

Mapping the EORTC-QLQ-C30 to the EQ-5D-3L: An assessment of existing and newly developed algorithms

Fionn Woodcock M.Sc.¹, Brett Doble Ph.D.², on behalf of the CANCER 2015 Consortium

¹School of Arts and Social Sciences, Department of Economics, City University, London,
EC1V 0HB, United Kingdom

²Health Economics Research Centre, Nuffield Department of Population Health, University
of Oxford, Oxford, OX3 7LF, United Kingdom

Address correspondence to: Brett Doble, Health Economics Research Centre, Nuffield
Department of Population Health, University of Oxford, Oxford, OX3 7LF, United Kingdom.
Tel: +44 1865289193; Email: brett.doble@dph.ox.ac.uk

Running head: Mapping the EORTC QLQ-C30 to the EQ-5D-3L

Keywords: cancer, condition-specific non-preference-based measures, external validation,
generic preference-based measures, regression models, quality of life

Financial support: No financial support was provided for the conduct of this study.

Word count: 5,085

ABSTRACT

Objectives: To assess the external validity of mapping algorithms for predicting EQ-5D-3L utility values from EORTC QLQ-C30 responses not previously validated and to assess whether statistical models not previously applied are better suited for mapping the EORTC QLQ-C30 to the EQ-5D-3L.

Methods: 3,866 observations for 1,719 patients from a longitudinal study (Cancer 2015) were used to validate existing algorithms. Predictive accuracy was compared to previously validated algorithms using root mean squared error, mean absolute error across the EQ-5D-3L range and for ten tumour-type specific samples as well as using differences between estimated QALYs. Thirteen new algorithms were estimated using a subset of the Cancer 2015 data (3,203 observations for 1,419 patients) applying various linear, response mapping, beta and mixture models. Validation was performed using two datasets composed of patients with varying disease severity not used in the estimation and all available algorithms ranked on their performance.

Results: None of the five existing algorithms offer an improvement in predictive accuracy over preferred algorithms from previous validation studies. Of the newly estimated algorithms, a two-part beta model performed the best across the validation criteria and in datasets composed of patients with different levels of disease severity. Validation results did, however, vary widely between the two datasets and the most accurate algorithm appears to depend on health state severity as the distribution of observed EQ-5D-3L values varies. Linear models performed better for patients in relatively good health, whereas beta, mixture and response mapping models performed better for patients in worse health.

Conclusion: The most appropriate mapping algorithm to apply in practice may depend on the disease severity of the patient sample whose utility values are being predicted.

INTRODUCTION

The collection of health-related quality-of-life (HRQoL) data from generic, preference-based measures, like the EQ-5D-3L,¹ SF-6D² and HUI III³ is not always prioritised in clinical trials of new interventions. This is particularly common in oncology trials where priority is given to the collection of HRQoL data from non-preference-based, condition-specific measures such as the EORTC-QLQ-C30 (henceforth QLQ-C30)⁴ or the Functional Assessment of Cancer – General (FACT-G).⁵ Many reimbursement agencies, however, recommend the use of generic, preference-based measures of HRQoL for estimating quality-adjusted life-years (QALYs) when conducting cost-utility analyses.⁶ Thus, researchers performing such analyses require a method for obtaining generic, health-state utility values from condition-specific, non-preference-based measures of HRQoL. One potential approach is through ‘mapping’, which uses a statistical algorithm to transform available health status data from a non-preference-based measure into utility values for a generic, preference-based measure, like the EQ-5D.⁷

A number of mapping algorithms capable of predicting EQ-5D-3L utility values from QLQ-C30 responses are available in the published literature,⁸⁻¹⁸ the majority of which have been externally validated.¹⁹⁻²² External validation of the existing algorithms highlight that they are largely inappropriate and may be due to not only limitations in the statistical methods used, but also the lack of conceptual overlap between the EQ-5D-3L and QLQ-C30. Two algorithms with the ‘best’ performance have, however, been consistently identified.^{19 20} These algorithms include an ordinary least squares (OLS) estimation model using QLQ-C30 scores as dummy variables (Veersteegh algorithm) and a multinomial logistic regression model to predict patients’ responses to the EQ-5D-3L questionnaire, based on their QLQ-C30 responses, age and sex (Longworth algorithm).^{15 16}

More recently, two additional studies reporting mapping algorithms for obtaining EQ-5D-3L utility values from QLQ-C30 responses have been published; neither of which have been externally validated. Marriott *et al.* report both mixed-effects and Tobit models, selecting the former as the ‘preferred’ algorithm.²³ Khan *et al.* report three algorithms, using beta-binomial, limited dependent variable mixture (LDVM) and random-effects models and select the beta-binomial model as the ‘preferred’ algorithm.²⁴ The two Marriott models and the three Khan models will be collectively referred to as the five existing mapping algorithms.

The algorithms presented by Khan *et al.* are some of the first mapping algorithms for the QLQ-C30 and the EQ-5D-3L that deviate from the use of linear regression methods. This

is not surprising, as a number of weaknesses of linear models when mapping have been identified, such as poor accuracy at the extremes of the EQ-5D-3L scale,⁷ assumptions that the EQ-5D-3L scale is continuous and unbounded²⁵ and that predictions from linear models are restricted to utility values for a single country-specific EQ-5D-3L tariff. Variants of more complex models that have been applied in the development of mapping algorithms for other HRQoL instruments have also largely been unexplored in the context of estimating EQ-5D-3L utility values from QLQ-C30 responses.^{26 27}

The objectives of this study are two-fold. First, to assess whether any of the five existing mapping algorithms, that have not been externally validated, offer an improvement in terms of predictive performance over previously identified, ‘best’ performing (Veersteegh and Longworth) algorithms. Second, to develop novel mapping algorithms that apply statistical methods that have not previously been used in the context of the QLQ-C30 and the EQ-5D-3L and assess whether such methods offer an improvement in predictive accuracy over those previously applied.

METHODS

Overview

Five existing mapping algorithms for predicting EQ-5D-3L utility values from QLQ-C30 responses that have not been assessed in an external validation study are validated using the complete Cancer 2015 dataset (validation dataset one).^{23 24} The performance of these five algorithms is compared to the ‘best’ performing algorithms previously identified in published external validation studies (Veersteegh and Longworth algorithms). The Cancer 2015 dataset is then split in two, with 300 patients randomly sampled to form validation dataset two and the remaining patients forming an estimation dataset. Novel mapping algorithms are developed using the estimation dataset. Validation dataset two is then split into a low baseline EQ-5D-3L utility value subset (validation dataset two A) and a high baseline EQ-5D-3L utility value subset (validation dataset two B) and used to validate the existing and novel mapping algorithms (Figure 1). All statistical analyses were performed using Stata version 13 (StataCorp LP, Texas USA).

Dataset

The Cancer 2015 dataset is derived from a longitudinal, prospective cohort of newly diagnosed adult cancer patients in Australia.²⁸ The dataset includes patients with over 20 different tumour types; the most common cancers being breast, prostate, head and neck,

colorectal and lung cancer. Patients have up to five follow-up points. Among the outcomes collected are the QLQ-C30 Version 3 and the EQ-5D-3L. The Cancer 2015 dataset was previously used to assess the external validity of ten existing mapping algorithms, but has not been used to estimate any mapping algorithms.¹⁹

The initial dataset contained 3,872 observations for 1,719 patients. After removing six observations with missing data on the QLQ-C30 or EQ-5D-3L the dataset used to externally validate the existing mapping algorithms (i.e., validation dataset one) contained 3,866 observations for 1,719 patients. To create the estimation dataset (1,419 patients, 3,203 observations), a random sample of 300 patients were removed. Data from these 300 patients (validation dataset two) were reserved to validate both the existing and newly developed mapping algorithms. The sample size of validation dataset two (300 patients) was chosen pragmatically to maximise the sample size of the estimation dataset and ensure sufficient patient numbers were still available to form two validation samples that could mimic the sample size of a small clinical trial (e.g., 150 patients). As potentially, similar sample sizes of QLQ-C30 data may be used to populate a mapping algorithm in practice. Validation dataset two was split in two by sorting the patients on baseline EQ-5D-3L utility value (low to high) and then numbering the patients 1 to 300. The patients were then randomly sorted into two groups, A and B, with the probability of the i^{th} patient being in group B being $i/300$. Group A contained 144 patients and group B 156. These two validation samples represent a low EQ-5D-3L dataset and a high EQ-5D-3L dataset respectively. Validation dataset two A has an average baseline utility of 0.68 and validation dataset two B has an average baseline utility of 0.87, the average baseline utility in the estimation dataset is 0.78. This allowed a comparison of the accuracy of algorithms in patients with above and below average utility values. Table 1 presents the baseline characteristics and a summary of the EQ-5D-3L utility values and QLQ-C30 summary scores for all four datasets.

Instruments

EORTC QLQ-C30

The latest version of the QLQ-C30 is version 3.0^{†,30} and describes 15 summary scales (six functioning and nine symptom scales), which can be produced from the 30 item responses.

[†] Version 3 of the QLQ-C30 in that it has a four-point scale for scale for items 1-7, while version 1 has a two-point scale for items 1-7⁴. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993;85 doi: 10.1093/jnci/85.5.365 and version 2 has a two-point scale for items 1-5.²⁹ Osoba D, Aaronson N, Zee B, et al. Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. The Study Group on Quality of Life of the EORTC and the Symptom Control and Quality of Life Committees of the NCI of

The functioning scales are physical, role, cognitive, emotional, social and global quality of life. The symptom scales are fatigue, nausea and vomiting, pain, dyspnoea, sleep disturbance, appetite loss, constipation, diarrhoea and financial impact.

EQ-5D

The EQ-5D-3L elicits information on five aspects of HRQoL, mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The EQ-5D-3L gives patients a choice of three responses for each of the five dimensions, resulting in a total of 243 possible health states. More recently, the EQ-5D-5L has been developed to improve the questionnaire's sensitivity,³¹ however, our study is limited to the validation and development of mapping algorithms for the EQ-5D-3L. Although it will be important to pursue the development of mapping algorithms for the EQ-5D-5L, in the foreseeable future algorithms for predicting EQ-5D-3L utility values will still form an important source of information for cost-utility analyses. Especially in the UK, where the National Institute for Health and Care Excellence (NICE) have explicitly stated a preference for analyses to use utility values derived from the EQ-5D-3L.

Observed EQ-5D-3L utility values for the external validation of the existing mapping algorithms were estimated using the same country-specific tariff as applied during the algorithms' development (e.g., UK and Dutch tariffs)^{32 33} to avoid accounting for any differences in social preferences in the analyses. In contrast, the newly developed algorithms were estimated and validated using the UK tariff, which results in utility values within the range of -0.594 to 1.

External validation of existing algorithms

The external validity of the five existing algorithms^{23 24} and two 'preferred' algorithms (Veersteegh and Longworth algorithms) selected from a previous external validation study¹⁹ as being the most appropriate (Table 2) was assessed in several steps as outlined below.

1. Estimate the predicted EQ-5D-3L utility values for all patients using the selected algorithms;
2. Plot the predicted and observed utility values, with lines to represent the best and worst QLQ-C30 health states when these values are entered into the algorithms as well as lines of best and perfect fit;
3. Calculate the errors for each algorithm;

4. Calculate the MAE and RMSE for each algorithm;
5. Repeat steps 3 and 4 for subsets of the EQ-5D-3L range;
6. Estimate the MAE across ten tumour sites; and
7. Compare the mean observed and predicted QALYs for each algorithm.

The ten tumour sites are those listed in Table 1. The subsets of the EQ-5D-3L range being used are $u \leq 0$, $0 < u \leq 0.5$, $0.5 < u \leq 0.75$ and $0.75 < u \leq 1$. Further details of each step are provided in Appendix A.

Estimation of new mapping algorithms

The models to be estimated fall into four categories, linear, response, mixture and beta models. Four types of linear mapping algorithms are estimated: OLS, fixed effects, random effects, and mixed effects models. Each of the linear models are also used alongside a logistic model to estimate perfect health. For each individual the logistic model produces a probability \hat{p}_i of perfect health and if this probability is greater than 0.5 the individual is assumed to have a utility value of 1. For individuals not predicted to be in perfect health (i.e., $\hat{p}_i \leq 0.5$) a utility value is assigned by the associated linear regression.

Two approaches to estimating a response mapping algorithm are applied, ordinal logistic regression and multinomial logistic regression. Both response mapping approaches avoid direct prediction of utility values and instead predict responses to each of the five EQ-5D-3L domains.^{25 34} The ordinal logistic regression approach takes into account the ordinal nature of the data, but relies on the data meeting the proportional odds assumption.³⁵ A multinomial logistic model does not require this assumption, but ignores the ordinal nature of the data and instead treats the data as categorical. These models account for the fact that utility values are not continuous by avoiding a direct prediction of the values and instead predicting an individual's response to each of the five EQ-5D-3L domains.

Mixture models are estimated using the 'aldvmm' command in Stata.³⁶ This command fits an adjusted limited dependent variable mixture model (ALDVMM), which takes into account the multimodality of the EQ-5D-3L distribution. It assumes that the EQ-5D-3L distribution is made up of C classes, where C is a user-defined number, to which it fits a linear model. These models are fit using both constant probabilities of class membership and using the combined health and quality of life scale from the QLQ-C30. The model is first fit using two classes and then the number of classes is increased until the models no longer converge.

For ease of readability, beyond this point mixture models will be referred to using the abbreviation ALDVMM, followed by the number of classes in the form ‘N’C, then by the letters CP or NCP for constant probabilities and non-constant probabilities respectively. Thus the 3 class model with variable mixing probabilities is known as ALDVMM-3C-NCP.

Two beta models are also estimated.²⁷ A generalised linear model is estimated using a logit link function and a binomial-link variance. Utility values are transformed onto the 0 to 1 scale prior to model estimation using the following formula $\frac{EQ5D - (-.594)}{1 - -.594}$. Predicted values must then be transformed back onto the -0.594 to 1 scale using the inverse transformation. This model is fit alone and with a logistic predictor for perfect health. These are referred to as the one-part and two-part beta models respectively.

Detailed descriptions of each statistical model used to estimate the newly developed algorithms are provided in Appendix B. All models estimated use the 15 summary scales as independent variables, either EQ-5D-3L utility values or responses to the five EQ-5D-3L domains (only for the ordinal and multinomial logistic regressions; response mapping algorithms) as dependent variables and no variables are eliminated from the models based on statistical significance.

Validation of all mapping algorithms and preferred model selection

There is little guidance in the literature as how to select the ‘best’ performing model. The existing and newly developed algorithms are validated based on seven criteria using validation datasets two A and B. Drawing on Longworth *et al.*¹⁶ each algorithm is ranked (a value of 1 representing the ‘best’ performing model) on the following criteria, for both validation datasets two A and B:

1. MAE;
2. RMSE;
3. MAE on the 0.75-1 subset of the EQ-5D-3L;
4. MAE on the 0.5-0.75 subset of the EQ-5D-3L;
5. MAE on the 0-0.5 subset of the EQ-5D-3L;
6. MAE on the <0 subset of the EQ-5D-3L; and
7. Error in the mean predicted QALYs compared to the mean observed QALYs.

The average rank across the criteria is then calculated for each model and the model with the best average rank across the two datasets (i.e., lowest value) will be selected as the ‘best’ performing algorithm.

RESULTS

External validation of existing mapping algorithms

Plots of the predicted EQ-5D-3L utility values against observed utility values highlight that the line of best fit diverges from the line of perfect fit as EQ-5D-3L utility values decrease, meaning that across all algorithms, model fit is worse for more severe disease states (Appendix C; Figures S1 to S7). All algorithms except the Versteegh algorithm¹⁵ predict EQ-5D-3L utility values outside the range of values defined by the best and worst QLQ-C30 health states. With the exception of the Khan beta-binomial model, average errors (range -0.016 to 0.031), MAE (range 0.096 to 0.116) and RMSE (0.091 to 0.113) are very similar for all the existing algorithms, with no algorithm clearly superior (Table 2). The MAEs across the range of EQ-5D-3L utility values show that all algorithms tend to be less accurate for lower EQ-5D utility values (Table 2). All algorithms have the smallest MAE on the prostate sub-group (range 0.065 to 0.121) and five of the seven algorithms have the largest MAE on the lung cancer sub-group, range 0.135 to 0.297 (Appendix D; Table S1).

The accuracy of each algorithm varies by tumour site, but it seems that the majority of algorithms consistently perform poorly when data from certain tumour site specific subgroups are applied. Tumour sites where all algorithms perform well tend to have higher average EQ-5D-3L utility values (Appendix D; Table S2). Both the Longworth¹⁶ and Khan LDVM²⁴ algorithms under predict mean QALYs when compared to the observed mean QALYs (-0.009 and -0.030 respectively) and the Longworth algorithm¹⁶ has the smallest error term (0.009) when compared to the observed QALYs (Appendix D; Table S3). It can also be seen that the error is significantly different from zero at the 5% level for all algorithms (Appendix D; Table S3).

It should be highlighted that the Khan beta-binomial model²⁴ performs poorly across all validation criteria. Upon further investigation we have identified some discrepancies in the model coefficients published in the original paper. The original paper provided utility values predicted from the model for the best and worst QLQ-C30 health states. We tried to replicate these results by inputting the respective best and worst QLQ-C30 scores into the model provided, but the results were not consistent with the utility values reported by Khan *et al.* (see

Appendix E and Table S4 for more details). It is believed that the model has been misspecified in the paper or insufficient detail has been provided to estimate the model correctly, therefore affecting the predictive accuracy of the algorithm in our external validation.

Overall, none of the five existing algorithms appear to represent a substantial improvement over those identified in the previous external validation study, highlighting that further refinement of the employed statistical models may be required to improve the prediction accuracy of algorithms.

Estimation of new mapping algorithms

Model coefficients for each algorithm are presented in Appendix F; Tables S5 to S15. Below are details of a number of statistical tests that can be used to guide selection of the most appropriate modelling approach within each of the four model categories (i.e., linear, response mapping, mixture and beta models) based on model fit.

Linear models

Linear models include the OLS model (R-squared=0.67, adjusted R-squared=0.67), the fixed (R-squared=0.67), random (R-squared=0.67) and mixed effects models, as well as all of the two part models that use a logistic map to predict perfect health. Model fit according to R-squared values is similar across the models and these values are comparable to those reported for the Veersteegh, Khan and Marriott models.^{15 23 24} The Hausmann test applied to the random-effects model rejects the null hypothesis ($p=0.0048$), indicating that the random-effects model is an inconsistent estimator and the fixed-effects model is the preferred specification. The logistic model of perfect health has a sensitivity of 81.8%, a specificity of 88.1% and correctly predicted if a patient was in perfect health for 86.0% of observations.

Response mapping

Ordinal and multinomial logistic models were fitted for each dimension of the EQ-5D-3L. Pseudo R-squared values for each domain of the EQ-5D-3L ranged from 0.396 to 0.474 and 0.403 to 0.482 for the ordinal and multinomial models respectively. Values were slightly better for the multinomial model across the domains and are comparable to those reported for the Longworth algorithm.¹⁶ Tests of the proportional odds assumption show that it holds for the self-care dimension of the EQ-5D-3L, but not for the mobility, usual activities, pain and discomfort or anxiety and depression dimensions, with significance assessed at the 5% level. This indicates that an ordinal model is inappropriate and the multinomial model is the preferred specification.

Mixture models

Six mixture models were estimated, two, three and four class models with both fixed and variable probabilities; beyond this, models would not converge. Adding the overall health and quality of life summary scale from the QLQ-C30 to the mixing probabilities for the two, three and four class models improved model fit. AIC and BIC values (AIC range -1894 to -2669 and BIC range -1682 to -2312) decreased when more classes were used and non-constant probabilities were used instead of constant probabilities. The ALDVMM-4C-NCP algorithm had the lowest BIC (-2311.9) and AIC (-2668.7), indicating that it is the best fitting of the ALDVMM models.

Beta models

The one-part beta model had a BIC of -25,445 and the two-part beta model had a BIC of -16,153, indicating that the former is the best fitting of the beta models.

Based on the statistical tests reported above, the random effects models, response mapping using ordinal logistic regression and ALDVMM models with a small number of classes and constant probabilities are unlikely to be the most appropriate model specifications to employ given that alternative modelling approaches can be used with improved model fit.

Validation of newly developed mapping algorithms

The mean error, RMSE, MAE and MAE across the EQ-5D-3L range for each algorithm are discussed below for both validation datasets two A and B and presented in Tables 3 and 4 respectively, with full validation results available in Appendix G; Tables S16 to S19.

Validation dataset two A – low baseline EQ-5D-3L utility value subgroup

The newly developed algorithms seem to perform better than the existing algorithms with RMSE values ranging from 0.107 to 0.114 and 0.115 to 0.252 respectively and MAE values ranging from 0.115 to 0.126 and 0.113 to 0.305 respectively. However, the overall variance in the RMSE and MAE values across all the algorithms is small. For all the newly developed linear models, the inclusion of a logistic predictor of perfect health results in both higher MAE and RMSE values. When the results are broken down across the EQ-5D-3L range, the logistic function improves fit in the EQ-5D-3L utility value range of 0.5 to 0.75, but does not seem to have an effect on the health states outside of this range.

When all criteria are accounted for together (see Methods section ‘Validation of all mapping algorithms and preferred model selection’), the multinomial response mapping

algorithm performs best on the low EQ-5D-3L dataset, as it has the lowest average rank (Table S18). The newly developed ALDVMM-4C-NCP, ALDVMM-3C-NCP, ordinal logistic and two-part beta models also perform well.

Validation dataset two B – high baseline EQ-5D-3L utility value subgroup

Overall the newly developed algorithms seem to perform better than the existing algorithms, though there is little difference between RMSE and MAE values. For all the newly developed linear models, the inclusion of a logistic predictor of perfect health results in an improvement in MAE but worse RMSE. When the results are broken down across the EQ-5D-3L range, the logistic function improves fit in the range 0-1, but it is important to note that this dataset does not include any observations with an EQ-5D-3L utility value below 0.

When all criteria are accounted for together (see Methods section ‘Validation of all mapping algorithms and preferred model selection’), the newly developed two-part beta mapping algorithm performs best on the high EQ-5D-3L dataset, as it has the lowest average rank (Table S19). The newly developed linear models with a logistic predictor of perfect health also perform well, as does the multinomial response mapping algorithm.

Preferred model selection

The algorithm with the lowest average rank across both validation dataset two A and B, and therefore the algorithm with best average performance, is the newly developed two-part beta algorithm (Table 5). This model, however, does not perform substantially better than the alternatives and a model performing well on the validation criteria for the low EQ-5D-3L dataset (validation dataset two A) does not appear to mean it will perform well on the high EQ-5D-3L dataset (validation dataset two B). This could potentially be explained by the differences in average utility values across the two datasets.

The second and third best performing models overall are the newly developed multinomial response mapping algorithm and the OLS+logit algorithm respectively. The multinomial response mapping algorithm performs well on the low EQ-5D-3L dataset and the OLS+logit algorithm performs well on the high EQ-5D-3L dataset.

DISCUSSION

The first aim of our research was to assess whether any of the five existing algorithms for mapping the QLQ-C30 to the EQ-5D-3L offer an improvement in predictive accuracy compared to algorithms that have previously undergone external validation. None of the five

existing algorithms appear to represent a substantial improvement over those identified in the previous external validation study.¹⁹ They perform worse on the lower end of the EQ-5D-3L range, where there is most need for more accurate prediction, and do not offer improvement in either the average errors or the estimated QALYs. This highlights that the more complex methods applied to date may need to be modified to further improve the prediction accuracy of algorithms for mapping the QLQ-C30 to the EQ-5D-3L.

Our second aim, therefore, attempted to assess whether any of the statistical methods for estimating mapping algorithms that have not previously been applied for predicting EQ-5D-3L utility values from QLQ-C30 responses perform better than the existing algorithms. Overall, the two-part beta model has the best fit across the two validation datasets (validation datasets two A and B). It should be noted, however, that the validation results did vary widely across the two validation datasets. One potential reason for this is that the distribution of utility values is different across the two datasets. It seems to be that simpler models (i.e., linear models) perform better, where a large number of patients are in perfect health or have relatively high utility values, whereas in datasets with generally worse health the more complex models (i.e., multinomial response mapping, ALDVMM-3C-NCP, ALDVMM-4C-NCP and two-part beta models) perform better.

A two-part beta model has previously been applied to the problem of mapping the Parkinson's Disease Questionnaire (PDQ-39) to the EQ-5D-3L.²⁷ These authors found the two-part beta model performed well when compared to other, more complex techniques, although they did not select it as the best performing model. Mixture models have also been shown to perform well in mapping to the EQ-5D-3L. Hernandez *et al.* found these models to be superior to linear models in mapping the Health Assessment Questionnaire (HAQ) to the EQ-5D-3L in a rheumatoid arthritis population.²⁶ The dataset they used shows a very complex distribution of utility values, with very clear evidence for a multimodal distribution, more so than in the dataset used for our study. Thus, it may be that different methods are suitable for producing mapping algorithms, depending on the complexity of the distribution. Simpler methods appear to be preferable when patients' utility values are gathered near one and the distribution is not multimodal. However, as EQ-5D-3L utility values decrease and the distribution becomes more variable, more complex methods become preferable. Future studies estimating mapping algorithms should take into account the range and complexity of the distribution of utility values in their dataset when deciding on the methods to apply.

More importantly, the results of our study highlight the influence of potentially unobservable factors on the accuracy of a mapping algorithm for a particular dataset. The validation of both existing and new algorithms has shown the importance of having datasets with comparable utility values for accurate prediction, as algorithms perform better on target datasets with a similar profile of EQ-5D-3L utility values as were observed in the estimation dataset regardless of tumour site. This is problematic as in practice the normal reason for needing to apply a mapping algorithm is that the dataset being used has not collected information on utility values, making a comparison of values across datasets impossible. This means that an important focus for research in this area going forward may be to identify variables that could be collected alongside non-preference-based measures of HRQoL like the QLQ-C30 that would give an indication of the most appropriate algorithm to apply.

The use of mixture models with mixing probabilities that are dependent on QLQ-C30 responses starts to answer this question and is potentially the reason for the good performance of the ALDVMM-3C-NCP and ALDVMM-4C-NCP algorithms. An area for future research would be in selecting the optimal set of variables for predicting class membership. For example, future researchers may wish to consider whether the inclusion of more of the QLQ-C30 summary scales as predictors of class membership improves model fit. Alternatively, commonly collected variables such as age, gender, Eastern Cooperative Oncology Group (ECOG) status or disease stage, which are not collected as part of the QLQ-C30, may be considered.

We have used a comprehensive approach to validate existing algorithms, estimate new algorithms and compare their performance to the existing algorithms using two datasets composed of patients with different levels of disease severity. Despite our approach our research does have some limitations. First, this research was limited by the fact that utility values in the Cancer 2015 dataset are relatively high. One of the key issues for mapping algorithms generally and for algorithms mapping the QLQ-C30 to EQ-5D-3L in particular, is poor fit on lower EQ-5D-3L ranges. In order to mitigate this, we have tested the validity of algorithms on datasets with both high and low average utility values.

Second, our assessment of mapping algorithms for the QLQ-C30 is specific to the EQ-5D-3L and our conclusions might not be applicable to algorithms developed for predicting EQ-5D-5L utility values from QLQ-C30 responses. In particular, the poor prediction of low values on the EQ-5D-3L scale may be less of an issue when predicting EQ-5D-5L values, given EQ-

5D-5L values are more evenly spread across the scale. However, further research using alternative datasets that have collected both EQ-5D-5L and QLQ-C30 responses will be necessary to address this issue.

Third, a more general limitation of all mapping algorithms is that their predictive accuracy will be limited by the level of overlap between the two instruments in terms of what they attempt to measure. Given the generic nature of the EQ-5D-3L, it is likely that some important aspects of HRQoL captured by the QLQ-C30 and specific to cancer patients will not be incorporated into the mapped EQ-5D-3L utility values. This ultimately means that application of more complex statistical models will potentially only marginally improve the predictive accuracy of mapping algorithms and that some level of discordance will always be present.

Finally, while this research sets aside a cohort of patients from the dataset for the purposes of validation, it would be preferable to have a completely external dataset for validation, as the validation samples may still be more similar to the estimation sample than a completely external dataset. It would be beneficial for the algorithms estimated as part of this research to be externally validated before they are applied in an economic evaluation.

In conclusion, this research has shown that none of the existing mapping algorithms that have not been previously externally validated offer a significant improvement in terms of predictive accuracy. It also highlights that the most appropriate algorithm to apply in practice is dependent on the disease severity of the patient sample whose utility values are being predicted. Linear models tend to perform better for patients in relatively good health, whereas beta, mixture and response mapping models perform better for patients in worse health. However, the two-part beta model estimated in our study has been shown to offer good predictions in patient samples with different disease severity and should therefore be of primary consideration when mapping the QLQ-C30 to the EQ-5D-3L.

ACKNOWLEDGMENTS

Cancer 2015 is funded by the Victorian Cancer Agency Translational Research Program. We would like to sincerely thank all the cancer patients who agreed to participate in the cohort. We acknowledge the contributions of the following staff and collaborators of this multi-site cohort: John P Parisot, Kristy Barnes-Cullen, Kate Crough, Jessica McDonald, Emma Galligan, Ann Officer, Anne Fennessy, and Sonia Mailer from the Peter MacCallum Cancer Centre; Kate Richards, Laura Zamurs, Kate Hurford, Rachel Osborne and Jennifer MacIndoe from Cabrini Health; Carolyn Wielens, Lea-Anne Harrison, Judi Broad, Robert Swiger, Tina Smith and Anne Woollett from The Andrew Love Cancer Centre, Barwon Health; Sandra Robinson, Marcelle Hennig and Ashlin Keane from Oncology Trials Department, South West Healthcare; Monica Merceica, Stefanie Hartley, Pat Bugeja, Lidia Veca, Christopher Bates and Nicole Ng from The Royal Melbourne Hospital, Melbourne Health.

CANCER 2015 Consortium

Stephen B Fox, Division of Cancer Research, Peter MacCallum Cancer Centre; Department of Pathology, Peter MacCallum Cancer Centre. Ian Collins, Warrnambool Hospital, South West Healthcare. Theresa Hayes, Warrnambool Hospital, South West Healthcare. Madhu Singh, The Andrew Love Cancer Centre, Geelong Hospital, Barwon Health. Gary Richardson, Department Haematology & Oncology, Cabrini Health. Lara Lipton, Department of Medical Oncology, The Royal Melbourne Hospital. So-Young Moon, Division of Cancer Research, Peter MacCallum Cancer Centre. Mark Lucas, Clinical Data Management Systems, Monash University. Andrew Fellowes, Department of Pathology, Peter MacCallum Cancer Centre. Huiling Xu, Department of Pathology, Peter MacCallum Cancer Centre. Heather Thorne, Division of Cancer Research, Peter MacCallum Cancer Centre. John J McNeil, Department of Epidemiology and Preventative Medicine, Alfred Centre, Monash University. Paula Lorgelly, Office of Health Economics. David M Thomas, Division of Cancer Research, Peter MacCallum Cancer Centre; The Kinghorn Cancer Centre and Garvan Institute. Paul A James, Division of Cancer Medicine, Peter MacCallum Cancer Centre. Tomas John, Department of Medical Oncology, Olivia Newton John Cancer and Wellness Centre, Austin Health. Gail Risbridger, Division of Cancer Research, Peter MacCallum Cancer Centre. Gavin Wright, Department Surgical Oncology, St Vincent's Hospital. Raymond Snyder, Department of Oncology, St Vincent's Hospital.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interest.

REFERENCES

1. Brooks R. EuroQol: the current state of play. *Health Policy* 1996;37(1):53-72. [published Online First: 1996/06/06]
2. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Med Care* 2004;42(9):851-9. [published Online First: 2004/08/21]
3. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care* 2002;40(2):113-28. [published Online First: 2002/01/22]
4. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993;85 doi: 10.1093/jnci/85.5.365
5. Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 1993;11(3):570-9. doi: 10.1200/jco.1993.11.3.570 [published Online First: 1993/03/01]
6. NICE. Guide to the methods of technology appraisal. Available at: <https://www.nice.org.uk/article/pmg9/resources/non-guidance-guide-to-the-methods-of-technology-appraisal-2013-pdf> Last accessed 09/06/2016. 2013
7. Brazier JE, Yang Y, Tsuchiya A, et al. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010;11 doi: 10.1007/s10198-009-0168-z
8. McKenzie L, van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2009;12(1):167-71. doi: 10.1111/j.1524-4733.2008.00405.x [published Online First: 2008/07/19]
9. Kontodimopoulos N, Aletras VH, Paliouras D, et al. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2009;12(8):1151-7. doi: 10.1111/j.1524-4733.2009.00569.x [published Online First: 2009/06/30]
10. Jang RW, Isogai PK, Mittmann N, et al. Derivation of utility values from European Organization for Research and Treatment of Cancer Quality of Life-Core 30

- questionnaire values in lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* 2010;5(12):1953-7. [published Online First: 2010/12/16]
11. Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. *Eur J Health Econ* 2010;11(4):427-34. doi: 10.1007/s10198-010-0233-7 [published Online First: 2010/05/18]
 12. Versteegh MM, Rowen D, Brazier JE, et al. Mapping onto Eq-5 D for patients in poor health. *Health Qual Life Outcomes* 2010;8:141. doi: 10.1186/1477-7525-8-141 [published Online First: 2010/11/30]
 13. Kim EJ, Ko SK, Kang HY. Mapping the cancer-specific EORTC QLQ-C30 and EORTC QLQ-BR23 to the generic EQ-5D in metastatic breast cancer patients. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2012;21(7):1193-203. doi: 10.1007/s11136-011-0037-y [published Online First: 2011/10/21]
 14. Kim SH, Jo M-W, Kim H-J, et al. Mapping EORTC QLQ-C30 onto EQ-5D for the assessment of cancer patients. *Health and Quality of Life Outcomes* 2012;10(1):151. doi: 10.1186/1477-7525-10-151
 15. Versteegh MM, Leunis A, Luime JJ, et al. Mapping QLQ-C30, HAQ, and MSIS-29 on EQ-5D. *Medical decision making : an international journal of the Society for Medical Decision Making* 2012;32(4):554-68. doi: 10.1177/0272989x11427761 [published Online First: 2011/11/25]
 16. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health technology assessment (Winchester, England)* 2014;18(9):1-224. doi: 10.3310/hta18090 [published Online First: 2014/02/15]
 17. Proskorovsky I, Lewis P, Williams CD, et al. Mapping EORTC QLQ-C30 and QLQ-MY20 to EQ-5D in patients with multiple myeloma. *Health Qual Life Outcomes* 2014;12:35. doi: 10.1186/1477-7525-12-35 [published Online First: 2014/03/13]
 18. Khan I, Morris S. A non-linear beta-binomial regression model for mapping EORTC QLQ- C30 to the EQ-5D-3L in lung cancer patients: a comparison with existing approaches. *Health Qual Life Outcomes* 2014;12:163. doi: 10.1186/s12955-014-0163-7 [published Online First: 2014/11/13]
 19. Doble B, Lorgelly P. Mapping the EORTC QLQ-C30 onto the EQ-5D-3L: assessing the external validity of existing mapping algorithms. *Quality of life research : an*

- international journal of quality of life aspects of treatment, care and rehabilitation* 2016;25(4):891-911. doi: 10.1007/s11136-015-1116-2 [published Online First: 2015/09/24]
20. Arnold DT, Rowen D, Versteegh MM, et al. Testing mapping algorithms of the cancer-specific EORTC QLQ-C30 onto EQ-5D in malignant mesothelioma. *Health and Quality of Life Outcomes* 2015;13(1):6. doi: 10.1186/s12955-014-0196-y
 21. Crott R, Versteegh M, Uyl-de-Groot C. An assessment of the external validity of mapping QLQ-C30 to EQ-5D preferences. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2013;22(5):1045-54. doi: 10.1007/s11136-012-0220-9 [published Online First: 2012/06/30]
 22. Crott R. Mapping algorithms from QLQ-C30 to EQ-5D utilities: no firm ground to stand on yet. *Expert review of pharmacoeconomics & outcomes research* 2014;14(4):569-76. doi: 10.1586/14737167.2014.908711 [published Online First: 2014/06/10]
 23. Marriott ER, van Hazel G, Gibbs P, et al. Mapping EORTC-QLQ-C30 to EQ-5D-3L in patients with colorectal cancer. *Journal of medical economics* 2016:1-7. doi: 10.1080/13696998.2016.1241788 [published Online First: 2016/09/28]
 24. Khan I, Morris S, Pashayan N, et al. Comparing the mapping between EQ-5D-5L, EQ-5D-3L and the EORTC-QLQ-C30 in non-small cell lung cancer patients. *Health and Quality of Life Outcomes* 2016;14(1):1-15. doi: 10.1186/s12955-016-0455-1
 25. Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical decision making : an international journal of the Society for Medical Decision Making* 2006;26(1):18-29. doi: 10.1177/0272989x05284108 [published Online First: 2006/02/24]
 26. Hernández Alava M, Wailoo A, Wolfe F, et al. The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis. *Rheumatology (Oxford, England)* 2013;52(5):944-50. doi: 10.1093/rheumatology/kes400
 27. Kent S, Gray A, Schlackow I, et al. Mapping from the Parkinson's Disease Questionnaire PDQ-39 to the Generic EuroQol EQ-5D-3L: The Value of Mixture Models. *Medical decision making : an international journal of the Society for Medical Decision Making* 2015;35(7):902-11. doi: 10.1177/0272989x15584921 [published Online First: 2015/05/01]
 28. Parisot JP, Thorne H, Fellowes A, et al. "Cancer 2015": A Prospective, Population-Based Cancer Cohort-Phase 1: Feasibility of Genomics-Guided Precision Medicine in the

- Clinic. *Journal of personalized medicine* 2015;5(4):354-69. doi: 10.3390/jpm5040354 [published Online First: 2015/11/04]
29. Osoba D, Aaronson N, Zee B, et al. Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. The Study Group on Quality of Life of the EORTC and the Symptom Control and Quality of Life Committees of the NCI of Canada Clinical Trials Group. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 1997;6(2):103-8. [published Online First: 1997/03/01]
 30. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993;85(5):365-76. [published Online First: 1993/03/03]
 31. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2011;20(10):1727-36. doi: 10.1007/s11136-011-9903-x [published Online First: 2011/04/12]
 32. Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997;35 doi: 10.1097/00005650-199711000-00002
 33. Lamers LM, McDonnell J, Stalmeier PF, et al. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. *Health economics* 2006;15(10):1121-32. doi: 10.1002/hec.1124 [published Online First: 2006/06/21]
 34. Young MK, Ng SK, Mellick G, et al. Mapping of the PDQ-39 to EQ-5D scores in patients with Parkinson's disease. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2013;22(5):1065-72. doi: 10.1007/s11136-012-0231-6 [published Online First: 2012/07/12]
 35. Stata. Stata: Release 13. Statistical Software. College Station, TX: StataCorp LP. 2013
 36. Hernández Alava M, Wailoo A. Fitting adjusted limited dependent variable mixture models to EQ-5D. *Stata Journal* 2015;15(3):737-50.
 37. Winter B. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. Available from <http://arxiv.org/pdf/1308.5499.pdf>. Last accessed 13/12/2016. 2013
 38. Dakin H, Gray A, Murray D. Mapping analyses to estimate EQ-5D utilities and responses based on Oxford Knee Score. *Quality of life research : an international journal of*

quality of life aspects of treatment, care and rehabilitation 2013;22(3):683-94. doi: 10.1007/s11136-012-0189-4

39. Hernandez Alava M, Wailoo AJ, Ara R. Tails from the peak district: adjusted limited dependent variable mixture models of EQ-5D questionnaire health state utility values. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2012;15(3):550-61. doi: 10.1016/j.jval.2011.12.014 [published Online First: 2012/05/16]

Figure 1. Overview of the validation and estimation datasets

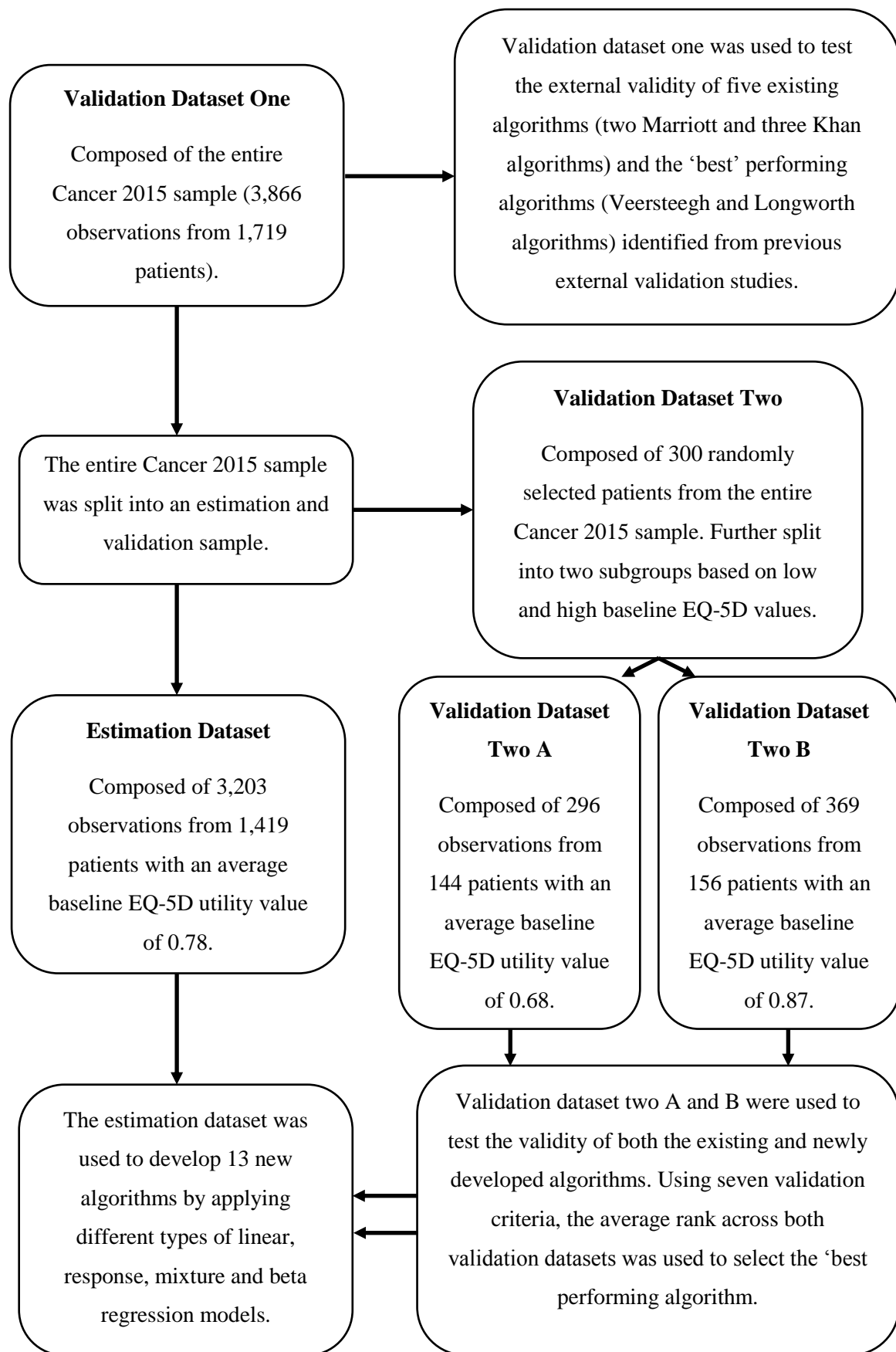


Table 1. Baseline characteristics and summary EQ-5D-3L utility values and QLQ-C30 summary scores for all for datasets

		Validation dataset one (N=1,719)		Estimation dataset (N=1,419)		Validation dataset two A (N=144)		Validation dataset two B (N=156)	
Variable	Label	Mean/N	SD/%	Mean/N	SD/%	Mean/N	SD/%	Mean/N	SD/%
Age at consent		61.55	12.54	61.45	12.45	62.74	13.78	61.42	12.09
Charlson comorbidity index		2.16	1.67	2.14	1.67	2.34	1.6	2.14	1.71
Sex	Male	938	54.57	779	54.9	73	50.69	86	55.13
	Female	781	45.43	640	45.1	71	49.31	70	44.87
Recruiting hospital	Public	1,430	83.19	1,182	83.3	122	84.72	126	80.77
	Private	289	16.81	237	16.7	22	15.28	30	19.23
Smoking status	Current smoker	255	14.83	217	15.29	21	14.58	17	10.9
	Ex-smoker	794	46.19	659	46.44	64	44.44	71	45.51
	Never smoked	670	38.98	543	38.27	59	40.97	68	43.59
ECOG score	Normal Activity	1,199	69.75	992	69.91	85	59.03	122	78.21
	Limited	379	22.05	308	21.71	44	30.56	27	17.31
	Self-care	108	6.28	91	6.41	10	6.94	7	4.49
	Limited	32	1.86	27	1.9	5	3.47	0	0
	No	1	0.06	1	0.07	0	0	0	0
Tumour site	Breast	348	20.24	297	20.93	22	15.28	29	18.59
	Prostate	278	16.17	231	16.28	20	13.89	27	17.31
	Head and neck	261	15.18	214	15.08	24	16.67	23	14.74
	Colorectal	179	10.41	148	10.43	13	9.03	18	11.54
	Lung	148	8.61	128	9.02	12	8.33	8	5.13
	Oesophagogastric	59	3.43	45	3.17	5	3.47	9	5.77
	Renal	76	4.42	67	4.72	4	2.78	5	3.21
	Bone and soft tissue	55	3.2	41	2.89	6	4.17	8	5.13
	Cancer of unknown primary	29	1.69	21	1.48	6	4.17	2	1.28
	Cervical	39	2.27	30	2.11	3	2.08	6	3.85
	Other	247	14.37	197	13.88	29	20.14	21	13.46
Stage	0	14	0.97	12	1.01	1	0.88	1	0.73
	1	377	26.22	315	26.54	31	27.19	31	22.63
	2	424	29.49	353	29.74	31	27.19	40	29.2

		Validation dataset one (N=1,719)		Estimation dataset (N=1,419)		Validation dataset two A (N=144)		Validation dataset two B (N=156)	
Variable	Label	Mean/N	SD/%	Mean/N	SD/%	Mean/N	SD/%	Mean/N	SD/%
	3	307	21.35	259	21.82	23	20.18	25	18.25
	4	280	19.47	217	18.28	27	23.68	36	26.28
	Other	36	2.5	31	2.61	1	0.88	4	2.92
EQ-5D-3L utility value		0.78	0.24	0.78	0.24	0.68	0.28	0.87	0.16
OQL		72.51	23.35	72.43	23.35	65.71	24.74	78.64	20.54
PF		84.22	19.4	84.17	19.43	77.68	22.25	89.88	14.39
RF		77.09	29.48	76.89	29.52	68.81	32.75	85.55	23.71
EF		78.5	22.19	78.31	22.21	74.72	23.31	83.18	20.24
CF		82.85	21.13	82.79	21.04	78.32	23.85	86.99	18.58
SF		78.87	27.15	78.96	27.02	69.82	31.49	85.37	22.11
FA		28.46	24.78	28.48	24.64	36.04	28.2	22.19	21.16
NV		6.13	14.4	6.07	14.22	9.63	19.94	3.79	9.23
PA		20.64	26.37	20.81	26.4	28.89	30.78	12.47	18.99
DY		14.55	23.69	14.84	23.99	18.47	26.26	8.85	17.02
SL		28.41	30.04	28.22	29.62	34.12	33.32	25.47	30.42
AP		14.87	26.16	14.72	25.81	19.37	31.18	12.65	24.38
CO		11.9	23.11	11.84	23.11	17.68	27.84	7.77	17.2
DI		8.15	18.39	8.15	18.51	8.9	18.4	7.59	17.28
FI		17.86	28.51	17.96	28.7	24.32	32.38	11.74	21.44

Abbreviations: ECOG, Eastern Cooperative Oncology Group; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problem

Table 2. External validation results of the ‘best’ performing mapping algorithms from previous external validation studies and five existing mapping algorithms that have not been externally validated using validation dataset one

Algorithm ^{a,b} (Method applied to estimate algorithm)	Undergone previous external validation	Observations	Error	RMSE	MAE	MAE across the EQ-5D-3L range			
						0.75-1	0.5-0.75	0-0.5	<0
Longworth¹⁶ (Response mapping)	Yes	3866	-0.007	0.093	0.096	0.065	0.121	0.238	0.245
Veersteegh¹⁵ (OLS regression)	Yes	3866	0.031	0.113	0.103	0.070	0.128	0.302	0.356
Khan Random Effects²⁴ (Random effects linear regression)	No	3866	0.023	0.095	0.106	0.072	0.115	0.297	0.443
Khan Beta-Binomial²⁴ (Beta-binomial regression)	No	3866	0.212	0.194	0.212	0.086	0.319	0.717	1.014
Khan LDVMM²⁴ (LDVMM)	No	3866	-0.016	0.097	0.116	0.094	0.104	0.282	0.433
Marriott Mixed Effects²³ (Mixed effects linear regression)	No	3866	0.027	0.093	0.101	0.066	0.099	0.316	0.497
Marriott Tobit²³ (Tobit model)	No	3866	0.023	0.091	0.099	0.065	0.099	0.302	0.472

Abbreviations: RMSE, Root mean standardised error; MAE, Mean absolute error; AE, Absolute error; OLS, Ordinary least squares, LDVMM, Limited dependent variable mixture model.

^aAll the existing algorithms were estimated using responses from QLQ-C30 version 3.0, with the exception of the Veersteegh algorithm, which used responses from QLQ-C30 version 2.0.

^bAll the existing algorithms were estimated using observed utility values derived using the United Kingdom EQ-5D-3L value set (tariff), with the exception of the Veersteegh algorithm, which used observed utility derived using the Dutch EQ-5D-3L value set (tariff).

Table 3. Average error, RMSE, MAE and MAE across the EQ-5D-3L range for each algorithm when validated using the low EQ-5D-3L dataset (validation dataset two A)

	Observations	Error	RMSE	MAE	MAE across EQ-5D-3L range			
					0.75-1	0.5-0.75	0-0.5	<0
Longworth	296	0.006	0.119	0.120	296	0.006	0.119	0.120
Veersteegh	296	0.034	0.123	0.113	296	0.034	0.123	0.113
Khan RE	296	0.060	0.122	0.133	296	0.060	0.122	0.133
Khan BB	296	0.305	0.252	0.305	296	0.305	0.252	0.305
Khan MM	296	0.026	0.115	0.131	296	0.026	0.115	0.131
Marriott FE	296	0.072	0.121	0.128	296	0.072	0.121	0.128
Marriott Tobit	296	0.065	0.118	0.126	296	0.065	0.118	0.126
OLS	296	0.031	0.108	0.120	296	0.031	0.108	0.120
OLS + logit	296	0.025	0.110	0.122	296	0.025	0.110	0.122
FE	296	0.038	0.112	0.124	296	0.038	0.112	0.124
FE + logit	296	0.029	0.113	0.124	296	0.029	0.113	0.124
RE	296	0.031	0.108	0.121	296	0.031	0.108	0.121
RE + logit	296	0.024	0.110	0.122	296	0.024	0.110	0.122
ME	296	0.031	0.108	0.120	296	0.031	0.108	0.120
ME + logit	296	0.024	0.110	0.122	296	0.024	0.110	0.122
Ordinal	296	0.026	0.112	0.116	296	0.026	0.112	0.116
Multinomial	296	0.024	0.110	0.115	296	0.024	0.110	0.115
ALDVMM-2C-CP	296	0.042	0.110	0.124	296	0.042	0.110	0.124
ALDVMM-2C-NCP	296	0.038	0.110	0.122	296	0.038	0.110	0.122
ALDVMM-3C-CP	296	0.041	0.110	0.124	296	0.041	0.110	0.124
ALDVMM-3C-NCP	296	0.035	0.108	0.119	296	0.035	0.108	0.119
ALDVMM-4C-CP	296	0.035	0.108	0.121	296	0.035	0.108	0.121
ALDVMM-4C-NCP	296	0.035	0.107	0.118	296	0.035	0.107	0.118
One-part beta	296	0.037	0.114	0.126	296	0.037	0.114	0.126
Two-part beta	296	0.028	0.109	0.117	296	0.028	0.109	0.117

Abbreviations: RMSE, Root mean squared error; MAE, Mean absolute error; OLS, Ordinary least squares; FE, Fixed effects; RE, Random effects; ME, Mixed effects.

Table 4. Average error, RMSE, MAE and MAE across the EQ-5D-3L range for each algorithm when validated using the high EQ-5D-3L dataset (validation dataset two B)

	Observations	Error	RMSE	MAE	MAE across EQ-5D-3L range			
					0.75-1	0.5-0.75	0-0.5	<0
Longworth	369	-0.024	0.076	0.076	0.062	0.116	0.249	0.000
Veersteegh	369	-0.011	0.080	0.071	0.057	0.119	0.278	0.000
Khan RE	369	-0.005	0.073	0.086	0.069	0.128	0.316	0.000
Khan BB	369	0.126	0.126	0.126	0.068	0.308	0.705	0.000
Khan MM	369	-0.051	0.080	0.105	0.100	0.105	0.287	0.000
Marriott FE	369	-0.009	0.069	0.078	0.063	0.108	0.329	0.000
Marriott Tobit	369	-0.011	0.069	0.077	0.063	0.107	0.322	0.000
OLS	369	-0.024	0.069	0.077	0.065	0.105	0.265	0.000
OLS + logit	369	-0.022	0.074	0.072	0.060	0.097	0.257	0.000
FE	369	-0.030	0.068	0.081	0.071	0.097	0.285	0.000
FE + logit	369	-0.023	0.075	0.072	0.063	0.089	0.266	0.000
RE	369	-0.027	0.069	0.078	0.066	0.104	0.266	0.000
RE + logit	369	-0.023	0.075	0.072	0.061	0.097	0.256	0.000
ME	369	-0.025	0.069	0.077	0.065	0.105	0.265	0.000
ME + logit	369	-0.022	0.075	0.072	0.061	0.097	0.256	0.000
Ordinal	369	-0.029	0.071	0.079	0.069	0.102	0.251	0.000
Multinomial	369	-0.030	0.068	0.078	0.068	0.099	0.239	0.000
ALDVMM-2C-CP	369	-0.029	0.069	0.084	0.074	0.103	0.280	0.000
ALDVMM-2C-NCP	369	-0.029	0.070	0.084	0.072	0.108	0.281	0.000
ALDVMM-3C-CP	369	-0.029	0.069	0.083	0.073	0.103	0.278	0.000
ALDVMM-3C-NCP	369	-0.030	0.070	0.079	0.070	0.094	0.276	0.000
ALDVMM-4C-CP	369	-0.031	0.069	0.083	0.073	0.101	0.271	0.000
ALDVMM-4C-NCP	369	-0.029	0.069	0.078	0.069	0.094	0.275	0.000
One-part beta	369	-0.031	0.076	0.094	0.080	0.130	0.286	0.000
Two-part beta	369	-0.020	0.074	0.070	0.059	0.093	0.251	0.000

Abbreviations: RMSE, Root mean squared error; MAE, Mean absolute error; OLS, Ordinary least squares; FE, Fixed effects; RE, Random effects; ME, Mixed effects.

Table 5. Average rank and overall rank for each algorithm across validation dataset two A and B for all validation criteria (RMSE, MAE, MAE across the EQ-5D-3L range and QALY error)

Algorithm	Average rank on validation dataset two A	Average rank on validation dataset two B	Average rank on both validation datasets	Overall rank
Two part beta	7.3	4.8	6.2	1
Multinomial	5.6	9.0	7.2	2
OLS + logit	11.1	6.8	9.2	3
OLS	9.7	10.2	9.9	4
ME + logit	11.1	8.5	9.9	4
RE + logit	12.0	7.8	10.1	6
ALDVMM-4C-NCP	8.0	12.7	10.2	7
ME	10.3	10.5	10.4	8
Ordinal	8.6	12.7	10.5	9
Veersteegh	10.7	10.8	10.8	10
RE	11.0	11.3	11.2	11
Longworth	11.7	10.8	11.3	12
ALDVMM-3C-NCP	9.0	14.3	11.5	13
FE + logit	14.3	8.8	11.8	14
ALDVMM-4C-CP	10.7	15.5	12.9	15
Marriott Tobit	16.7	11.2	14.2	16
Marriott FE	17.1	11.8	14.7	17
FE	15.9	14.2	15.1	18
ALDVMM-2C-NCP	12.7	18.2	15.2	19
ALDVMM-3C-CP	15.9	15.2	15.5	20
ALDVMM-2C-CP	15.6	16.3	15.9	21
One part beta	15.4	21.5	18.2	22
Khan RE	21.6	16.5	19.2	23
Khan MM	18.0	22.5	20.1	24
Khan BB	25.0	23.0	24.1	25

Abbreviations: RMSE, Root mean squared error; MAE, Mean absolute error; QALY, Quality-adjusted life-year; OLS, Ordinary least squares; FE, Fixed effects; RE, Random effects; ME, Mixed effects

SUPPLEMENTAL INFORMATION

TABLE OF CONTENTS

APPENDIX A - EXTERNAL VALIDATION APPROACH	32
APPENDIX B - DESCRIPTIONS OF THE STATISTICAL MODELS USED TO ESTIMATE THE NEWLY DEVELOPED MAPPING ALGORITHMS	35
<i>Linear Regression</i>	35
<i>Fixed, random and mixed effects</i>	35
<i>Logistic models of perfect health</i>	36
<i>Response mapping</i>	37
<i>Mixture models</i>	39
<i>Beta models</i>	41
APPENDIX C - SCATTER PLOTS FOR THE EXTERNAL VALIDATION OF EXISTING MAPPING ALGORITHMS	42
Figure S1. Scatter plot of observed vs. predicted utility values for the Longworth response mapping algorithm	42
Figure S2. Scatter plot of observed vs. predicted utility values for the Veersteegh OLS mapping algorithm	42
Figure S3. Scatter plot of observed vs. predicted utility values for the Khan random effects mapping algorithm	43
Figure S4. Scatter plot of observed vs. predicted utility values for the Khan beta-binomial mapping algorithm	43
Figure S5. Scatter plot of observed vs. predicted utility values for the Khan LDVM mapping algorithm	44
Figure S6. Scatter plot of observed vs. predicted utility values for the Marriott mixed effects mapping algorithm.....	44
Figure S7. Scatter plot of observed vs. predicted utility values for the Marriott Tobit mapping algorithm	45

APPENDIX D - ADDITIONAL EXTERNAL VALIDATION RESULTS FOR EXISTING MAPPING ALGORITHMS	46
Table S1. Mean absolute error for each algorithm for each of the ten tumour sites from external validation dataset one	46
Table S2. Rankings by mean absolute error for each algorithm using tumour site specific sub-groups from validation dataset one and the mean observed EQ-5D-3L utility value for each tumour site	47
Table S3. Observed and predicted QALYs for each algorithm using validation dataset one	48
APPENDIX E - REPLICATION OF RESULTS PRESENTED BY KHAN ET AL. FOR THE BETA BINOMIAL MODEL	49
Table S4. Data used to replicate the predicted utilities reported by Khan et al.....	50
APPENDIX F- MODEL COEFFICIENTS FOR THE NEWLY DEVELOPED MAPPING ALGORITHMS	51
Table S5. Linear model coefficients.....	51
Table S6. Ordinal logistic model coefficients	52
Table S7. Multinomial logistic model coefficients – response 2 versus response 1	53
Table S8. Multinomial logistic model coefficients – response 3 versus response 1	54
Table S9. ALDVMM-2C-CP coefficients.....	55
Table S10. ALDVMM-2C-NCP coefficients.....	56
Table S11. ALDVMM-3C-CP coefficients.....	57
Table S12. ALDVMM-3C-NCP coefficients	58
Table S13. ALDVMM-4C-CP coefficients.....	59
Table S14. ALDVMM-4C-NCP coefficients.....	60
Table S15. Beta models	61
APPENDIX G - ADDITIONAL RESULTS FROM THE VALIDATION OF THE NEWLY DEVELOPED MAPPING ALGORITHMS USING VALIDATION DATASET TWO A AND B.....	62
Table S16. Validation dataset two A - Observed and predicted QALYs	63

Table S17. Validation dataset two B - Observed and predicted QALYs	64
Table S18. Ranking of algorithms on each validation criteria using validation dataset two A	65
Table S19. Ranking of algorithms on each validation criteria using validation dataset two B	66

APPENDIX A - EXTERNAL VALIDATION APPROACH

Step 1: Estimating predicted utility values

This step involves using the selected algorithms to predict patients' EQ-5D-3L utility values.

Step 2: Plotting the predicted and observed utility values

A scatter plot showing the observed utility values along the x-axis and the predicted utility values along the y-axis, a line of best fit, a line of perfect fit and line at the EQ-5D-3L utility values predicted for the best and worst possible QLQ-C30 health states is created.

The best possible health state is defined to be the health state to which the patient has responded with the most positive response to each of the QLQ-C30 questions, thus scores 100 on each of the functioning scales and 0 on each of the symptom scales. Conversely, the worst possible health state is defined to be the health state where the patient has given the worst possible response to each of the QLQ-C30 questions, so scores 0 on each of the functioning scales and 100 on each of the symptom scales.

Step 3: Calculating the errors

The error is the observed EQ-5D-3L utility value, minus the predicted utility value.

Step 4: Calculating the MAE and RMSE

The MAE is the average values of the error term. The RMSE is the average of the squared error term, divided by the range of possible utility values for the EQ-5D-3L tariff in question. This is done to ensure comparability of results across tariffs.

These measures provide an estimate of the total deviation from the true EQ-5D-3L utility values and are more useful than looking only at the mean error which should have a mean of zero. The difference between MAE and RMSE comes from the squared term in the RMSE formula. This means that the RMSE gives extra weight to larger errors. The RMSE will always be at least as large as the MAE and they will only be equal if all the errors are equal. The difference between the MAE and RMSE gives an indication of the variance in the error term.

Step 5: Repeat for subsets of the EQ-5D-3L scale

The predictive accuracy of a mapping algorithm may not be constant across the range of possible EQ-5D-3L utility values, especially if the dataset used to estimate the algorithm did

not have a large amount of data on patients in certain ranges. Thus, it is important to assess the accuracy of an algorithm on subsets of the EQ-5D-3L scale.

The ranges being used are $u \leq 0$, $0 < u \leq 0.5$, $0.5 < u \leq 0.75$ and $0.75 < u \leq 1$.

Step 6: Estimate the MAE across 10 tumour sites

Algorithms estimated in patients with a specific cancer site may not be generalizable to other cancer populations. Results are compared across tumour sites to assess the validity of an algorithm for predicting utility values for patients with a tumour site that was not present in the dataset used to estimate the algorithm. The tumour sites studied here are listed below:

- Breast;
- Prostate;
- Head and neck;
- Colorectal;
- Lung;
- Oesophagogastric;
- Renal;
- Bone and soft tissue;
- Cancer of unknown primary;
- Cervical; and
- Other.

“Other” cancers include anal, biliary, bladder, cancer of the central nervous system, endometrial, hepatic, lymphoma, melanoma, ovarian, pancreatic, testicular and thyroid.

Step 7: Compare observed and predicted QALYs for each patient

Ultimately, algorithms will be used to predict QALYs in a health economic model. Thus, it is important to attempt to assess the effect each algorithm may have on these results. In order to assess an algorithm’s accuracy in predicting QALYs, the observed QALYs must be calculated first. This can only be done for patients with at least 2 observations in the dataset. To estimate the QALYs gained between two observations, the time in years between observations is calculated and then multiplied by the average of the observed utility value for that patient at the two time points. This is repeated for all sets of consecutive follow-up points and the QALYs between each pair are summed to find the total observed QALYs. This process is then repeated using the predicted utility values for each algorithm, rather than the observed utility values.

It should be noted that this differs from the QALY outcomes used in economic evaluation, which is interested in the incremental QALYs between treatment options. However, the Cancer 2015 dataset does not directly compare treatments, thus this approach is not feasible.

APPENDIX B - DESCRIPTIONS OF THE STATISTICAL MODELS USED TO ESTIMATE THE NEWLY DEVELOPED MAPPING ALGORITHMS

Linear Regression

Linear regression models estimate an additive model of the form

$$Y_i = \beta X_i + \epsilon_i,$$

where Y_i is the EQ-5D-3L utility value for patient i , X_i is a vector of covariates for patient i , β is a vector of regression coefficients and ϵ_i is the error term. In the context of this paper, the vector of covariates is individual i 's scores on the summary scales of the QLQ-C30 questionnaire.

The linear model is fitted for comparison with other methods. EQ-5D-3L utility values are predicted by taking the sum product of the vector of QLQ-C30 summary scores X_i and the vector of regression coefficients β .

$$\hat{Y}_i = \beta X_i$$

Fixed, random and mixed effects

Fixed effects and random effects are variants on linear regression models that control for factors that vary between units but that are constant over time. In the Cancer 2015 dataset, a unit is an individual and the factors this allows one to control for are any time-invariant omitted variables that may determine a patient's utility value, but are not captured by the QLQ-C30.

The fixed effects model assumes the true relationship between an EQ-5D-3L utility value and the QLQ-C30 is of the form:

$$Y_{i,t} = \alpha + \beta X_{i,t} + Z_i + u_{i,t}.$$

Here $Y_{i,t}$ is individual i 's utility value at time t , α is a constant, β is a vector of regression coefficients, $X_{i,t}$ is a vector of QLQ-C30 summary scores for individual i at time t , Z_i a unit specific intercept and $u_{i,t}$ is the error term.

A fixed effects model does not seek to estimate Z_i and instead uses the fact that this is constant over time to avoid its estimation. Instead, only the value of β is estimated and the predicted utility value for a patient is then,

$$\hat{Y}_{i,t} = \beta X_{i,t}.$$

The random effects model works in a similar manner and assumes the true relationship takes the same form as in the fixed effects model above. However, this model also imposes the additional assumption that $X_{i,t}$ and Z_i are not correlated. If this is the case, a random effects model will produce more efficient estimates. To test whether or not the random effects assumption is appropriate, a Hausman test is applied using the Stata command ‘xtoverid’.

Winter³⁷ provides a useful introduction to mixed effects models, which allow for both fixed and random effects. The fixed effects are viewed as an omitted variable, while the random effects term is viewed as the innate difference between individuals. Multiple responses from an individual cannot be considered independent and each subject is assumed to have a random intercept, meaning that the constant term for each patient varies. The Stata ‘mixed’ command is used to fit the mixed effects model. This command does not estimate the random intercept directly and the same formula is applied to calculate predicted utilities as in other linear models:

$$\hat{Y}_{i,t} = \beta X_{i,t}.$$

Logistic models of perfect health

Logistic models are used to predict the outcome of a binary dependent variable given a set of regressors. In this scenario, the outcome variable is 1 if a patient has perfect health and 0 otherwise. Letting this variable be called p , the following model is estimated

$$\hat{p}_i = P(p_i = 1) = \frac{\exp(\beta' X_i)}{1 + \exp(\beta' X_i)}.$$

This model predicts the probability that a patient is in perfect health on the EQ-5D-3L scale, based on their QLQ-C30 responses. A linear regression model is then estimated using all observed utility values less than one. If the probability a patient is in perfect health is greater than or equal to 0.5, they are assigned a utility value of 1. If this is less than 0.5, they are assigned a utility value using the linear regression model:

$$\hat{Y}_i = \begin{cases} 1, & \text{if } \hat{p}_i \geq 0.5 \\ \beta X_i, & \text{if } \hat{p}_i < 0.5 \end{cases}$$

A logistic model is fit in conjunction with an OLS model, a random effects model, a fixed effects model and a mixed effects model.

Response mapping

Response mapping accounts for the fact that utility values are not continuous by avoiding a direct prediction of the values and instead predicting a patient's response to each of the five EQ-5D-3L domains. Depending on the method used, this approach treats the outcome of each question as either an ordinal or categorical variable with three possible outcomes, and then estimates the probability that a patient responds 1, 2 or 3 to that question based on their QLQ-C30 responses.

The advantages of this approach are that it makes no assumptions about the continuity or boundedness of utility values and the results can be applied using the EQ-5D-3L tariff for any country. However, these models are more complicated than other methods, requiring the estimation of five separate models. Additionally, response mapping requires a large sample size to produce reliable algorithms.³⁸

There are two methods for predicting the responses, ordinal logistic regression and multinomial logistic regression. Both are used to develop mapping algorithms in this paper.

Ordinal logistic regression

Ordinal logistic models are used to predict the value of an ordinal dependent variable, such as EQ-5D-3L responses.³⁵ In an ordinal logistic regression with N possible outcomes, the set of independent variables are used to produce a linear predictor and a set of cut points. The probability of observing outcome j corresponds to the probability that the linear predictor plus an error term is between cut point $j-1$ and cut point j .

$$\Pr(\text{outcome}_i = j) = \Pr(\varphi_{j-1} < \beta X_i + u_i < \varphi_j),$$

where $\varphi_0, \varphi_1, \dots, \varphi_N$ are the cut points, βX_i is the linear predictor and u_i is a logistically distributed error term. Here $\varphi_0 = -\infty$ and $\varphi_N = \infty$.

Therefore, the probability that $\text{outcome}_i \leq j$ is equal to the probability that $\beta X_i + u_i < \varphi_j$, or equivalently that $u_i < \varphi_j - \beta X_i$.

The cumulative distribution function for a logistically distributed random variable takes the form

$$F(x) = 1/(1 + \exp(-x))$$

and so the probability of the variable taking a value ranked j or below is then

$$P_{\leq i} = 1/[1 + \exp(-(\varphi_i - \beta X_j))].$$

Thus $P_j = P_{\leq j} - P_{\leq j-1}$, where P_j is the probability of outcome j .

In the case of the EQ-5D-3L there are three possible outcomes, and so two cut points are required.

This model relies on the data meeting the proportional odds assumptions. This assumes that the coefficients that describe the probability of moving between each category and the categories above it are constant. In other words, this means that the explanatory variables have the same effect on the odds of being in category one rather than category two or three as they do on the odds of being in category one or two, rather than category three. It is this assumption that allows for a single set of coefficients and if it is violated, which is common, it may be preferred to consider a multinomial logistic model. The proportional odds assumption is tested in Stata using the commands ‘omodel’ and ‘brant’.

Multinomial logistic regression

Multinomial logistic regression ignores that the data are ordinal, instead treating the data as categorical. When using Stata³⁵ to estimate such a model, a base category must be specified. The model then estimates the risk of being in each of the other categories, relative to the base category. A multinomial logistic regression with three categories will estimate two sets of parameters. If response one is chosen as the base category, the model will produce vectors of parameters β_2 and β_3 , corresponding to outcomes two and three respectively on a specific domain of the EQ-5D-3L. The probability of each response is then

$$P(\text{response} = j) = \frac{\exp(\beta_j X_i)}{\sum_{k=1}^3 \exp(\beta_k X_i)},$$

where β_1 is a vector of zeros, so that $\exp(\beta_1 X_i) = 1$. The choice of base categories is not important and is only required so that the parameters for this category can be set to zero, or else there would be an infinite number of solutions to the problem.

Calculating utility values

Once these models have been estimated there will be 15 probabilities, three per EQ-5D-3L dimension, which will need to be transformed into utility values. The method applied here follows the approach previously implemented by the Longworth¹⁶ response mapping algorithm, which uses the probabilities to produce the expected EQ-5D-3L utility values. By

this method a patient's utility is calculated by multiplying each coefficient in the tariff by the probability of that outcome. Using the UK tariff³²:

$$\begin{aligned}\hat{Y}_i = & 1 - 0.069 * P_{1,2} - 0.314 * P_{1,3} - 0.104 * P_{2,2} - 0.214 * P_{2,3} - 0.036 * P_{3,2} \\ & - 0.094 * P_{3,3} - 0.123 * P_{4,2} - 0.386 * P_{4,3} - 0.071 * P_{5,2} - 0.236 * P_{5,3} \\ & - 0.081 * (1 - P_{perfect}) - 0.269 * N_3.\end{aligned}$$

Here $P_{i,j}$ represents the probability of having outcome j on category i , where categories 1 through 5 are mobility self-care, usual activities, pain and discomfort and anxiety and depression respectively. $P_{perfect}$ is the probability of being in perfect health and N_3 is the probability of having at least one category 3 response:

$$N_3 = 1 - (1 - P_{1,3}) * (1 - P_{2,3}) * (1 - P_{3,3}) * (1 - P_{4,3}) * (1 - P_{5,3}).$$

Mixture models

In this paper mixture models are estimated using the 'aldvmm' command in Stata.³⁶ This command fits an adjusted limited dependent variable mixture model (ALDVMM), which takes into account not just the multimodality of the EQ-5D-3L distribution, but also that the EQ-5D-3L is not on a continuous unbounded scale. It assumes that the EQ-5D-3L distribution is made up of C classes, where C is a user-defined number, and that conditional on being in class c , the EQ-5D-3L utility value is

$$y_i|c = \begin{cases} 1, & \text{if } y_i^*|c > \Psi_1 \\ \Psi_2, & \text{if } y_i^*|c \leq \Psi_2 \\ y_i^*|c, & \text{otherwise} \end{cases}$$

where Ψ_1 is the largest possible EQ-5D-3L score below 1, Ψ_2 is the minimum possible EQ-5D-3L score, and

$$y_i^*|c = x_i\beta_c + \epsilon_{ic}.$$

Here x_i is a vector of covariates, β_c is a vector of regression coefficients and ϵ_{ic} is the error term. Once each of these models is estimated, taking the conditional expectation from each class, weighting the score by the probability of class membership, the following equation provides the expected EQ-5D-3L score.

$$\begin{aligned}
E(y_i|x'_i w'_i) = & \sum_{c=1}^c \frac{\exp(w'_i \delta_c)}{\sum_{s=1}^c \exp(w'_i \delta_s)} \left\{ \left[1 - \Phi \left(\frac{\Psi_1 - x'_i \beta_c}{\sigma_c} \right) \right] + \left[\Phi \left(\frac{\Psi_2 - x'_i \beta_c}{\sigma_c} \right) \right] \Psi_2 \right. \\
& + \left[\Phi \left(\frac{\Psi_1 - x'_i \beta_c}{\sigma_c} \right) - \Phi \left(\frac{\Psi_2 - x'_i \beta_c}{\sigma_c} \right) \right] \left[x'_i \beta_c \right. \\
& \left. \left. + \sigma_c \frac{\phi \left(\frac{\Psi_1 - x'_i \beta_c}{\sigma_c} \right) - \phi \left(\frac{\Psi_2 - x'_i \beta_c}{\sigma_c} \right)}{\Phi \left(\frac{\Psi_2 - x'_i \beta_c}{\sigma_c} \right) - \Phi \left(\frac{\Psi_1 - x'_i \beta_c}{\sigma_c} \right)} \right] \right\}
\end{aligned}$$

Here x_i represents the covariates for the model fitted to each component class and w_i represents the covariates for the multinomial logit model predicting class membership. The β_c are the vectors of regression coefficients within each class and the δ_c are the coefficients for predicting class membership. The standard deviation of the error term in each class is σ_c . Φ and ϕ represent the standard cumulative normal function and standard normal density function respectively.

These models are more complicated than those previously described and can be difficult to apply in practice. Consideration must be given to the number of classes to fit and to the variables used to predict class membership. More classes means the model will fit the data more closely, but this comes with a trade-off in terms of generalisability. Hernandez Alava et al.³⁹ advise that the Bayesian information criteria (BIC) may be a useful method of assessing whether an additional class improves model fit, as it penalises additional components.

The models use the 15 QLQ-C30 summary scores to predict EQ-5D-3L within each component. Including these as predictors of class membership will increase the complexity of the model, without necessarily improving the model fit. A model with constant probabilities can be estimated which will simplify the model, but this may come at the cost of reduced accuracy.

Variables not captured in the QLQ-C30 could also be used as predictors of class membership, for example if it were hypothesised that the latent classes could be defined by disease stage, this could be used as a predictor. However, including such external variables reduces the generalisability of the algorithm.

A key practical consideration when weighing both of these issues is that the more complex the model grows, the less likely it is that it will converge. For this reason, a simple model with two

latent classes and constant probabilities is estimated first and the complexity is built from there. The combined health and quality of life scale is added as a predictor of class membership and if these models converge the process is repeated with three latent classes, then four, and so on until either it is judged that the addition of further classes does not improve the model or until models stop converging.

Beta models

The two beta models are estimated using the generalised linear model framework. Utility values are first transformed onto 0-1 scale using the function $Y'_i = \frac{Y_i - -0.594}{1 - -0.594}$. Then a model is fit using a logit link function $\log(\mu/(1 - \mu))$ and a binomial-like variance $\phi\mu(1 - \mu)$. Thus to estimate utility values the linear predictor is transformed using the inverse of the logit link function.

$$\hat{Y}'_i = \frac{\exp(\hat{W}_i)}{1 + \exp(\hat{W}_i)}$$

$$\hat{W}_i = \beta X_i.$$

This then needs to be transformed back to the -0.594-1 scale using the transformation $\hat{Y}_i = \hat{Y}'_i(1 - -0.594) - 0.594$.

APPENDIX C - SCATTER PLOTS FOR THE EXTERNAL VALIDATION OF EXISTING MAPPING ALGORITHMS

Figure S1. Scatter plot of observed vs. predicted utility values for the Longworth response mapping algorithm

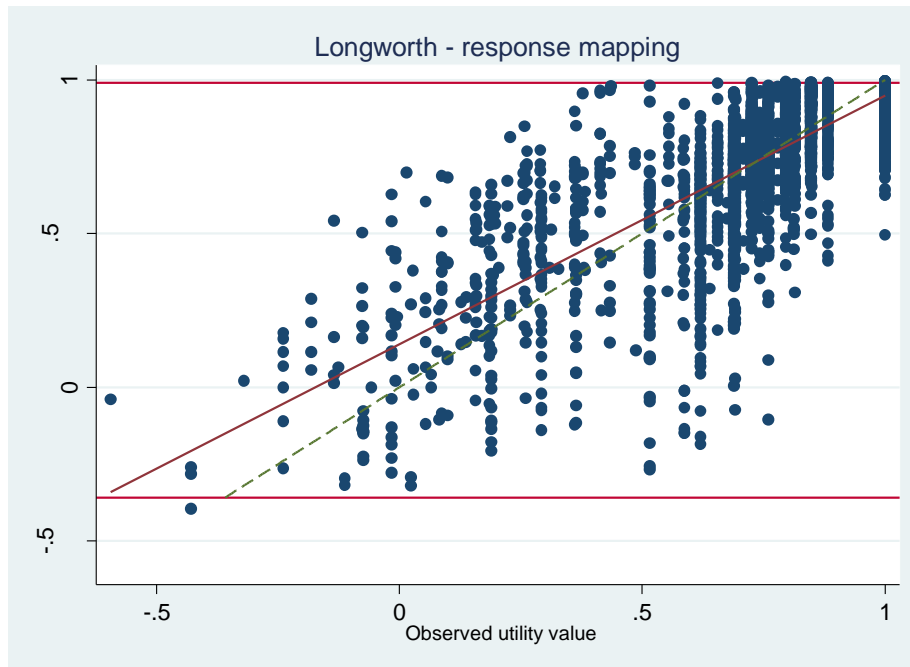


Figure S2. Scatter plot of observed vs. predicted utility values for the Veersteegh OLS mapping algorithm

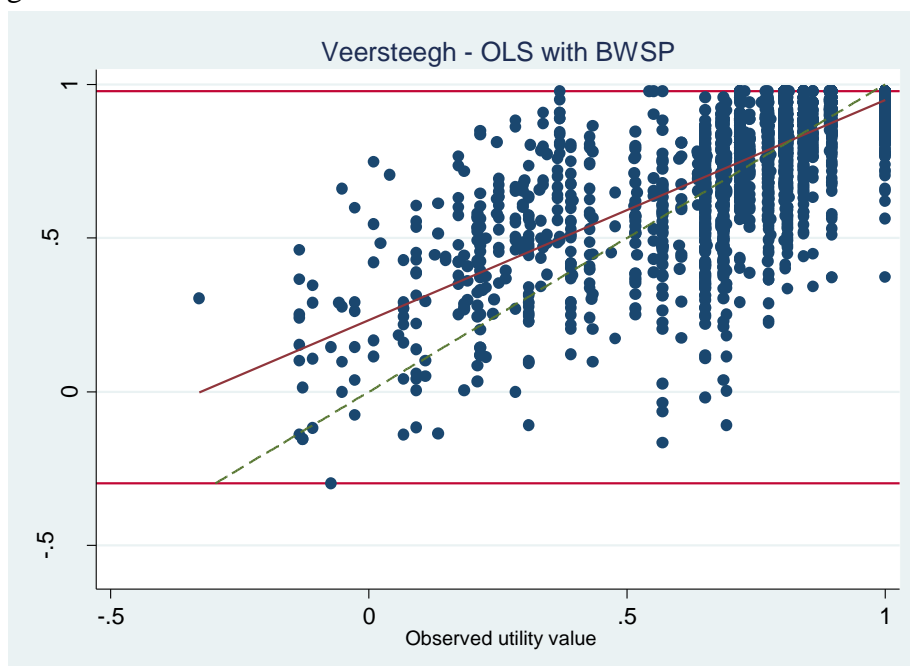


Figure S3. Scatter plot of observed vs. predicted utility values for the Khan random effects mapping algorithm

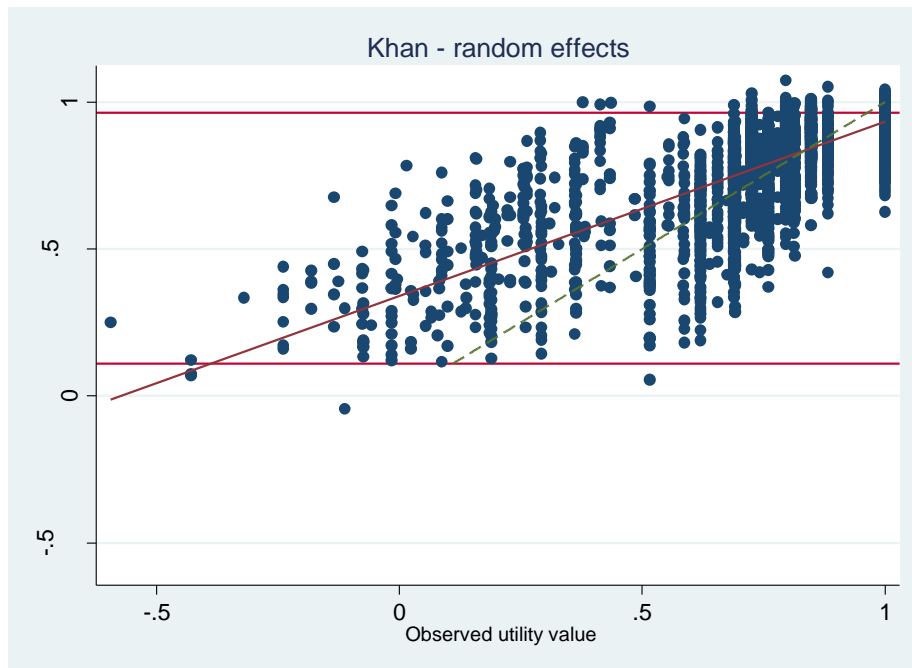


Figure S4. Scatter plot of observed vs. predicted utility values for the Khan beta-binomial mapping algorithm

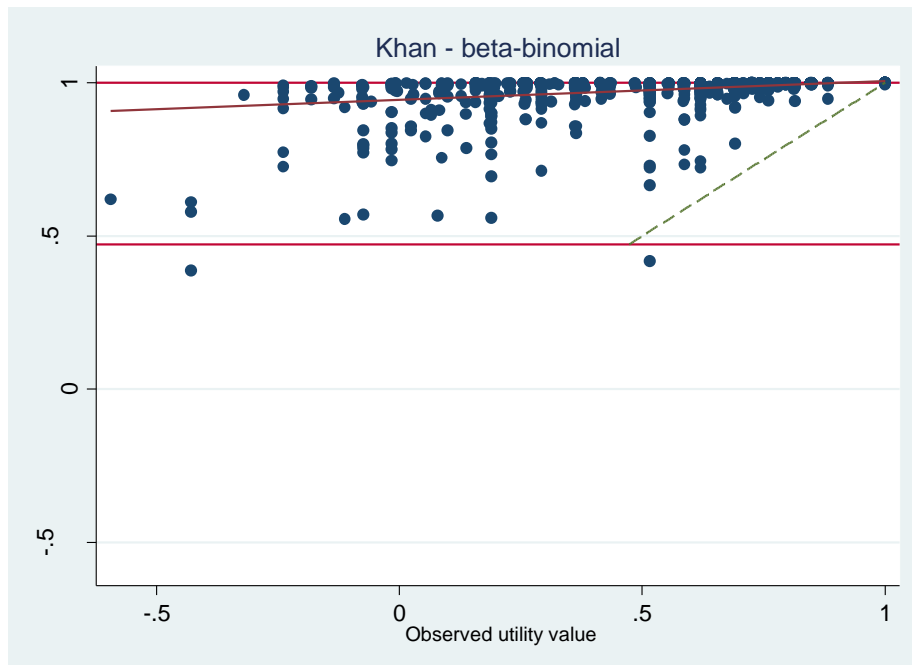


Figure S5. Scatter plot of observed vs. predicted utility values for the Khan LDVM mapping algorithm

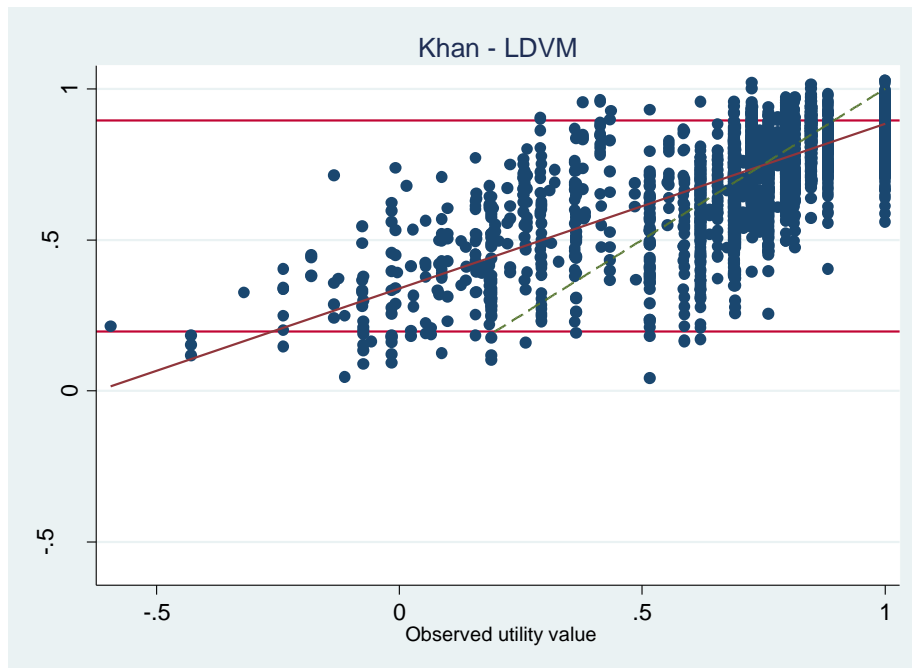


Figure S6. Scatter plot of observed vs. predicted utility values for the Marriott mixed effects mapping algorithm

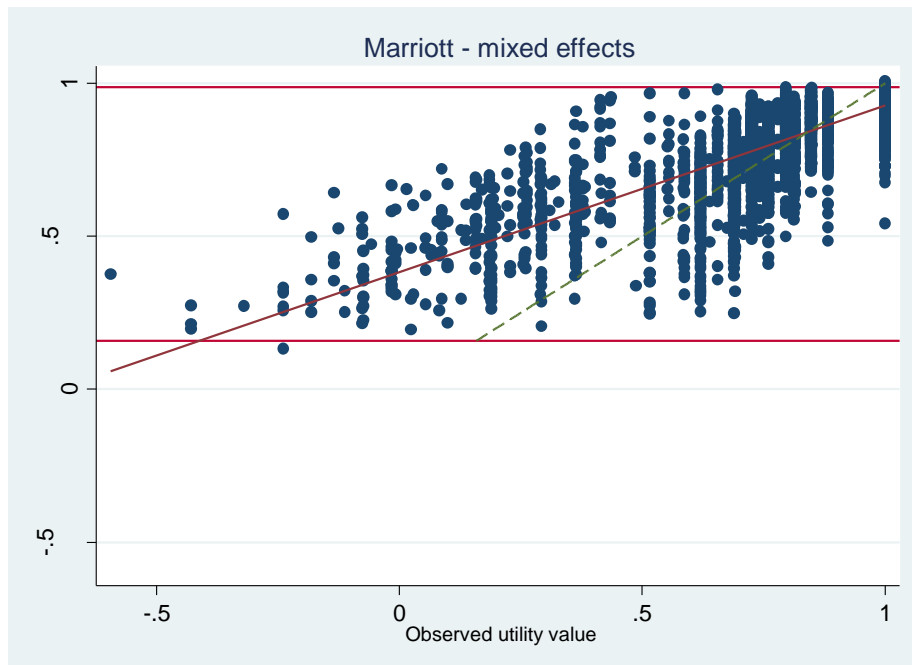
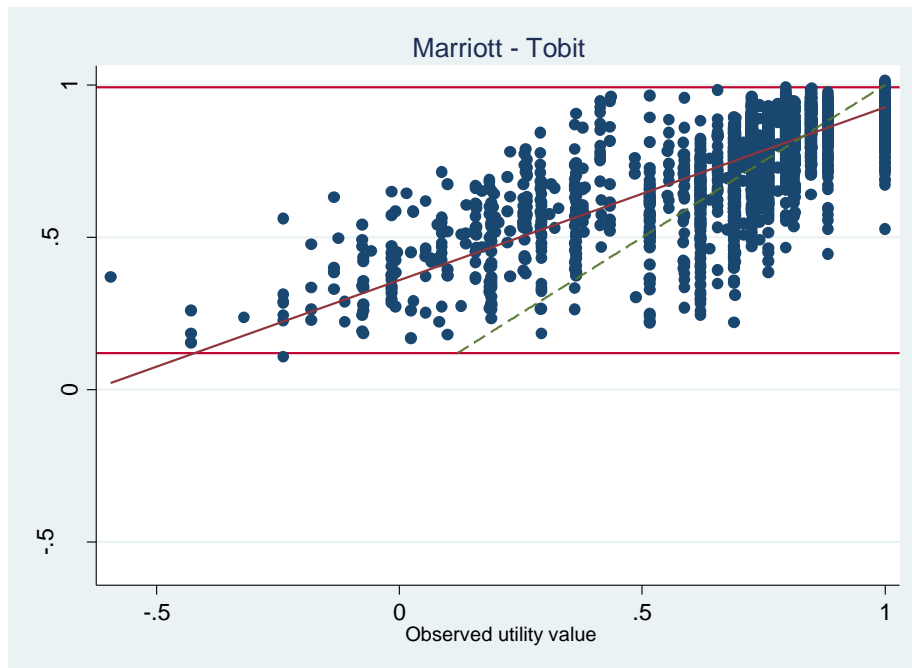


Figure S7. Scatter plot of observed vs. predicted utility values for the Marriott Tobit mapping algorithm



APPENDIX D - ADDITIONAL EXTERNAL VALIDATION RESULTS FOR EXISTING MAPPING ALGORITHMS

Table S1. Mean absolute error for each algorithm for each of the ten tumour sites from external validation dataset one

Tumour Site	Longworth¹⁶	Veersteegh¹⁵	Khan Random Effects²⁴	Khan Beta-Binomial²⁴	Khan LDVMM²⁴	Marriott Mixed Effects²³	Marriott Tobit²³
Bone and Soft Tissue	0.112	0.130	0.121	0.305	0.128	0.123	0.120
Breast	0.096	0.097	0.103	0.196	0.112	0.095	0.094
Cancer of Unknown Primary	0.117	0.137	0.112	0.268	0.113	0.116	0.114
Cervical	0.085	0.108	0.098	0.182	0.107	0.087	0.086
Colorectal	0.103	0.105	0.110	0.204	0.123	0.104	0.103
Head and Neck	0.098	0.108	0.113	0.243	0.113	0.104	0.103
Lung	0.135	0.135	0.135	0.297	0.138	0.130	0.128
Oesophagogastric	0.106	0.112	0.120	0.263	0.127	0.116	0.115
Other	0.101	0.108	0.113	0.228	0.123	0.109	0.108
Prostate	0.065	0.075	0.080	0.121	0.103	0.074	0.072
Renal	0.096	0.100	0.112	0.238	0.120	0.106	0.103

Abbreviations: LDVMM, Limited dependent variable mixture model

Table S2. Rankings by mean absolute error for each algorithm using tumour site specific sub-groups from validation dataset one and the mean observed EQ-5D-3L utility value for each tumour site

Tumour Site	Longworth¹⁶	Veersteegh¹⁵	Khan Random Effects²⁴	Khan Beta-Binomial²⁴	Khan LDVMM²⁴	Marriott Mixed Effects²³	Marriott Tobit²³	Mean Observed EQ-5D-3L
Prostate	1	1	1	1	1	1	1	0.88
Cervical	2	7	2	2	2	2	2	0.81
Breast	4	2	3	3	3	3	3	0.80
Renal	3	3	5	6	6	6	5	0.76
Colorectal	7	4	4	4	7	4	6	0.79
Head and Neck	5	6	8	7	4	5	4	0.75
Other	6	5	7	5	8	7	7	0.77
Cancer of Unknown Primary	10	11	6	9	5	9	8	0.72
Oesophagogastric	8	8	9	8	9	8	9	0.72
Bone and Soft Tissue	9	9	10	11	10	10	10	0.68
Lung	11	10	11	10	11	11	11	0.68

Abbreviations: LDVMM, Limited dependent variable mixture model

Table S3. Observed and predicted QALYs for each algorithm using validation dataset one

	Observations	Mean QALYs	Std. Dev.	Minimum	Maximum	Error	p-value for error = 0
Observed QALYs (UK tariff)	924	0.933	0.701	-0.115	3.012	-	-
Observed QALYs (NL tariff)	924	0.952	0.703	-0.065	3.012	-	-
Longworth¹⁶	924	0.922	0.701	-0.054	2.976	-0.010	0.014
Veersteegh¹⁵	924	0.959	0.701	0.061	2.943	0.007	0.004
Khan Random Effects²⁴	924	0.950	0.698	0.073	2.985	0.018	0.000
Khan Beta-Binomial²⁴	924	1.151	0.780	0.189	3.189	0.219	0.000
Khan LDVMM²⁴	924	0.903	0.658	0.080	2.794	-0.029	0.000
Marriott Mixed Effects²³	924	0.955	0.692	0.111	2.958	0.022	0.000
Marriott Tobit²³	924	0.950	0.692	0.104	2.965	0.018	0.000

Abbreviations: QALY, Quality-adjusted life-year; LDVMM, Limited dependent variable mixture model.

APPENDIX E - REPLICATION OF RESULTS PRESENTED BY KHAN ET AL. FOR THE BETA BINOMIAL MODEL

We had difficulty replicating the results presented by Khan et al. for the beta-binomial model. In the paper by Khan et al., predicted utilities for the beta binomial model developed using the EQ-5D-3L are presented in Table 5 '*Predicted utilities from 3 scenarios*'. When the best QLQ-C30 scores for all the function (scores of 100) and symptom (scores of 0) scales are imputed into the beta-binomial mapping algorithm for the EQ-5D-3L, Khan et al. report a predicted utility value of 0.901. When the worst QLQ-C30 scores for all function (scores of 0) and symptom (scores of 100) scales are imputed into the beta-binomial mapping algorithm for the EQ-5D-3L, Khan et al. report a predicted utility value of 0.097. We attempted to replicate these results given the poor performance of the Khan et al. beta-binomial model in our external validation. Using the coefficients for the beta-binomial for the EQ-5D-3L reported by Khan et al. in Table 6 '*Results from Statistical Modelling (BB Model)*' and the best and worst QLQ-C30 scores described above (coefficients and best/worst QLQ-C30 scores are provided in Table S4 below for reference) we tried to estimate the respective predicted EQ-5D-3L utility values by taking the logistic transformation of the sum product $[1/(1+\exp(-\beta X))]$. This resulted in a predicted EQ-5D-3L utility value of 0.9999 when imputing the best QLQ-C30 scores and a utility value of 0.4735 when imputing the worst QLQ-C30 scores. As you can see our replicated results are not consistent with those reported by Khan et al. in Table 5. It is, therefore, believed that the beta-binomial model has been miss-specified in the paper by Khan et al. or insufficient detail has been provided to estimate the model correctly. Thus, affecting the predictive accuracy of the algorithm in our external validation.

Table S4. Data used to replicate the predicted utilities reported by Khan et al.

QLQ-C30 Functioning/ Symptom Scales	Coefficients for beta-binomial model reported by Khan et al. in Table 6 for the EQ-5D-3L	Best QLQ-C30 scores	Worst QLQ-C30 scores
Intercept	-0.0123	1	1
Physical	0.08711	100	0
Role	0.00421	100	0
Emotional	0.00661	100	0
Cognitive	-0.00425	100	0
Social	-0.00035	100	0
Global	-0.00197	100	0
Fatigue	0.00443	0	100
Nausea	-0.00146	0	100
Pain	-0.01039	0	100
Dyspnoea	0.00015	0	100
Insomnia	0.00193	0	100
Appetite	0.0002	0	100
Constipation	0.0014	0	100
Diarrhoea	0.00393	0	100
Financial	-0.00113	0	100

APPENDIX F- MODEL COEFFICIENTS FOR THE NEWLY DEVELOPED MAPPING ALGORITHMS

Table S5. Linear model coefficients

	OLS	Logit	OLS + Logit	Fixed effects	FE + Logit	Random effects	RE + Logit	Mixed effects	ME + Logit
OQL	0.0011***	0.0181***	0.001***	0.0009***	0.0011***	0.001***	0.0011***	0.001***	0.0011***
PF	0.0036***	0.0559***	0.0033***	0.0033***	0.0036***	0.0033***	0.0036***	0.0033***	0.0036***
RF	0.0005***	0.0208***	0.0003	0.0005*	0.0004**	0.0002	0.0005***	0.0003	0.0005***
EF	0.0025***	0.056***	0.0022***	0.0023***	0.0025***	0.0022***	0.0025***	0.0022***	0.0025***
CF	-0.0002	0.0002	-0.0002	-0.0003	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
SF	0.0003*	0.0073	0.0003	0.0003	0.0003*	0.0003	0.0003*	0.0003	0.0003*
FA	0.0004*	-0.0053	0.0007**	0.0001	0.0003	0.0005*	0.0003	0.0006*	0.0004*
NV	0.0005*	0.0089	0.0003	0.0005	0.0004*	0.0003	0.0005*	0.0003	0.0005*
PA	-0.0029***	-0.069***	-0.0025***	-0.0026***	-0.0028***	-0.0024***	-0.0028***	-0.0024***	-0.0029***
DY	0	-0.0022	0	-0.0001	0	0.0001	0	0.0001	0
SL	-0.0001	-0.0023	-0.0001	0.0001	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
AP	-0.0002	-0.0037	-0.0002	0.0001	-0.0001	-0.0002	-0.0002	-0.0002	-0.0002
CO	-0.0001	0.0008	-0.0002	0	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
DI	0	0.0005	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0
FI	-0.0003**	-0.0067*	-0.0002*	-0.0002	-0.0003**	-0.0002*	-0.0003**	-0.0002*	-0.0003**
Constant	0.2151***	-13.527***	0.2304***	0.2622***	0.216***	0.2315***	0.2145***	0.2304***	0.2151***

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S6. Ordinal logistic model coefficients

	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
OQL	-0.0192***	-0.0043	-0.0182***	-0.0117***	-0.0157***
PF	-0.0859***	-0.0636***	-0.0356***	-0.0307***	-0.0132***
RF	0.0022	-0.0128**	-0.0415***	0.013***	0.0023
EF	0.0071*	0.006	0.0022	-0.0082**	-0.1024***
CF	-0.0073*	-0.0067	0.001	0.0113***	0.0008
SF	0.0011	-0.0019	-0.0128***	0.0035	-0.001
FA	-0.006	-0.0183**	0.0075*	0.0046	-0.0038
NV	-0.0051	-0.0007	-0.003	-0.0044	-0.011**
PA	0.0126***	0.0065	-0.0001	0.1023***	-0.0051*
DY	0.0034	-0.0011	0.0017	0	0
SL	-0.0034	0.0082**	-0.0007	0.0038	0.0033
AP	-0.0042	0.0057	-0.0005	-0.0008	-0.0002
CO	0.0056*	0.0027	0.0025	-0.0005	-0.0046*
DI	-0.0005	0.0025	0.0003	0.0013	-0.0012
FI	-0.0023	0.0038	0.0045*	0.0053**	0.0048**
Cut point 1	-6.5319	-2.7766	-7.507	0.0793	-9.7692
Cut point 2	0.8218	1.3401	-2.1368	8.6377	-3.641

Abbreviations: OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S7. Multinomial logistic model coefficients – response 2 versus response 1

Response 2 versus 1					
	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
OQL	-0.0188***	-0.006	-0.0147***	-0.0105**	-0.0121**
PF	-0.0853***	-0.0603***	-0.0481***	-0.0333***	-0.0172***
RF	0.0021	-0.0136**	-0.0376***	0.0103**	0.0031
EF	0.0063	0.0058	0.0011	-0.0082*	-0.1096***
CF	-0.0066	-0.0065	-0.0015	0.0065	0.0011
SF	0.0012	-0.0001	-0.0163***	0.0025	-0.0023
FA	-0.0061	-0.0175**	0.0123**	0.0034	-0.0036
NV	-0.0054	-0.0001	-0.0023	-0.004	-0.013**
PA	0.0128***	0.006	0.0042	0.1101***	-0.0041
DY	0.0036	-0.0013	0.0044	-0.0002	0.0014
SL	-0.0032	0.0089**	-0.001	0.0037	0.0036
AP	-0.004	0.0064	0.0033	-0.002	0.0002
CO	0.0056*	0.0036	-0.0023	0.0002	-0.0057*
DI	-0.0015	0	-0.0037	0.0022	-0.0009
FI	-0.0027	0.0038	0.005*	0.0048*	0.0057**
Constant	6.4467***	2.5157**	8.348***	0.7074	10.3745***

Abbreviations: OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S8. Multinomial logistic model coefficients – response 3 versus response 1

Response 3 versus 1					
	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety and depression
OQL	-0.0541	0.0298	-0.0348***	-0.0213*	-0.0378***
PF	-0.1676**	-0.1241***	-0.0699***	-0.0627***	-0.0226*
RF	-0.824	0.006	-0.0758***	0.0311***	0.0046
EF	0.0554	0.0135	0.0052	-0.0178**	-0.1785***
CF	-0.0561	-0.011	0.002	0.0264***	-0.0051
SF	-0.0199	-0.0375	-0.0261***	0.0085	0.0031
FA	0.0311	-0.0453	0.0013	0.0006	-0.0132
NV	-0.0173	0.0034	-0.0022	-0.0019	-0.0183*
PA	0.0074	0.022	-0.0006	0.1902***	-0.0069
DY	0.0134	0.0068	0.0026	0.0006	-0.0032
SL	-0.0055	0.005	-0.0018	0.0076	0.0049
AP	-0.0306	-0.002	0	0.0018	0.0028
CO	0.0042	-0.0124	0.0071	-0.001	-0.008
DI	0.0455	0.0231	0.0007	-0.0013	-0.0042
FI	0.0188	-0.0043	0.0083*	0.0123**	0.0077
Constant	2.4946	1.7401	10.7855***	-6.9707***	13.3566***

Abbreviations: OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S9. ALDVMM-2C-CP coefficients

	Class 1	Class 2
OQL	0.0015***	0.0007***
PF	0.0042***	0.002***
RF	0.0014***	-0.0001
EF	0.0037***	0.0011***
CF	0.0006	-0.0003**
SF	0.0005	-0.0001
FA	0.0006	-0.0003
NV	0.0009	0.0004*
PA	-0.004***	-0.0016***
DY	0	0
SL	-0.0005	0
AP	-0.0004	0.0002
CO	0.0004	-0.0001
DI	-0.0003	0
FI	-0.0004	-0.0002*
Constant	-0.0729	0.5736***
sigma	0.1719	0.0594
Multinomial logistic model for class membership		
Constant	-0.1145	
Probability of class membership	0.4714	0.5286

Abbreviations: OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S10. ALDVMM-2C-NCP coefficients

	Class 1	Class 2
OQL	-0.0012	0.001***
PF	0.0053***	0.0019***
RF	0.0013**	0.0001
EF	0.0044***	0.0012***
CF	0.0004	-0.0003*
SF	0.0003	0
FA	0.0011	-0.0003
NV	0.0005	0.0006***
PA	-0.0035***	-0.0019***
DY	0.0001	0
SL	-0.0004	-0.0001
AP	-0.0005	0.0001
CO	0.0004	-0.0002
DI	-0.0002	-0.0001
FI	-0.0005	-0.0002*
Constant	-0.0808	0.5339***
sigma	0.2005	0.0694
Multinomial logistic model for class membership		
OQL	-0.0492***	
Constant	1.9336***	

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S11. ALDVMM-3C-CP coefficients

	Class 1	Class 2	Class 3
OQL	-0.0003***	0.0018***	0.0007***
PF	0.0006***	0.0052***	0.002***
RF	0.0007***	0.0015***	-0.0001
EF	0.0018***	0.0041***	0.0011***
CF	-0.0001	0.0007	-0.0003**
SF	0.0006***	0.0003	-0.0001
FA	0.0003**	0.001	-0.0003
NV	0.0005***	0.001*	0.0004**
PA	-0.007***	-0.0032***	-0.0017***
DY	-0.0001*	0.0001	0
SL	0	-0.0005*	0
AP	-0.0009***	-0.0003	0.0002
CO	0.0008***	0.0003	-0.0001
DI	-0.0002*	-0.0003	0
FI	-0.0002***	-0.0005*	-0.0002*
Constant	0.5935***	-0.2209**	0.574***
sigma	0.0085	0.1803	0.0596
Multinomial logistic model for class membership			
Constant	-1.9331***	-0.2231*	
Probability of class membership	0.0744	0.4114	0.5142

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S12. ALDVMM-3C-NCP coefficients

	Class 1	Class 2	Class 3
OQL	-0.0006	-0.0012	0.0004*
PF	0.0031***	0.005***	0.0018***
RF	0.0006	0.0017***	-0.0001
EF	0.0028***	0.0044***	0.0008***
CF	-0.0005	0.0005	-0.0001
SF	0.0007*	0.0002	0
FA	-0.0002	0.0011	0
NV	0.0015	0.0002	0.0002
PA	-0.0053***	-0.0032***	-0.0014***
DY	-0.0008**	0.0004	0
SL	-0.0004	-0.0004	0
AP	-0.0006	-0.0005	0.0002
CO	-0.0001	0.0005	-0.0002
DI	-0.0001	-0.0001	0
FI	-0.0002	-0.0006	-0.0001
Constant	0.4178***	-0.0839	0.5823***
Sigma	0.0836	0.2115	0.0554
Multinomial logistic model for class membership			
OQL	0.1173***	-0.0447***	
Constant	-9.1346***	1.7776***	

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S13. ALDVMM-4C-CP coefficients

	Class 1	Class 2	Class 3	Class 4
OQL	0.0016***	0.0007***	-0.0008***	- 0.0008***
PF	0.0042***	0.002***	0.0005***	0.0005***
RF	0.002***	-0.0002	0.0002**	0.0002**
EF	0.0037***	0.001***	0.0004***	0.0004***
CF	0.0008*	-0.0002	-0.0006***	- 0.0006***
SF	0.0004	-0.0002	0.0001**	0.0001**
FA	0.001*	-0.0001	-0.0021***	- 0.0021***
NV	0.0009	0.0002	0.001***	0.001***
PA	-0.0039***	-0.0015***	-0.0064***	- 0.0064***
DY	-0.0001	-0.0001	-0.0002***	- 0.0002***
SL	-0.0008***	-0.0001	-0.0001*	-0.0001*
AP	-0.0003	0.0002*	-0.0003***	- 0.0003***
CO	0.0001	-0.0001	-0.0001***	- 0.0001***
DI	-0.0005	0.0002	-0.0002***	- 0.0002***
FI	-0.0004	-0.0002**	-0.0001**	-0.0001**
Constant	-0.1372*	0.5789***	0.9307***	0.9307***
Sigma	0.0586	0.1782	0.0586	0.005
Multinomial logistic model for class membership				
Constant	0.906***	0.891***	-1.2798***	
Probability of class membership	0.3998	0.3938	0.0449	0.1615

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S14. ALDVMM-4C-NCP coefficients

	Class 1	Class 2	Class 3	Class 4
OQL	-0.0002	0.0003	-0.0015**	0.0125
PF	0.0027***	0.0018***	0.0041***	0.0093
RF	0.0005*	-0.0002*	0.0019***	0.0124
EF	0.0022***	0.0008***	0.0039***	0.0111
CF	-0.0001	-0.0001	0.0003	0.0077
SF	0.0004*	0	0.0005	0.0103
FA	-0.0004	0.0002	0.0008	0.0013
NV	0.0005	0.0001	0.0003	0.0053
PA	-0.0036***	-0.0013***	-0.0038***	0
DY	0.0001	-0.0001	0.0002	0.0059
SL	-0.0002	0	-0.0007*	0.0024
AP	-0.0003	0.0001	-0.0003	0.0019
CO	-0.0002	-0.0001	0.0004	0.0043
DI	0.0001	0.0001	-0.0003	0.0043
FI	-0.0002	-0.0001	-0.0004	-0.0014
Constant	0.431***	0.5895***	0.0316	1.464
sigma	0.0636	0.0558	0.2049	0
Multinomial logistic model for class membership				
OQL	-0.1765*	-0.2788***	-0.3105***	
Constant	18.8203**	26.2164***	27.5053***	

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

Table S15. Beta models

	One-part model	Two-part model
OQL	0.0085*	0.0044
PF	0.0144***	0.0114**
RF	0.0027	0.0012
EF	0.0118***	0.0075**
CF	-0.0018	-0.0009
SF	0	0.0008
FA	-0.0006	0.0021
NV	0.004	0.0014
PA	-0.0113***	-0.0086***
DY	0.0006	0.0003
SL	-0.0011	-0.0005
AP	0	-0.0006
CO	-0.0005	-0.0005
DI	-0.0006	0.0001
FI	-0.0013	-0.0009
Constant	-0.4088	-0.1291

Abbreviations: OLS, Ordinary least squares; RE, Random effects; ME, Mixed effects; OQL, Overall health and quality of life; PF, Physical functioning; RF, Role functioning; EF, Emotional functioning; CF, Cognitive functioning; SF, Social functioning; FA, Fatigue; NV, Nausea and vomiting; PA, Pain; DY, Dyspnoea; SL, Sleep disturbance; AP, Appetite loss; CO, Constipation; DI, Diarrhoea; FI, Financial problems.

**APPENDIX G - ADDITIONAL RESULTS FROM THE VALIDATION OF THE
NEWLY DEVELOPED MAPPING ALGORITHMS USING VALIDATION
DATASET TWO A AND B**

Table S16. Validation dataset two A - Observed and predicted QALYs

Algorithm	Observations	Mean QALYs	Standard deviation	Min	Max	Error
Observed QALYs – UK tariff	71	0.7579559	0.6969343	-0.1082779	2.811116	
Observed QALYs – Dutch tariff	71	0.7810267	0.6997136	-0.0294853	2.834605	
Longworth	71	0.7493322	0.6981716	-0.0301589	2.932139	0.009
Veersteegh	71	0.8021307	0.6998625	0.0255209	2.860327	-0.021
Khan RE	71	0.7976079	0.7123811	0.0731958	2.984646	-0.040
Khan BB	71	1.013499	0.7837895	0.2335624	3.013245	-0.256
Khan MM	71	0.7611456	0.6725971	0.0796952	2.786683	-0.003
Marriott FE	71	0.8043181	0.7017308	0.1109985	2.913831	-0.046
Marriott Tobit	71	0.7978702	0.7011688	0.1040306	2.918756	-0.040
OLS	71	0.7681308	0.6994119	0.0654517	2.933806	-0.010
OLS + logit	71	0.7615119	0.705723	0.0750667	3.000684	-0.004
FE	71	0.7716973	0.6903721	0.0761751	2.865282	-0.014
FE + logit	71	0.7646071	0.7039432	0.0866044	3.000684	-0.007
RE	71	0.7668288	0.6971946	0.0664427	2.921357	-0.009
RE + logit	71	0.7602837	0.7050455	0.0757224	3.000684	-0.002
ME	71	0.7675458	0.6986005	0.0658831	2.929733	-0.010
ME + logit	71	0.7592869	0.7025497	0.0754205	2.950979	-0.001
Ordinal	71	0.7649288	0.6958415	0.0037734	2.84618	-0.007
Multinomial	71	0.762997	0.6935873	0.0310516	2.861092	-0.005
ALDVMM-2C-CP	71	0.7753663	0.6932946	0.075512	2.83973	-0.017
ALDVMM-2C-NCP	71	0.7734976	0.6976967	0.0162497	2.845034	-0.016
ALDVMM-3C-CP	71	0.774375	0.6928025	0.0751294	2.840843	-0.016
ALDVMM-3-NCP	71	0.7685078	0.6898124	0.0350094	2.920963	-0.011
ALDVMM-4C-CP	71	0.7692601	0.6915992	0.0732476	2.844716	-0.011
ALDVMM-4-NCP	71	0.7672897	0.6882122	0.0511385	2.917761	-0.009
One-part beta model	71	0.7725621	0.7008396	-0.0056055	2.782479	-0.015
Two-part beta model	71	0.7649612	0.7099776	0.0598475	3.000684	-0.007

Table S17. Validation dataset two B - Observed and predicted QALYs

Algorithm	Observations	Mean QALYs	Standard deviation	Min	Max	Error
Observed QALYs – UK tariff	91	1.001048	0.7189816	0.1889117	3.011636	
Observed QALYs – Dutch tariff	91	1.013484	0.7262484	0.1889117	3.011636	
Longworth	91	0.9805506	0.6995562	0.1678483	2.908004	0.020
Veersteegh	91	1.007986	0.7123305	0.177577	2.906365	0.005
Khan RE	91	0.997469	0.7066677	0.1746219	2.919594	0.004
Khan BB	91	1.137801	0.7857034	0.1888817	3.063223	-0.137
Khan MM	91	0.9443612	0.6655379	0.1625867	2.735534	0.057
Marriott FE	91	0.9942787	0.7017283	0.1760405	2.907748	0.007
Marriott Tobit	91	0.9919526	0.7004281	0.1759714	2.912088	0.009
OLS	91	0.9789373	0.6979112	0.1762814	2.955403	0.022
OLS + logit	91	0.9833808	0.7090119	0.1732811	3.011636	0.018
FE	91	0.9700488	0.6857678	0.1729153	2.884515	0.031
FE + logit	91	0.9796653	0.7049233	0.1709613	3.011636	0.021
RE	91	0.9756051	0.6951022	0.1754419	2.941842	0.025
RE + logit	91	0.981783	0.7083012	0.1727543	3.011636	0.019
ME	91	0.9775822	0.6969206	0.1759449	2.950732	0.023
ME + logit	91	0.9764382	0.6971117	0.1729417	2.941116	0.025
Ordinal	91	0.9724439	0.6875966	0.1726732	2.868258	0.029
Multinomial	91	0.9719193	0.6888325	0.1736223	2.886721	0.029
ALDVMM-2C-CP	91	0.9713951	0.6871741	0.1734094	2.851207	0.030
ALDVMM-2C-NCP	91	0.9710432	0.6854171	0.1730018	2.853633	0.030
ALDVMM-3C-CP	91	0.9714612	0.6875973	0.1733641	2.856647	0.030
ALDVMM-3C-NCP	91	0.968886	0.6801271	0.1750478	2.915325	0.032
ALDVMM-4C-CP	91	0.9696622	0.6862922	0.1732109	2.860686	0.031
ALDVMM-4C-NCP	91	0.9704317	0.6849439	0.1742652	2.915964	0.031
One-part beta model	91	0.971463	0.688097	0.1714668	2.79137	0.030
Two-part beta model	91	0.9854843	0.7101463	0.1710772	3.011636	0.016

Table S18. Ranking of algorithms on each validation criteria using validation dataset two A

	RMSE	MAE	MAE 0.75-1	MAE 0.5- 0.75	MAE 0-0.5	MAE <0	QALY error	Average rank
Longworth	7	21	19	22	3	1	9	11.7
Veersteegh	1	24	1	23	1	4	21	10.7
Khan RE	24	23	17	21	22	22	22	21.6
Khan BB	25	25	25	25	25	25	25	25.0
Khan MM	23	19	24	18	20	19	3	18.0
Marriott FE	22	22	2	2	24	24	24	17.1
Marriott Tobit	21	20	3	4	23	23	23	16.7
OLS	8	3	9	20	5	10	13	9.7
OLS + logit	12	9	20	10	8	15	4	11.1
FE	16	16	15	7	21	20	16	15.9
FE + logit	17	17	23	1	15	21	6	14.3
RE	10	6	12	17	10	12	10	11.0
RE + logit	15	11	22	8	9	17	2	12.0
ME	9	4	11	19	6	11	12	10.3
ME + logit	14	10	21	9	7	16	1	11.1
Ordinal	3	15	6	14	12	3	7	8.6
Multinomial	2	8	4	13	2	5	5	5.6
ALDVMM-2C-CP	19	12	13	12	19	14	20	15.6
ALDVMM-2C-NCP	13	14	7	15	16	6	18	12.7
ALDVMM-3C-CP	18	13	14	11	18	18	19	15.9
ALDVMM-3C-NCP	6	5	8	5	17	8	14	9.0
ALDVMM-4C-CP	11	2	5	16	13	13	15	10.7
ALDVMM-4C-NCP	5	1	10	6	14	9	11	8.0
One-part beta	20	18	16	24	11	2	17	15.4
Two-part beta	4	7	18	3	4	7	8	7.3

Abbreviations: RMSE, Root mean squared error; MAE, Mean absolute error; QALY, Quality-adjusted life-year; OLS, Ordinary least squares; FE, Fixed effects; RE, Random effects; ME, Mixed effects.

Table S19. Ranking of algorithms on each validation criteria using validation dataset two B

	RMSE	MAE	MAE 0.75-1	MAE 0.5-0.75	MAE 0-0.5	MAE <0	QALY error	Average Rank
Longworth	7	21	6	21	2	-	8	10.8
Veersteegh	2	23	1	22	15	-	2	10.8
Khan RE	22	15	16	23	22	-	1	16.5
Khan BB	25	25	13	25	25	-	25	23.0
Khan MM	24	24	25	17	21	-	24	22.5
Marriott FE	12	3	9	20	24	-	3	11.8
Marriott Tobit	10	4	8	18	23	-	4	11.2
OLS	8	9	10	16	8	-	10	10.2
OLS + logit	3	17	3	5	7	-	6	6.8
FE	17	1	19	8	19	-	21	14.2
FE + logit	6	20	7	1	10	-	9	8.8
RE	11	7	12	14	11	-	13	11.3
RE + logit	5	19	5	6	5	-	7	7.8
ME	9	8	11	15	9	-	11	10.5
ME + logit	4	18	4	7	6	-	12	8.5
Ordinal	16	14	17	11	4	-	14	12.7
Multinomial	13	2	14	9	1	-	15	9.0
ALDVMM-2C-CP	21	6	23	13	17	-	18	16.3
ALDVMM-2C-NCP	20	13	20	19	18	-	19	18.2
ALDVMM-3C-CP	19	5	22	12	16	-	17	15.2
ALDVMM-3C-NCP	15	12	18	4	14	-	23	14.3
ALDVMM-4C-CP	18	10	21	10	12	-	22	15.5
ALDVMM-4C-NCP	14	11	15	3	13	-	20	12.7
One-part beta	23	22	24	24	20	-	16	21.5
Two-part beta	1	16	2	2	3	-	5	4.8

Abbreviations: RMSE, Root mean squared error; MAE, Mean absolute error; QALY, Quality-adjusted life-year; OLS, Ordinary least squares; FE, Fixed effects; RE, Random effects; ME, Mixed effects.