% of the data lie) would result in an underestimation of health benefit of 0.104 using the van Hout method compared to the observed data and 0.016 using the ALDVMM method. For improvements of 8 to 0 (where 96% of the data lie) would result in an underestimation of 0.126 using the van Hout method and an overestimation of benefit of 0.003 using the ALDVMM.

It is worth noting that the limitations of the linear regression are particularly noticeable at the extremes of the pain distribution. For example, where pain is equal to zero, the

linear regression underestimates health utility by 0.059. Where pain is equal to 0.9 or 1 the linear regression overestimates health utility by 0.098 and 0.187 respectively.

# 5. DISCUSSION

There is no single method for assessing the performance of different mapping models. It is always important to consider a range of criteria and plots showing how models perform at different parts of the disease severity spectrum, and by different measures of disease. The importance of this general advice is reflected in the findings here where the use of standard summary measures of model fit and different measures of disease impact (HAQ and pain) illustrate contradictory findings. The assessment of any mapping model should be undertaken with some consideration as to how its results would be propagated through a cost effectiveness model. In this situation, plots of mean EQ-5D-3L as a function of HAQ and pain (which are typically used to define patient severity in cost-effectiveness models which assess technologies for people with RA) are of particular importance.

Findings show that it is not possible to draw conclusions solely based on whether a mapping is "generic" or "disease specific". In particular, many models used to conduct disease specific mappings are known to systematically undervalue health utility for mild health states and overvalue them for severe health states. This is evident in the RA case study presented here, with very poor performance evident in relation to the use of the linear regression. The ISPOR Good Practice Guide on mapping states: *"that it is wise to use a model type for which there is existing empirical evidence of good performance, and that respects the key features of the target utility measure, particularly the limited range of feasible utility values"*[13]. . There is ample empirical and theoretical evidence that linear regression is inappropriate for mapping[14]. Furthermore, the NICE Methods Guide states in relation to mapping models that *"its choice justified, and it should be adequately demonstrated how well the function fits the data"*[15]. For analysts that face the choice between using an "off-the-shelf" disease specific mapping, or the generic van Hout et al mapping approach, it is essential that these criteria are fully complied with. In the absence of convincing evidence of model performance (which in the case of linear regression methods is likely to reveal poor performance) then we recommend the generic method be used.

Our evidence shows that disease specific mapping approaches, conducted using methods that have been designed to be appropriate for EQ-5D-3L, and that have demonstrated good performance empirically, are more reliable than generic methods that map from EQ-5D-5L to 3L for use in cost effectiveness analysis. Differences between methods are substantial when considered on the limited scale of health utilities and within the marginal health gains observed for most health technologies. Well-constructed disease specific mapping methods make use of outcome measures that reflect the dimensions of health considered relevant to patients in that disease area. They do so based on data that was collected from patients with the same condition and using models that reflect the characteristics of the underlying data. They do not rely on assumptions that are at odds with the collected data. Generic methods for mapping from EQ-5D-5L do so using explanatory variables that are intended to be of relevance to a broad range of health conditions and based on data that may contain responses from groups of patients / the general population that may have little in common with the condition of interest. We know from the range of methods for mapping between EQ-5D-5L and EQ-5D-3L that the dataset matters[8]. However, in this case we know that a proportion of respondents contributing data used to inform the van Hout et al method did come from patients with arthritis of some type. Van Hout et al report that 250 respondents had “arthritis” (6.8%) and 122 (3.3%) had “rheumatoid arthritis”[7]. There is therefore some overlap with the patient samples. It may be the case that the difference between generic and disease specific mapping methods is greater when the respondent groups are more diverse.

This is only one case study but obtaining datasets of sufficient magnitude to allow further comparisons of methods to be performed is unlikely to be a realistic option in the absence of specific research initiatives.

# 6. REFERENCES


[1] NICE (2013) *Guide to the Methods of Technology Appraisal*, available at https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781 (last accessed 11th June 2019)

[2] Kearns B, Ara R, Wailoo A, et al. (2013) *Good practice guidelines for the use of statistical regression models in economic evaluations*. Pharmacoeconomics, Vol:31:643–52.

[3] Devlin N, Shah K, Feng Y et al. (2018) *Valuing health-related quality of life: an EQ-5D-5L value set for England*. Health Economics 27(1): 7-22.

[4] Hernandez Alava, M., Wailoo A., Grimm S., Pudney S., Gomes M., Sadique Z., Meads D., O'Dwyer J., Barton G., Irvine L. (2018) *EQ-5D-5L versus 3L: the impact on cost-effectiveness in the UK*, Value in Health, Vol.21(1):49-56

[5] Hernandez Alava M, Pudney S, Wailoo A. (2018) *Quality review of a proposed EQ5D-5L value set for England*, EEPRU Report, available at http://www.eepru.org.uk/wp-content/uploads/2017/11/eepru-report-eq-5d-5l-27-11-18-final.pdf (last accessed 25th Feb 2019)

[6] NICE (2019) *Position statement on use of the EQ-5D-5L valuation set for England* (updated October 2019). Available at https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l (last accessed 12th November 2019)

[7] Van Hout B, Janssen M, Feng Y et al. (2012) Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health, 15: 708-15.

[8] Hernandez Alava M, Wailoo A, Pudney S. (2017) *Methods for Mapping Between the EQ-5D-5L and the 3L for Technology Appraisal*. NICE DSU Report. Available at http://nicedsu.org.uk/wp-content/uploads/2017/05/Mapping-5L-to-3L-DSU-report.pdf (last accessed 12th June 2019)

[9] Wolfe, F. & Michaud, K. (2011), *The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank*, Rheumatology 50, 16-24.

[10] Hernandez 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, Vol.55:139-152. Available at: https://www.sciencedirect.com/science/article/pii/S0167629616305070?via%3Dihub (last accessed 12th June 2019)

[11] Hernandez Alava, M., Wailoo, A., Wolfe, F., Michaud, K. (2014) *A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes*, Medical Decision Making, Vol: 34:919–930.

[12] Dolan, P. (1997), *Modeling valuations for EuroQol health states*. Medical Care, 35:1095-1108.

[13] 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

[14] Hernandez Alava M, Wailoo A, Pudney S, Gray L, Manca A. (forthcoming) *Modelling generic preference based outcome measures - development and comparison of methods*, Health Technology Assessment, forthcoming.

[15] NICE (2013) *Guide to the Methods of technology Appraisal*, NICE London. Available at https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781 (last accessed 18th November 2019)