A comparison of Bayesian and frequentist approaches for mapping between the Roland Morris Disability Questionnaire and generic preference-based measures

Authors: J Madan, K Khan, SE Lamb, S Petrou

Warwick Clinical Trials Unit, Division of Health Sciences, Warwick Medical School, University of Warwick

Aims: The Roland Morris Disability Questionnaire (RMQ) is a widely used health status measure for low back pain. However, it is not preference-based, and there are currently no established algorithms for mapping between the RMQ and preference-based health-related quality-of-life measures. Using data from randomised controlled trials of treatment for low back pain, we sought to i) develop algorithms for mapping between RMQ scores and health utilities or dimension-level responses for either the EQ-5D or SF-6D; and ii) explore a range of methodological issues surrounding the development of mapping algorithms, including mapping to dimension-level responses using repeated observations, and Bayesian Markov Chain Monte Carlo (MCMC) methods for estimation of algorithms and quantification of uncertainty.

Methods: Data are primarily drawn from the Back Skills Training (BeST) trial, a trial that evaluated the clinical and cost effectiveness of adding cognitive behavioural therapy to current active management of low back pain. Data was collected from 701 patients at baseline and subsequently at 3, 6 and 12 months post-randomisation, using a range of outcome measures including the RMQ, EQ-5D and SF-12. We begin by using baseline trial data to develop direct utility mapping algorithms between RMQ scores and the EQ-5D and SF-6D derived utility scores using a number of estimators, including OLS, CLAD and GLM, and subsequently compare the fit and predictions of these alternative regression models. We then explore algorithms based on mapping between the RMQ scores and the dimension responses for the EQ-5D and SF-6D (response mapping) using baseline trial data. We go on to explore different techniques and mapping models that make use of repeated follow-up observations in the data. Analysis was carried out using both frequentist (STATA v11) and Bayesian MCMC (WinBUGS v1.4) methods, allowing for a comparison of the feasibility of these approaches, model predictions, and estimates of parameter uncertainty. The estimated mapping algorithms are provisionally validated using data from the UK Back Pain Exercise and Manipulation (BEAM) trial.

Results: We present results for direct mapping and response mapping models between the RMQ and either the EQ-5D or SF-6D fitted using alternative techniques. The models exploit both the cross-sectional and longitudinal features of the underpinning trial data. We also present results for direct mapping and response mapping models between the RMQ and the EQ-5D using Bayesian MCMC sampling.

Conclusions: Preliminary results suggest that mapping between the RMQ and generic preference-based measures are feasible, using either utility scores or dimension responses. We demonstrate alternative novel approaches for handling repeated observations and the quantification of uncertainty in mapping studies.

PAPER PRESENTED AT UK HEALTH ECONOMISTS' STUDY GROUP MEETING
UNIVERSITY OF EXETER, 9TH – 11TH JANUARY 2013
WORK IN PROGRESS, PLEASE DO NOT CITE

Introduction

Low back pain (LBP) is a major public health problem in western industrialised societies. In the UK, the annual prevalence is approximately 37% and LBP is so common that it affects almost everyone at some time during his or her lifetime [1, 2]. The direct health care costs associated with LBP in 1998 were £1,628 million (1998 prices); the majority of this expenditure was on physiotherapy and general practice. Between the years of 1994 and 1995, 116 million production days were lost in the UK due to LBP costing an estimated £10,668 million (1998 prices) in production and informal care costs [1, 3].

In order to treat individuals with LBP, new treatments and therapies are being developed and clinical trials have been run to test the efficacy of these new treatments [4-6]. Outcome measures within these trials include disease specific and generic health-related quality of life (HRQoL) instruments. The Roland Morris Questionnaire (RMQ) is one of the most extensively used and recognised disease specific outcome measures in back pain studies [7]. However, it is not preference-based, i.e. it cannot be used in the calculation of quality-adjusted life years (QALYs). Many decision makers, such as the National Institute for Health and Clinical Excellence (NICE) in England and Wales, recommend the use of the QALY as a standard measure of benefit for economic evaluation purposes and the EQ-5D as the integral preference-based measure of HRQoL [8]. In situations where the RMQ, but no preference-based measure is collected, a potential solution is to apply a mapping (or 'cross-walk') function to convert RMQ scores into preference-based (or utility) values. 'Mapping' involves the development and use of an algorithm (or algorithms) to predict health-state utility values using data on other indicators or measures of health [9]. The algorithm can be applied to data from clinical trials, other studies or economic models containing the source predictive measure(s) to predict utility values even though the target preference-based measure was not included in the original source study of effectiveness. The predicted utility values can then be analysed using standard methods for trial-based analyses or summarised for each health state within an economic model [9]. The aim of this study is to examine the relationship between the RMQ and generic preference-based measures collected as part of clinical trials and to develop well-fitting mapping algorithms, with quantitative estimates of uncertainty in predictions. These algorithms can be used to derive utilities for secondary analyses such as economic evaluations where the supporting dataset only includes RMQ information (in which case correctly quantifying mapping uncertainty is required for probabilistic sensitivity analyses). In addition, we use this mapping exercise for exploring methodological innovations exploiting the information available from the repeated observations in the datasets, including the use of Bayesian methods to aid development of flexible algorithms and quantification of prediction uncertainty. We start off by describing the data and outcome measures used in the analyses and move on to briefly explain the statistical methods employed in the study, the results of the analyses are then presented, and finally we discuss the findings and identify areas for future research.

Methods

Data

The data for this study are primarily drawn from the Back Skills Training Trial (BeST) [4]. BeST was a pragmatic, multicentre, randomised controlled trial that recruited 701 participants from 56 general practices in seven regions across England. Individuals were eligible for inclusion if they were aged 18 years or older, had at least moderately troublesome sub-acute or chronic low back pain of a minimum of 6 weeks duration, and had consulted for low-back pain in primary care within the preceding 6 months. The main aim of the BeST study was to estimate the clinical effectiveness of two interventions: active management and active management combined with a cognitive

behavioural approach in reducing disability associated with low back pain. The outcome measures in BeST included the RMQ, EQ-5D and SF-12v2 and these were collected at baseline, 3, 6 and 12 months post-treatment. Of 701 trial participants, only 675, 506, 516, and 491 participants provided RMQ, EQ-5D, age and gender data at baseline, 3, 6 and 12 months, respectively. For the RMQ and SF-12, the equivalent figures were 664, 492, 506 and 476 participants.

Preliminary external validation of the algorithms has been performed using data from the UK Back pain Exercise and Manipulation trial (UK BEAM) [10]. The UK BEAM study recruited and randomised 1334 participants to one of four interventions: manipulation, exercise, manipulation combined with exercise or best care in general practice. The main aim of this trial was to assess the effectiveness of adding these interventions to best care in general practice settings. The outcome measures in UK BEAM included the RMQ, EQ-5D and SF36v2 and these were collected at baseline, 1, 3 and 12 months.

Outcome measures

The RMQ was originally derived from the Sickness Impact Profile [11] and it contains 24 items relating to a range of functions commonly affected by low back pain and disability. Completion of the RMQ scale comprises responses to of 24 statements each with binary 'Yes'/'No' options. The total number of positive responses is summed to form a score (out of 24), and a low score is associated with less disability.

The EQ-5D [12] is the most widely used generic preference-based measure of health outcomes [13]. It consists of two principal measurement components, a descriptive system and a visual analogue scale. This study concentrates on mapping onto information provided by the descriptive system, which is comprised of five dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension is assessed by a single question on a three-point ordinal scale (no problems, some or moderate problems, severe or extreme problems). This generates a total number of 243 (3^5) possible EQ-5D health states. In the UK, utility valuations for all 243 EQ-5D health states are commonly based on the York A1 tariff set derived from a survey of the UK population (n = 3337), which had used the time trade-off valuation method to estimate utility scores for a subset of 45 EQ-5D health states and a multivariate model to value the remainder of the EQ-5D health states [14]. Alternative tariffs for the EQ-5D have been developed for other countries [15].

The SF-6D [16] is a preference-based instrument derived from the Short Form 36 (SF-36) [17], a widely used instrument that measures HRQoL. Like the EQ-5D, the SF-6D consists of a descriptive system and a valuation of health states. There are two versions of the SF-6D, one based on the complete 36-item version of the SF-36 [16], and the other based on its 12-item (SF-12) derivative [18]. The SF-6D has 6 dimensions (pain, mental health, physical functioning, social functioning, role limitations, and vitality) and each dimension has between three and six levels, depending on the response choice categories of the original items from the SF-36. The SF-6D was valued by a representative sample of 836 members of the UK general population using the standard gamble valuation method.

Statistical methods

Models fitted to baseline data using frequentist approaches

For the first stage of our analyses we estimated mapping algorithms using only the baseline data from BeST, and made use of both direct utility and response mapping approaches to estimate EQ-5D and SF-6D utility scores based on the RMQ. The direct utility approach makes use of regression equations to predict the values of one outcome measure using scores/values from a second measure as regressors. The coefficients of the models can then be used to carry out the conversion from the starting measure to the target measure in the required dataset. The use of linear regression can be

problematic because of underpinning assumptions of normality and homoscedasticity in the error term. Therefore, as well as Ordinary Least Squares (OLS) regression, we also explore more flexible regression models including Tobit [19], Fractional logistic (FLOGIT) [20, 21], Censored Least Absolute Deviations (CLAD) [22, 23] and Generalized Linear Models (GLM) [24]. We also used response mapping to predict the responses to either the EQ-5D or SF-6D dimensions as opposed to predicting the summary utility score directly [25]. These models were estimated by fitting a separate multinomial logistic regression model to predict the probability of each response level for each EQ-5D or SF-6D dimension, as described elsewhere [25].

For each model we included either the overall RMQ score (from here on referred to as model 1) or individual responses to each of the 24 RMQ questions (from here onwards referred to as model 2) as explanatory variables. All models also included age and gender variables as covariates and were estimated in STATA version 11 (Stata-Corp, College Station, TX).

Analysing repeated measurement data

For the second part of the analyses, we explored mapping algorithms for the EQ-5D and SF-6D utility scores based on the RMQ scores using the repeated measurements available in the BeST dataset at baseline, 3 month, 6 months and 12 months. This type of longitudinal data can be seen to have a two-tier structure where the measurement occasions (level 1 units) are nested within subjects (level 2 units). By analysing the results at level 1 and ignoring the multi-level nature of the data, we risk overlooking the importance of group effects, which in turn can render invalid the model estimated coefficients. In order to address the issue of multiple observations per subject, two main methods were used; random-effects models and OLS with robust standard errors. The XTMIXED command was used in STATA to create multi-level models. We fitted both random intercept models (in which the predicted utility for an RMQ score of 0 varies between individuals) and the more flexible random-coefficient model (in which the change in utility per unit change in RMQ also varies between individuals). We also explored the use of OLS with robust standard errors (equivalent to White's standard errors in the presence of heteroskedasticity), which involves relaxing the assumption of independence of the observations, grouping the data according to the subject identifier and only treating those observations with different subject identifiers as truly independent. Within STATA, this is achieved by using the cluster command on the subject identifiers.

Assessing model performance

Predicted EQ-5D utilities were estimated for each mapping model. For the direct utility models, the predictions were generated by using the 'predict' post-estimation command, with direct back-transformations applied to predictions from FLOGIT and GLM models. Any utilities predicted to be >1 using the OLS models were set to one. For the response mapping models, the predictions were generated using the expected value method [26]. This is equivalent to the Monte Carlo method [25] given a large number of repeated Monte Carlo draws.

In line with external guidance [9, 27], the mean squared error (MSE) and the mean absolute error (MAE) were used to measure the goodness of fit of the models. The MSE equals the mean of squared differences between observed and predicted EQ-5D or SF-6D utility scores, whilst the MAE is the mean of absolute differences between observed and predicted EQ-5D or SF-6D utility scores. The performance of the RMQ to EQ-5D models using baseline data was also assessed using the UK BEAM dataset to see how well the models performed with out of sample data (external validation).

The MAE was also calculated for response mapping models in terms of the predicted response level rather than the utility score. This was done by weighting each possible level by its predicted probability, and taking the difference between the sums of these weighted levels and the actual

response. For example, if the mapping predicts a 20% probability of a level 1 response and an 80% probability of a level 2 response, the predicted response would be 1.8.

The distributions of the predicted and observed EQ-5D or SF-6D utility scores were also examined to see how closely the predicted values matched the observed scores [28], and the proportions of predictions deviating from observed values by <0.10 or <0.25 were also calculated since these indicate the distribution of errors and how often the models fail to produce useful predictions [29].

Estimation of mapping algorithms using Bayesian MCMC

Estimation of relationships between non-preference-based measures and overall health utilities or responses to individual dimensions of preference-based measures, such as the EQ-5D or SF-6D, can be challenging due to issues such as bounded values and non-Gaussian error distributions. Estimation of distributions for mapping parameters using Bayesian updating via Markov Chain Monte Carlo (MCMC) sampling is a promising alternative approach that has the additional advantage of formally quantifying parameter uncertainty. To explore the potential of this approach, we first explored its use in the direct mapping for the OLS and Fractional Logit models (other direct mapping models will be fitted in further work). We then explored a range of response mapping models using this approach. For the Bayesian models, we assume that individual i has a probability $p^{M}_{i,k}$ of responding positively to each of the 24 RMQ questions, so that the total score is a binomially distributed random variable, and use the log-odds of a positive response as the explanatory variable. The first response mapping involved a multinomial logit model using baseline observations from BeST to map RMQ scores to individual EQ-5D dimensions. Future work will extend these analyses to cover response mapping onto the SF-6D.

We then explored two modelling approaches for constructing mapping, based on all (including repeated) observations from BeST, between the RMQ total score and the individual dimensions of the EQ-5D. The first was to fit a random coefficient multi-level model as described above (except that the dependent variable is now the probability of a given response, on a log-odds scale, rather than health utility). The second approach involved relating changes between follow-ups in each EQ-5D dimension response to the change in RMQ score over that time. Let $Q_{i,j,k}$ be the response (1, 2 or 3) given by individual i for dimension j at the k^{th} follow-up. For k=1 (baseline), we assume the response can be predicted from the RMQ score at baseline using the multinomial logit model described above. For follow-up observations (k>1), we define $p^{\Delta_{XY}}_{i,j,k}$ as the probability individual i gives a response Y for dimension j at follow-up k+1, if they gave a response X at follow-up k, i.e.

$$p^{\Delta_{XY}}_{i,j,k} = P\ \ Q_{i,j,k+1} = Y \mid Q_{i,j,k} = X$$

It follows that, for each X:

$$\sum_{Y=1}^{3} p^{\Delta_{XY}}_{i,j,k} = 1$$

A multinomial logit model was assumed for these transition probabilities, with age, gender and change in probability of positive response to each RMQ question (on a log-odds scale) as explanatory variables:

$$\text{logit}\ \ p^{\Delta_{XY}}_{i,j,k}\ = \alpha^{\Delta_{XY}}_{j} + \beta^{\Delta_{XY}}_{j}\ \ \text{logit}\ \ p^{M}_{i,k}\ - \text{logit}\ \ p^{M}_{i,k-1}\ \ + \beta^{\Delta_{A}}_{j} A_i + \beta^{\Delta_{G}}_{j} G_i$$

A_i and G_i are the age and gender of participant i respectively, and $p_{i,k}^{M}$ is the probability of a positive response to each RMQ question at the k^{th} follow-up (as defined above).

For each model the explanatory variables were total the RMQ score, age and gender. MCMC sampling was carried out in WinBUGS version 1.4.2. Performance statistics such as the MAE and MSE were estimated for models fitted using Bayesian updating from MCMC samples generated using the 'coda' command in WinBUGS (further work will involve estimating the Deviance Information Criterion (DIC)). 25,000 simulations were used for estimation, after a burn-in sample of 25,000 had been discarded. All Bayesian models involved mapping the RMQ to EQ-5D utilities and dimensions. Further work is planned to extend this approach to cover the SF-6D.

Results

Study sample characteristics

Table 1 presents descriptive statistics for the BeST study population. The mean EQ-5D and SF-6D utility scores for the study population at baseline were 0.56 (SD 0.28) and 0.62 (SD 0.12), respectively. The mean EQ-5D and SF-6D utility scores for the validation sample (UK BEAM trial participants) at baseline were 0.58 (SD 0.25) and 0.64 (SD 0.07), respectively.

Table 2 and figure 1 shows the distribution of the outcome measures for the BeST study population at baseline. The EQ-5D utility scores are negatively skewed and bi-modally distributed. The RMQ scores have a right skewed distribution and the SF-6D utility scores appear to follow a normal distribution.

For the direct utility mapping of RMQ to either the EQ-5D or SF-6D, the Tobit model was not used as there was no mass of values at either end of the range of the utility scale. When using the GLM estimator for the EQ-5D data, the only model that converged was the GLM with Gaussian family and identity link, which produces the same coefficients as the OLS model and therefore the results are not reported. For the SF-6D utility data, the GLM model selected as being best suited to the data was that with a gamma distribution and log link.

When considering the use of a mapping algorithm, its ability to accurately predict the mean EQ-5D utility score is important, but should not be the only consideration. The performance of all of the models using the different estimators was assessed using the estimation sample. In addition to assessing how accurately the models estimated the mean EQ-5D or SF-6D utility score, we also examined the distributions of the predicted scores.

Results of baseline models

Table 3 presents the results of the mapping analyses between the RMQ and EQ-5D carried out using the baseline BeST data. The results show that most of the models were able to fairly accurately predict the mean EQ-5D utility score in the estimation sample (0.5646) with predicted mean EQ-5D utility scores ranging from 0.5560 to 0.5669. The exceptions were the CLAD models, which predicted mean EQ-5D utility scores of 0.6463 and 0.6211. The MLOGIT (1) and (2) models for each estimator were able to predict utility scores into the negative range, with the CLAD (2) model predicting the value closest to the most extreme (-0.3061 compared with -0.3490). None of the models were able to predict the highest EQ-5D utility score, although the Bayesian response mapping model came very close (0.991). Models using the individual RMQ items (model 2) generated consistently better predictions. On average, the CLAD (2) model gave the widest range of predicted scores and the FLOGIT (2) model gave the narrowest range of predictions. The SDs of the predicted mean EQ-5D utility scores estimated using the OLS (2), FLOGIT (2), CLAD (2) and MLOGIT (2) models were

respectively only 65%, 65%, 59% and 69% of the magnitude of the observed EQ-5D scores, on average. If we compare the 25th percentile, median and 75th percentile predicted scores to the observed EQ-5D utility scores, no single model can be selected as closely matching the observed scores. The MLOGIT (2) model generates the closest matching score for the 25th percentile, and the CLAD (2) model generates the closest matching score for the median and 75th percentile. Model performance varied when considering the proportions of predictions deviating from observed values. The OLS (2), FLOGIT (2), CLAD (2), and MLOGIT (2) models achieved 45%, 48%, 63% and 55% of individual estimations within 0.10 of the observed values. The models produced similar results when looking at deviations within 0.25 of the observed values with all models achieving around 78-80% of individual estimations. For OLS regression, results from Bayesian MCMC estimation were identical to those generated in STATA, and are not shown. However, Bayesian MCMC estimation of the FLOGIT model gave slightly better fit in terms of MAE and percentage of observations with accuracy <10%.

The performance of the mapping analyses between the RMQ and EQ-5D using the baseline data was also assessed using the validation sample (UK BEAM trial data not presented) and the results were consistent with those from the estimation sample. For example, the MSE and MAE for the OLS (2) model using the validation sample was 0.0460 and 0.1635 compared with 0.0448 and 0.1563 using the estimation sample. Also for the MLOGIT (2) model, the MSE in the validation sample was 0.0485 compared with 0.0429 in the estimation sample.

Table 4 presents the results of the mapping analyses between the RMQ and SF-6D carried out using the baseline BeST data. The results show that OLS and GLM estimators were able to fairly accurately predict the mean SF-6D utility score in the estimation sample (0.6236) with predicted mean SF-6D utility scores ranging from 0.6232 to 0.6236. The CLAD and MLOGIT models produced mean SF-6D utility scores ranging from 0.6080 to 0.619. The CLAD (2) model was able to predict the lowest SF-6D utility value closest to the most extreme (0.3570 compared with 0.3450). None of the models were able to predict the highest SF-6D utility score. As was the case for the mapping algorithms from the RMQ to the EQ-5D, the models with the individual RMQ questions as dummy variables performed better with both lower MSE and MAE values. On average, the CLAD (2) model gave the widest range of predicted scores and the MLOGIT (2) model gave the narrowest range of predictions. The SDs of the predicted mean SF-6D utility scores estimated using the OLS (2), CLAD (2), GLM (2) and MLOGIT (2) models were respectively 64%, 60%, 65% and 64% of the magnitude of the observed SF-6D scores, on average. If we compare the 25th percentile, median and 75th percentile predicted scores to the respective observed EQ-5D utility scores, no single model can be selected as closely matching the observed scores. The CLAD (2) model has the closest matching score for the 25th percentile and median, and the OLS (2) model has the closest matching score for the 75th percentile. All models achieved around 73-77% of individual estimations within 0.10 of the observed values and 99% of individual estimations within 0.25 of the observed values.

Table 5 gives the predicted probability of each response level for each EQ-5D dimension as a function of the RMQ total score. Predictions (median and 95% credible intervals) are generated using the Bayesian MCMC response mapping model fitted to BeST baseline data only. The relationship between the RMQ total score and probability of a given response level is non-linear, so that sensitivity of response level to RMQ score changes over the RMQ range, and the zone of greatest sensitivity to RMQ varies across dimensions. For example, mobility response is most sensitive to RMQ score for values between 3 and 9, whereas self-care response is most sensitive to RMQ score for values between 13 and 19. Credible interval widths vary substantially across dimensions and RMQ scores, although a typical interval width is around 0.1.

Results of repeated measures models

Table 6 presents the results of the mapping analyses between the RMQ and EQ-5D carried out using the repeated measurements. The results show that all models were able to fairly accurately predict the mean EQ-5D utility score in the estimation sample (0.5997) with predicted mean EQ-5D utility scores ranging from 0.6003 to 0.6004. At the lower end of the utility range, the random coefficient model was able to predict the closest EQ-5D utility score to the most extreme (-0.2861 compared with -0.4840). None of the models were able to predict the highest EQ-5D utility score, but the random intercept model came closest with a prediction of 0.9864 compared to 1.00. On average, the random intercept model gave the widest range of predicted scores and the OLS robust cluster model gave the narrowest range of predictions. The SDs of the predicted mean EQ-5D utility scores estimated using the OLS robust cluster, random intercept, and random coefficient models were respectively 66%, 75%, and 77% of the magnitude of the observed EQ-5D scores, on average. If we compare the 25th percentile, median and 75th percentile predicted scores to the observed EQ-5D utility scores, the random coefficient model has the closest matching score for the 25th percentile and median, and the random intercept model has the closest matching score for the 75th percentile. Model performance varied when considering the proportions of predictions deviating from observed values. The OLS robust cluster, random intercept, and random coefficient models achieved 44%, 61%, and 65% of individual estimations within 0.10 of the observed values and 65% and 80%, 88%, and 89% of individual estimations within 0.25 of the observed values. The random coefficient model had the lowest MSE and MAE of all of the RMQ to EQ-5D models using repeated measurements.

Table 7 presents the results of the mapping analyses between the RMQ and SF-6D carried out using the repeated measurements. The results show that all models were able to fairly accurately predict the mean SF-6D utility score in the estimation sample (0.6572) with predicted mean SF-6D utility scores ranging from 0.6575 to 0.6579. The random intercept model was able to predict the closest SF-6D utility value to the lowest observed (0.3518 compared with 0.3450). None of the models were able to predict the highest SF-6D utility score, but the random coefficient model came closest with a prediction of 0.9286 compared to 1.00. On average, the random intercept model gave the widest range of predicted scores and the OLS robust cluster model gave the narrowest range of predictions. The SDs of the predicted mean EQ-5D utility scores estimated using the OLS robust cluster, random intercept, and random coefficient models were respectively 67%, 79%, and 80% of the magnitude of the observed SF-6D utility scores, on average. If we compare the 25th percentile, median and 75th percentile predicted scores to the observed SF-6D utility scores, the random coefficient model generated the closest predicted score at the 25th percentile and median, whilst the random intercept model generated the closest predicted score at the 75th percentile. Model performance varied when considering the proportions of predictions deviating from observed values. The OLS robust cluster, random intercept, and random coefficient models achieved 68%, 88%, and 90% of individual estimations within 0.10 of the observed values, respectively. The proportion of predictions achieved within 0.25 of actual values was 99% for the OLS robust cluster model and 100% for the random intercept and random coefficient models. The random coefficient model had the lowest MSE and MAE of all of the RMQ to SF-6D models using the repeated measurements.

Results of Bayesian change from baseline models

Table 8 shows the MAE for each of the Bayesian MCMC response mapping models for predicted EQ-5D utility scores as well as responses to each dimension. MAEs are presented separately for each of the 4 time points for which data are available. In terms of utility predictions, model predictions are comparably accurate between baseline and multilevel models, and the model relating changes in RMQ to changes in response is less accurate, particularly at 3 and 12 months. There are noticeable differences in model prediction accuracy across dimensions, with anxiety/depression having the

highest MAEs. While the transition model is less accurate overall, it does give more accurate predictions for certain dimensions (self-care and anxiety/depression).

Comparison of model algorithms

As well as comparing predictive accuracy for alternative mapping models, it is useful to compare the algorithm each generates for mapping between RMQ score and EQ-5D (at an overall utility or dimension level). Table 9 presents predicted utility values for a given RMQ score using a number of different models. The algorithms are generally comparable, with the clearest exception being the CLAD algorithm, which generates substantially higher utility predictions at the upper end of the RMQ score range. The transition mapping model is not reported in table 9 since it generates a different type of mapping – from changes in RMQ to probabilities of changing response for each dimension. Figure 2 illustrates the mapping generated by this model for an individual with a starting RMQ score of 9. In general, transitions between levels 1 and 3 are extremely unlikely – self-care is the dimension where the transition from 1 to 3 is most sensitive to worsening RMQ, and anxiety or depression the dimension where the transition from 3 to 1 is most sensitive to improving RMQ. There are clear differences between dimensions in the degree to which changes in EQ-5D responses are sensitive to changes in RMQ score. For example, those reporting no problems with self-care at the preceding follow-up are unlikely to have worsened unless they experienced substantial increases in RMQ. Those reporting no pain at the preceding follow-up are almost certain to decline if they experience any increase in RMQ between follow-ups, whilst even small reductions in RMQ are associated with substantial increases in the probability of remaining pain-free.

Table 10 presents the impact of a reduction in RMQ score from 9 to 7 (chosen as the mean RMQ reduction in the BEST study) on the change in probability of reporting each response level for each EQ-5D dimension. Predictions are presented for each response mapping model fitted using Bayesian MCMC, in order to compare the predicted impact of the RMQ change. While all models predict that reduction in RMQ gives rise to a probability of an improved response across dimensions, the predicted impact is stronger for the transition model. The difference is most noticeable for usual activities, where the transition model predicts an increased probability of a level 1 response of 33% while the baseline model predicts an 8% increase. The only dimension where the baseline model predicts a (slightly) stronger impact than the transition model is self-care. As a result, the transition model predicts this reduction in RMQ will lead to a larger increase in utility score than the baseline model (0.078 vs 0.051).

Discussion

We have constructed a range of models to generate algorithms for mapping between the RMQ and the EQ-5D and SF-6D, all of which we were able to fit to the estimation dataset with accuracy comparable to other direct utility mappings in the literature [27]. We have also provisionally validated these algorithms against an additional dataset (observations from the UK BEAM trial) with promising results. Several criteria have been proposed for assessing the fit of such algorithms to the estimation sample, which we have estimated for our models. While these criteria can lead to different results, one consistent finding is that model accuracy can be improved by considering individual RMQ responses rather than the total RMQ score. Further work will explore, using approaches such as Rasch analysis, whether particular RMQ questions, singly or in combination, can be used to further improve prediction of utility and EQ-5D or SF-6D responses.

It can be instructive to compare the mapping algorithms generated by alternative models, as well as comparing the fit with the estimation dataset, particularly if predictive accuracy is comparable but the predictions themselves differ substantially. Considering the mappings between RMQ score and EQ-5D fitted using baseline observations as an example, while the CLAD model performs well in comparison with OLS and FLOGIT models, it also gives markedly different predictions of utility for

high RMQ scores. There are fewer observations at the upper end of the RMQ scale in the estimation dataset, but their utilities are predicted poorly by the CLAD model, suggesting that this model achieves small predictive gains at low-medium RMQ scores by allowing for significant prediction and it may not be appropriate to use this model for mapping populations with high RMQ scores. Further work is underway exploring the use of polynomial regression models to gain improved predictive accuracy across the RMQ scale.

We estimate models to predict responses to each dimension of the EQ-5D or SF-6D. Our results suggest that this leads to predictions of comparable accuracy to direct mapping, and it provides insights into the sensitivity of different health dimensions to the RMQ score. For example, response mapping using baseline observations suggests that predictions are most accurate for pain and self-care, and least accurate for anxiety or depression. This is consistent with the fact that the RMQ does not include questions on anxiety/depression (although it does include one question on being irritable or bad-tempered). The response mapping model also illustrates how responses worsen at different points of the RMQ scale for different dimensions. For both direct and response mapping, we explore the impact of including follow-up observations on the mapping. While this does appear to result in substantial improvements in model fit for the direct mapping, this may reflect use of correlations between repeated observations from the same individual, rather than improvements in the mapping itself. For the response mapping, multi-level modelling of all observations appears to offer comparable accuracy to models based on baseline observations alone, although it does shift utility predictions upwards at low RMQ scores and downwards at high RMQ scores.

We explore two approaches to using the follow-up observations for response mapping – multi-level models and models of change in response. The latter approach effectively models response levels as states, and attempts to predict the probability of transitions between levels at successive follow-ups in terms of RMQ score changes over that time. This approach leads to less accurate predictions of utility overall compared with other response mapping models. However, it is important to also consider the intended use of the mapping, as well as fit with the estimation dataset, when selecting the appropriate model. The transition response mapping estimates how predictable changes in EQ-5D responses are given changes in RMQ, whereas all other models estimate the correlation between outcome measures at a fixed point in time. These different relationships may have independent causal factors or confounders. For example, at a given time, factors other than back pain may influence anxiety or depression levels, reducing the correlation between this response and RMQ score. This may explain why the transition model predicts anxiety or depression responses more accurately than the other response mapping models, as it identifies the impact of improved back-related health on mental wellbeing. If the motivation for the mapping exercise is to translate treatment effects from one scale to another, as it will often be when mapping is carried out to inform economic evaluation, then results based on transition mapping models may give more accurate predictions of treatment effects. Our results suggest that, for a typical individual in the estimation dataset, using the 'stationary' mapping instead would underestimate the health gain from the typical change in RMQ.

We explore the use of both frequentist MLE and Bayesian MCMC methods to estimate mapping algorithms. We found apparent advantages from using Bayesian methods – it was easier to fit a wider range of models, including the transition response mapping, and easier to generate estimates of uncertainty around the predictions from the mapping (particularly for response mapping models).

References

1. Lamb, S.E., et al., *A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain: the Back Skills Training (BeST) trial.* Health Technology Assessment, 2010. **14**(41): p. 1-281.
2. Savigny, P., et al., *Low Back Pain: early management of persistent non-specific low back pain.* London: National Collaborating Centre for Primary Care and Royal College of General Practitioners, 2009. **14**.

3. Maniadakis, N. and A. Gray, *The economic burden of back pain in the UK.* Pain, 2000. **84**(1): p. 95-103.
4. Lamb, S.E., et al., *Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis.* The Lancet, 2010. **375**(9718): p. 916-923.
5. Sherman, K.J., et al., *A randomized trial comparing yoga, stretching, and a self-care book for chronic low back pain.* Archives of Internal Medicine, 2011: p. archinternmed. 2011.524 v1.
6. Hill, J.C., et al., *Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial.* The Lancet, 2011. **378**(9802): p. 1560-1571.
7. Kopec, J.A., *Measuring functional outcomes in persons with back pain: a review of back-specific questionnaires.* Spine, 2000. **25**(24): p. 3110-3114.
8. National Institute for Health and Clinical Excellence, (NICE), *Guide to the methods of technology appraisal.* National Institute for Health and Clinical Excellence (NICE), 2008.
9. Longworth, L. and D. Rowen, *NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values*, 2011.
10. Russell, I., et al., *United Kingdom back pain exercise and manipulation (UK BEAM) randomised trial: cost effectiveness of physical treatments for back pain in primary care.* BMJ, 2004. **329**(7479): p. 1381.
11. Roland, M. and R. Morris, *A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain.* Spine, 1983. **8**(2): p. 141-4.
12. The, E.Q.G., *EuroQol-a new facility for the measurement of health-related quality of life.* Health Policy, 1990. **16**(3): p. 199-208.
13. Räsänen, P., et al., *Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review.* International Journal of Technology Assessment in Health Care, 2006. **22**(02): p. 235-241.
14. Dolan, P., *Modeling valuations for EuroQol health states.* Medical Care, 1997: p. 1095-1108.
15. Szende, A., M. Oppe, and N. Devlin, *EQ-5D value sets: inventory, comparative review and user guide*. Vol. 2. 2006: Springer.
16. Brazier, J., J. Roberts, and M. Deverill, *The estimation of a preference-based measure of health from the SF-36.* Journal of Health Economics, 2002. **21**(2): p. 271-292.
17. Ware Jr, J.E. and C.D. Sherbourne, *The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection.* Medical Care, 1992: p. 473-483.
18. Brazier, J.E. and J. Roberts, *The estimation of a preference-based measure of health from the SF-12.* Medical Care, 2004. **42**(9): p. 851-859.
19. Tobin, J., *Estimation of relationships for limited dependent variables.* Econometrica: Journal of the Econometric Society, 1958: p. 24-36.
20. Papke, L.E. and J. Wooldridge, *Econometric methods for fractional response variables with an application to 401 (k) plan participation rates*, 1993, National Bureau of Economic Research Cambridge, Mass., USA.
21. Levy, A., T. Christensen, and J. Johnson, *Utility values for symptomatic non-severe hypoglycaemia elicited from persons with and without diabetes in Canada and the United Kingdom.* Health and Quality of Life Outcomes, 2008. **6**(1): p. 73.
22. Powell, J.L., *Least absolute deviations estimation for the censored regression model.* Journal of Econometrics, 1984. **25**(3): p. 303-325.
23. Chay, K.Y. and J.L. Powell, *Semiparametric censored regression models.* The journal of economic perspectives, 2001. **15**(4): p. 29-42.
24. McCullagh, P. and J.A. Nelder, *Generalized linear models*. Vol. 37. 1989: Chapman & Hall/CRC.
25. Gray, A.M., O. Rivero-Arias, and P.M. Clarke, *Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping.* Medical Decision Making, 2006. **26**(1): p. 18-29.
26. Le, Q.A.P.P. and J.N.P. Doctor, *Probabilistic Mapping of Descriptive Health Status Responses Onto Health State Utilities Using Bayesian Networks: An Empirical Analysis Converting SF-12 Into EQ-5D Utility Index in a National US Sample.* Medical Care, 2011. **49**(5): p. 451-460.
27. Brazier, J., et al., *A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures.* The European Journal of Health Economics, 2010. **11**(2): p. 215-225.
28. Kaambwa, B., L. Billingham, and S. Bryan, *Mapping utility scores from the Barthel index.* The European Journal of Health Economics: p. 1-11.
29. Dakin, H., et al., *Mapping analyses to estimate health utilities based on responses to the OM8-30 otitis media questionnaire.* Quality of Life Research, 2010. **19**(1): p. 65-80.

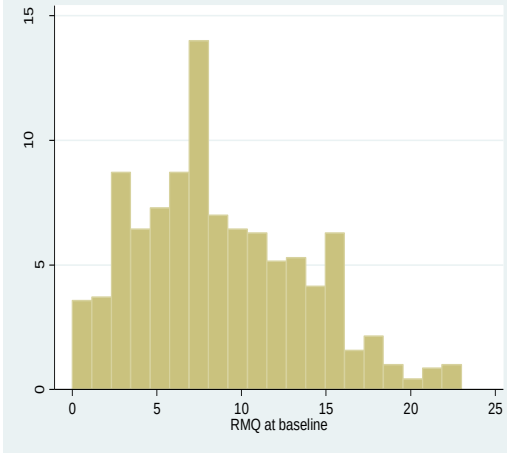
Table 1: Demographic characteristics for all BeST participants

Details	
Number of participants, N	701
Age (years), Mean (SD)	53.62 (14.69)
Missing, n (%)	2 (0.29%)
Gender, n (%)	
Male	279 (39.80%)
Female	420 (59.91%)
Missing	2 (0.29%)
Ethnic origin, n(%)	
White	618 (88.16%)
Mixed	7 (1.00%)
Asian	29 (4.14%)
Black	11 (1.57%)
Chinese	2 (0.29%)
Not specified	34 (4.85%)
Left full-time education (years), n (%)	
16 or less	387 (55.21%)
17-19	165 (23.54%)
20 or over	113 (16.12%)
Still in full-time education	3 (0.43%)
Missing	33 (4.71%)
Currently working, n (%)	
Yes, full-time	227 (32.38%)
Yes, part-time	121 (17.26%)
No	350 (49.93%)
Missing	3 (0.43%)
Received benefit payments (of any type), n (%)	
Yes	277 (39.51%)
No	416 (59.34%)
Missing	8 (1.14%)
Back pain trouble-someness, n (%)	
Slight	38 (5.42%)
Moderate	315 (44.94%)
Very	229 (32.67%)
Extreme	53 (7.56%)
Missing	66 (9.42%)
RMQ score at baseline, Mean (SD)	8.68 (4.90)
Missing, n (%)	1 (0.14%)
EQ-5D utility score at baseline, Mean (SD)	0.56 (0.28)
Missing, n (%)	22 (3.14%)
SF-6D utility score at baseline, Mean (SD)	0.62 (0.12)
Missing, n (%)	33 (4.71%)

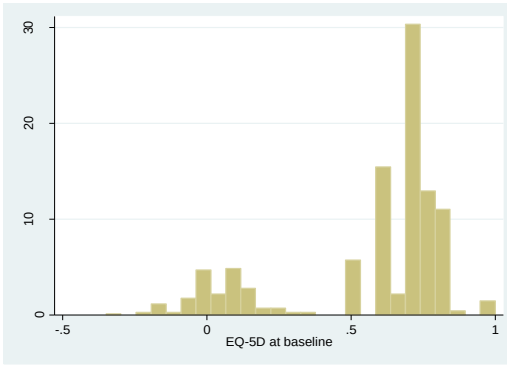
Table 2: Distribution of outcome measures across the BeST sample at baseline

Characteristic		N (%)
EQ-5D Dimension (N=679)		
Mobility, n (%)		
Level 1		276 (40.65%)
Level 2		403 (59.35%)
Level 3		0
Self-Care, n (%)		
Level 1		545 (80.27%)
Level 2		131 (19.29%)
Level 3		3 (0.44%)
Usual Activities, n (%)		
Level 1		147 (21.65%)
Level 2		502 (73.93%)
Level 3		30 (4.42%)
Pain/Discomfort, n (%)		
Level 1		18 (2.65%)
Level 2		544 (80.12%)
Level 3		117 (17.23%)
Anxiety/Depression, n (%)		
Level 1		349 (51.40%)
Level 2		296 (43.59%)
Level 3		34 (5.01%)
EQ-5D VAS Score, Mean(SD)		64.65 (18.66)
Range, (IQR)		(0, 100), 30
EQ-5D Utility Score, Mean(SD)		0.56 (0.28)
Range, (IQR)		(-0.35, 1.00), 0.24
SF-6D Dimension (N=668)		
Physical functioning		
Not limited at all		113 (16.92%)
Limited a little		331 (49.55%)
Limited a lot		224 (33.53%)
Role limitations		
None of the time		34 (5.09%)
A little of the time		178 (26.65%)
Some of the time		15 (2.25%)
Most of the time		441 (66.02%)
All of the time		0
Social functioning		
None of the time		205 (30.69%)
A little of the time		134 (20.06%)
Some of the time		237 (35.48%)
Most of the time		73 (10.93%)
All of the time		19 (2.84%)
Bodily pain		
Not at all		22 (3.29%)
A little bit		128 (19.16%)
Moderately		243 (36.38%)
Quite a bit		222 (33.23%)
Extremely		53 (7.93%)
Mental health		
None of the time		122 (18.26%)
A little of the time		201 (30.09%)
Some of the time		226 (33.83%)
Most of the time		94 (14.07%)
All of the time		25 (3.74%)
Vitality		
All of the time		16 (2.40%)
Most of the time		119 (17.81%)
Some of the time		246 (36.83%)
A little of the time		191 (28.59%)
None of the time		96 (14.37%)
SF-6D Utility Score, Mean(SD)		0.6236 (0.1207)
Range, (IQR)		(0.35, 1.00), 0.156

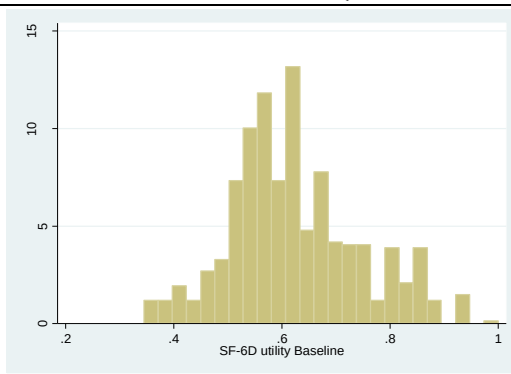
Figure 1: Distribution of the Outcome measures across the BeST data sets



Distribution of RMQ score



Distribution of EQ-5D



Distribution of SF-6D

Table 3: Model Performance using baseline BeST measurements- EQ-5D

	Predicted EQ-5D Values								Abs Diff	Abs Diff
Model	**Mean (SD)**	**Min**	**P.25**	**Median**	**P.75**	**Max**	**MSE**	**MAE**	**<0.10 (%)**	**<0.25 (%)**
Observed EQ-5D	0.5646 (.2803)	-0.3490	0.5160	0.6890	0.7600	1	-	-	-	-
OLS (1)	0.5665 (0.1676)	0.0688	0.4509	0.5882	0.6914	0.8882	0.0499	0.1674	41.93%	76.30%
OLS (2)	0.5668 (0.1816)	-0.1069	0.4523	0.6071	0.7020	0.8962	0.0448	0.1563	45.04%	77.78%
FLOGIT (1)	0.5666 (0.1683)	0.0308	0.4584	0.6035	0.6952	0.8289	0.0494	0.1620	45.33%	76.74%
Bayesian FLOGIT (1)	0.5875 (0.1740)	0.0218	0.4769	0.6300	0.7194	0.8471	0.0499	0.1582	49.63%	77.93%
FLOGIT (2)	0.5669 (0.1821)	-0.0814	0.4607	0.6185	0.7045	0.8361	0.0442	0.1518	48.44%	78.52%
CLAD (1)	0.6463 (0.0982)	0.3600	0.5800	0.6600	0.7200	0.8200	0.0613	0.1559	63.70%	79.10%
CLAD (2)	0.6211 (0.1643)	-0.3061	0.5767	0.6625	0.7202	0.8145	0.0545	0.1457	63.26%	80.74%
MLOGIT (1)	0.5601 (0.1773)	-0.0820	0.4728	0.6187	0.6911	0.7783	0.0499	0.1571	50.07%	77.48%
MLOGIT (2)	0.5560 (0.1941)	-0.1126	0.4787	0.6321	0.7031	0.8299	0.0429	0.1442	54.81%	79.85%
Bayesian Response mapping (1)	0.6020 (0.1719)	0.0510	0.5190	0.6499	0.7174	0.9905	0.0510	0.1521	55.30%	79.26%

Dependent variable for OLS, Tobit, Fractional Logistic and CLAD was EQ-5D utility score. Dependent variables for MLOGIT models were EQ-5D dimension scores.
Independent variable(s): Model (1) RMQ score, Model (2) RMQ questions entered as categorical variables. Age and gender variables present in all models.

Table 4: Model Performance using baseline BeST measurements- SF-6D (SF12 version)

	Predicted SF6D Values								Abs Diff	Abs Diff
Model	**Mean (SD)**	**Min**	**P.25**	**Median**	**P.75**	**Max**	**MSE**	**MAE**	**<0.10 (%)**	**<0.25 (%)**
Observed SF6D utility	0.6236 (0.1207)	0.345	0.538	0.615	0.694	1.000	-	-	-	-
OLS (1)	0.6236 (0.0724)	0.4062	0.5726	0.6313	0.6769	0.7803	0.0093	0.0761	70.64%	99.10%
OLS (2)	0.6233 (0.0778)	0.3777	0.5693	0.6342	0.6832	0.7738	0.0085	0.0726	74.85%	99.25%
CLAD (1)	0.6143 (0.0749)	0.3891	0.5625	0.6220	0.6699	0.7778	0.0094	0.0754	70.78%	98.95%
CLAD (2)	0.6080 (0.0730)	0.3570	0.5590	0.6171	0.6629	0.7565	0.0091	0.0717	73.04%	98.80%
GLM (1)	0.6235 (0.0727)	0.4327	0.5684	0.6273	0.6760	.8011	0.0093	0.0757	71.08%	98.95%
GLM (2)	0.6232 (0.0779)	0.3959	0.5669	0.6310	0.6814	0.7848	0.0085	0.0722	74.55%	99.25%
MLOGIT										
MLOGIT (1)	0.6179 (0.0701)	0.4280	0.5661	0.6220	0.6687	0.7742	0.0093	0.0752	71.99%	98.80%
MLOGIT (2)	0.6190 (0.0769)	0.4067	0.5670	0.6224	0.6742	0.7923	0.0081	0.0695	76.96%	99.10%

Dependent variable for OLS, Tobit, Fractional Logistic and CLAD was EQ-5D utility score. Dependent variables for MLOGIT models were EQ-5D dimension scores.
Independent variable(s): Model (1) RMQ score, Model (2) RMQ questions entered as categorical variables. Age and gender variables present in all models.
GLM model: Gamma distribution and log link

Table 5: Predicted probability of response levels for each EQ-5D dimension (median and 95% credible intervals). Probabilities are generated from the response mapping model fitted to baseline observations using Bayesian MCMC, and are given for females aged 53.6 (mean age of BEST participants).

RMQ score	Mobility			Self Care			Usual activities			Pain / Discomfort			Anxiety / Depression		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	0.98 (0.95,0.99)	0.02 (0.01,0.05)	0 (0,0.01)	1 (1,1)	0 (0,0)	0 (0,0)	0.91 (0.83,0.96)	0.09 (0.04,0.17)	0 (0,0)	0.15 (0.06,0.31)	0.85 (0.68,0.94)	0 (0,0)	0.9 (0.84,0.94)	0.1 (0.06,0.16)	0 (0,0)
3	0.85 (0.79,0.89)	0.15 (0.1,0.21)	0 (0,0)	0.99 (0.99,1)	0.01 (0,0.01)	0 (0,0)	0.6 (0.5,0.69)	0.4 (0.31,0.49)	0 (0,0)	0.05 (0.03,0.09)	0.94 (0.91,0.97)	0 (0,0.01)	0.75 (0.69,0.81)	0.25 (0.19,0.31)	0 (0,0)
5	0.66 (0.6,0.73)	0.33 (0.27,0.4)	0 (0,0)	0.98 (0.96,0.99)	0.02 (0.01,0.04)	0 (0,0)	0.35 (0.3,0.41)	0.64 (0.58,0.7)	0 (0,0.01)	0.03 (0.02,0.05)	0.95 (0.93,0.97)	0.02 (0.01,0.03)	0.63 (0.58,0.69)	0.36 (0.31,0.42)	0 (0,0.01)
7	0.49 (0.43,0.54)	0.51 (0.46,0.57)	0 (0,0)	0.95 (0.93,0.97)	0.05 (0.03,0.07)	0 (0,0)	0.21 (0.17,0.25)	0.78 (0.74,0.82)	0.01 (0,0.02)	0.02 (0.01,0.03)	0.94 (0.91,0.96)	0.04 (0.03,0.07)	0.54 (0.49,0.58)	0.46 (0.41,0.51)	0.01 (0,0.01)
9	0.34 (0.28,0.39)	0.66 (0.61,0.72)	0 (0,0)	0.9 (0.87,0.93)	0.09 (0.07,0.13)	0 (0,0)	0.13 (0.09,0.17)	0.86 (0.82,0.89)	0.02 (0.01,0.03)	0.01 (0.01,0.03)	0.89 (0.86,0.92)	0.09 (0.06,0.13)	0.45 (0.4,0.5)	0.54 (0.49,0.58)	0.01 (0.01,0.03)
11	0.22 (0.17,0.28)	0.78 (0.72,0.83)	0 (0,0)	0.83 (0.79,0.87)	0.17 (0.13,0.21)	0 (0,0.01)	0.08 (0.05,0.11)	0.89 (0.86,0.92)	0.03 (0.01,0.05)	0.01 (0,0.02)	0.82 (0.77,0.86)	0.17 (0.13,0.22)	0.38 (0.33,0.42)	0.6 (0.55,0.65)	0.03 (0.01,0.05)
13	0.14 (0.10,0.19)	0.86 (0.81,0.9)	0 (0,0)	0.72 (0.66,0.77)	0.28 (0.23,0.34)	0 (0,0.01)	0.05 (0.03,0.07)	0.9 (0.87,0.93)	0.05 (0.03,0.08)	0.01 (0,0.02)	0.7 (0.64,0.76)	0.29 (0.24,0.35)	0.31 (0.26,0.36)	0.64 (0.59,0.69)	0.05 (0.03,0.08)
15	0.09 (0.05,0.13)	0.91 (0.87,0.95)	0 (0,0)	0.57 (0.49,0.64)	0.42 (0.35,0.5)	0.01 (0,0.02)	0.03 (0.01,0.05)	0.89 (0.85,0.92)	0.08 (0.05,0.12)	0.01 (0,0.02)	0.54 (0.45,0.61)	0.46 (0.38,0.54)	0.25 (0.19,0.3)	0.66 (0.6,0.72)	0.09 (0.06,0.13)
17	0.05 (0.03,0.08)	0.95 (0.92,0.97)	0 (0,0.01)	0.39 (0.3,0.49)	0.59 (0.5,0.69)	0.01 (0,0.04)	0.01 (0.01,0.03)	0.84 (0.78,0.89)	0.14 (0.09,0.21)	0 (0,0.01)	0.35 (0.25,0.45)	0.65 (0.55,0.74)	0.19 (0.14,0.25)	0.65 (0.56,0.72)	0.16 (0.11,0.24)
19	0.02 (0.01,0.05)	0.98 (0.95,0.99)	0 (0,0.01)	0.21 (0.13,0.32)	0.75 (0.64,0.84)	0.03 (0.01,0.09)	0.01 (0,0.02)	0.74 (0.61,0.84)	0.25 (0.15,0.38)	0 (0,0.01)	0.18 (0.1,0.27)	0.82 (0.72,0.9)	0.13 (0.09,0.19)	0.57 (0.43,0.68)	0.3 (0.19,0.44)
21	0.01 (0.0,0.02)	0.99 (0.96,1)	0 (0,0.03)	0.08 (0.04,0.15)	0.83 (0.61,0.93)	0.09 (0.01,0.31)	0 (0,0.01)	0.53 (0.32,0.73)	0.46 (0.27,0.68)	0 (0,0.01)	0.06 (0.02,0.12)	0.94 (0.88,0.97)	0.08 (0.05,0.13)	0.38 (0.19,0.58)	0.54 (0.34,0.73)
23	0 (0,0)	1 (0.61,1)	0 (0,0.39)	0.01 (0,0.03)	0.6 (0.06,0.97)	0.39 (0.02,0.93)	0 (0,0)	0.15 (0.04,0.42)	0.84 (0.58,0.96)	0 (0,0)	0.01 (0,0.02)	0.99 (0.98,1)	0.03 (0.01,0.06)	0.09 (0.02,0.3)	0.88 (0.67,0.96)

Table: 6 Model performance using repeated BeST measurements- EQ-5D

	Predicted EQ-5D Values								Abs Diff	Abs Diff
Model	**Mean (SD)**	**Min**	**P.25**	**Median**	**P.75**	**Max**	**MSE**	**MAE**	**<0.10 (%)**	**<0.25 (%)**
Observed EQ-5D	0.5997 (0.2835)	-0.4840	0.5870	0.6910	0.7600	1.000	-	-	-	-
OLS Robust Cluster	0.6004 (0.1883)	-0.0167	0.4769	0.6340	0.7528	0.8855	0.0448	0.1575	43.80%	79.84%
Random Intercept	0.6003 (0.2126)	-0.1702	0.4943	0.6567	0.7548	0.9864	0.0235	0.1093	60.54%	88.34%
Random Coefficient	0.6004 (0.2186)	-0.2861	0.5246	0.6726	0.7529	0.8479	0.0231	0.1035	64.56%	89.44%
Dependent variable for OLS Robust Cluster, Random Intercept and Random Coefficient models: EQ-5D utility score. Independent variables: RMQ score, age and gender.										

Table 7: Model Performance using repeated BeST measurements- SF6D

	Predicted EQ-5D Values								Abs Diff	Abs Diff
Model	**Mean (SD)**	**Min**	**P.25**	**Median**	**P.75**	**Max**	**MSE**	**MAE**	**<0.10 (%)**	**<0.25 (%)**
Observed SF6D	0.6572 (0.1340)	0.3450	0.5650	0.6360	0.7400	1.000	-	-	-	-
OLS Robust Cluster	0.6575 (0.0893)	0.3566	0.5975	0.6737	0.7301	0.8075	0.0097	0.0803	68.48%	99.25%
Random Intercept	0.6579 (0.1058)	0.3518	0.5819	0.6548	0.7330	0.9159	0.0040	0.0503	88.40%	100%
Random Coefficient	0.6578 (0.1069)	0.3670	0.5804	0.6485	0.7317	0.9286	0.0037	0.0480	89.90%	100%
Dependent variable for OLS Robust Cluster, Random Intercept and Random Coefficient models: EQ-5D utility score. Independent variables: RMQ score, age and gender.										

Table 8: Prediction errors from response mapping models fitted using Bayesian MCMC

Predicted variable	Mean absolute error of prediction											
	Baseline			3 months			6 months			12 months		
	Baseline model	Multilevel model	Transition model	Baseline model	Multilevel model	Transition model	Baseline model	Multilevel model	Transition model	Baseline model	Multilevel model	Transition model
Utility	0.152	0.151	0.153	0.134	0.136	0.150	0.149	0.152	0.154	0.142	0.143	0.166
Mobility	0.328	0.298	0.325	0.287	0.262	0.298	0.280	0.255	0.268	0.269	0.241	0.279
Self-Care	0.221	0.206	0.215	0.185	0.172	0.177	0.182	0.168	0.162	0.176	0.164	0.159
Usual activities	0.312	0.301	0.310	0.350	0.330	0.378	0.317	0.295	0.326	0.320	0.296	0.336
Pain	0.231	0.223	0.230	0.230	0.234	0.251	0.270	0.277	0.259	0.271	0.271	0.287
Anxiety / Depression	0.473	0.462	0.471	0.433	0.417	0.384	0.428	0.410	0.357	0.438	0.421	0.378

Table 9: Predicted utility (mean and 95% confidence / credible interval) for a given RMQ score

RMQ score	No. Obs	Actual (both genders, baseline)	OLS Model	FLOGIT Model	Bayesian FLOGIT Model	CLAD Model	Baseline Response Mapping (Bayesian)	OLS robust cluster	Random Coefficient	Random Coefficient Response Mapping (Bayesian)
1	14	0.749 (0.716,0.783)	0.823 (0.787, 0.859)	0.789 (0.721, 0.857)	0.816 (0.792,0.838)	0.788 (0.768,0.807)	0.810 (0.792 , 0.843)	0.820 (0.803, 0.837)	0.790 (0.774, 0.805)	0.848 (0.827 , 0.873)
2	26	0.752 (0.727,0.777)	0.789 (0.755, 0.822)	0.767 (0.699, 0.835)	0.795 (0.771,0.818)	0.768 (0.750, 0.786)	0.780 (0.769 , 0.793)	0.785 (0.768, 0.801)	0.760 (0.745, 0.774)	0.802 (0.793 , 0.813)
3	61	0.735 (0.717,0.754)	0.754 (0.724, 0.785)	0.744 (0.676, 0.811)	0.773 (0.748,0.796)	0.749 (0.732, 0.765)	0.757 (0.748 , 0.767)	0.749 (0.734, 0.764)	0.730 (0.716, 0.743)	0.776 (0.769 , 0.784)
4	45	0.719 (0.69,0.749)	0.720 (0.692, 0.749)	0.718 (0.651, 0.785)	0.748 (0.724,0.771)	0.729 (0.714, 0.744)	0.736 (0.727 , 0.745)	0.714 (0.699, 0.728)	0.700 (0.687, 0.713)	0.752 (0.744 , 0.76)
5	51	0.707 (0.661,0.753)	0.686 (0.660, 0.712)	0.691 (0.625, 0.756)	0.722 (0.698,0.745)	0.710 (0.695, 0.724)	0.716 (0.706 , 0.725)	0.678 (0.664, 0.692)	0.670 (0.657, 0.683)	0.728 (0.720 , 0.737)
6	61	0.661 (0.624,0.697)	0.652 (0.627, 0.676)	0.662 (0.597, 0.726)	0.693 (0.669,0.716)	0.690 (0.677, 0.703)	0.695 (0.684 , 0.705)	0.642 (0.628, 0.657)	0.640 (0.627, 0.653)	0.704 (0.695 , 0.713)
7	64	0.651 (0.599,0.703)	0.618 (0.595, 0.641)	0.631 (0.567, 0.694)	0.663 (0.638,0.686)	0.671 (0.658, 0.683)	0.673 (0.66 , 0.685)	0.607 (0.592, 0.622)	0.610 (0.596, 0.624)	0.681 (0.672 , 0.69)
8	34	0.639 (0.58,0.697)	0.584 (0.561, 0.606)	0.598 (0.535, 0.661)	0.630 (0.605,0.654)	0.651 (0.639, 0.663)	0.650 (0.635 , 0.663)	0.571 (0.556, 0.587)	0.580 (0.565, 0.595)	0.660 (0.65 , 0.669)
9	49	0.516 (0.443,0.59)	0.550 (0.528, 0.571)	0.564 (0.500, 0.628)	0.595 (0.57,0.62)	0.632 (0.620, 0.643)	0.624 (0.607 , 0.64)	0.536 (0.519, 0.553)	0.550 (0.534, 0.567)	0.638 (0.627 , 0.648)
10	45	0.573 (0.495,0.65)	0.515	0.528	0.559 (0.532,0.585)	0.612	0.595	0.500	0.520	0.614

			(0.493, 0.537)	(0.463, 0.594)		(0.600, 0.624)	(0.576 , 0.613)	(0.482, 0.519)	(0.503, 0.538)	(0.601 , 0.627)
11	44	0.505 (0.422,0.588)	0.481 (0.459, 0.504)	0.491 (0.423, 0.560)	0.521 (0.492,0.549)	0.593 (0.580, 0.605)	0.562 (0.54 , 0.582)	0.465 (0.445, 0.485)	0.490 (0.471, 0.510)	0.586 (0.570 , 0.602)
12	36	0.439 (0.335,0.542)	0.447 (0.423, 0.471)	0.453 (0.380 ,0.526)	0.481 (0.449,0.512)	0.573 (0.560, 0.586)	0.524 (0.499 , 0.547)	0.429 (0.408, 0.451)	0.461 (0.440, 0.481)	0.551 (0.531 , 0.571)
13	37	0.412 (0.32,0.504)	0.413 (0.387, 0.438)	0.414 (0.335, 0.494)	0.440 (0.405,0.474)	0.554 (0.540, 0.567)	0.480 (0.453 , 0.506)	0.394 (0.370, 0.417)	0.431 (0.408, 0.453)	0.507 (0.480 , 0.532)
14	29	0.385 (0.270,0.500)	0.379 (0.351, 0.406)	0.375 (0.289, 0.461)	0.398 (0.36,0.436)	0.534 (0.519, 0.549)	0.431 (0.400 , 0.461)	0.358 (0.333,0.384)	0.401 (0.376, 0.425)	0.451 (0.419 , 0.483)
15	24	0.376 (0.247,0.506)	0.345 (0.315, 0.374)	0.335 (0.241, 0.429)	0.355 (0.313,0.397)	0.515 (0.499, 0.531)	0.377 (0.343 , 0.412)	0.323 (0.296, 0.350)	0.371 (0.344, 0.397)	0.386 (0.347 , 0.423)
16	20	0.368 (0.213,0.522)	0.310 (0.278, 0.343)	0.295 (0.193, 0.397)	0.312 (0.266,0.358)	0.495 (0.478, 0.512)	0.321 (0.283 , 0.358)	0.287 (0.258, 0.317)	0.341 (0.312, 0.369)	0.313 (0.272 , 0.355)
17	11	0.397 (0.213,0.58)	0.276 (0.241, 0.311)	0.255 (0.146, 0.365)	0.269 (0.22,0.319)	0.476 (0.457, 0.494)	0.263 (0.225 , 0.304)	0.252 (0.220, 0.283)	0.311 (0.280, 0.342)	0.241 (0.201 , 0.283)
18	15	0.131 (-0.018,0.279)	0.242 (0.205, 0.280)	0.216 (0.099, 0.333)	0.226 (0.174,0.279)	0.456 (0.436, 0.476)	0.207 (0.172 , 0.248)	0.216 (0.183, 0.250)	0.281 (0.248, 0.314)	0.174 (0.140, 0.213)
19	7	0.117 (-0.017,0.25)	0.208 (0.167, 0.249)	0.178 (0.054, 0.301)	0.184 (0.128,0.24)	0.437 (0.415, 0.459)	0.156 (0.125 , 0.193)	0.181 (0.145, 0.217)	0.251 (0.216, 0.286)	0.118 (0.088 , 0.15)
20	3	0.109 (-0.492,0.71)	0.174 (0.130, 0.217)	0.140 (0.011, 0.269)	0.142 (0.084,0.202)	0.417 (0.394, 0.441)	0.111 (0.083 , 0.143)	0.145 (0.107, 0.183)	0.221 (0.184, 0.258)	0.072 (0.043 , 0.100)
21	6	0.041 (-0.186,0.268)	0.140 (0.093, 0.186)	0.104 (-0.030, 0.238)	0.102 (0.042,0.164)	0.398 (0.372, 0.423)	0.072 (0.047 , 0.100)	0.110 (0.070, 0.150)	0.191 (0.152, 0.230)	0.031 (-0.007 , 0.060)
22	4	0.047 (-0.274,0.367)	0.106 (0.056, 0.155)	0.069 (-0.069, 0.206)	0.063 (0.002,0.127)	0.378 (0.351, 0.405)	0.037 (0.014 , 0.064)	0.074 (0.032, 0.117)	0.161 (0.120, 0.202)	-0.015 (-0.092 , 0.022)
23	3	-0.09 (-0.185,0.004)	0.071 (0.019, 0.124)	0.035 (-0.104, 0.175)	0.026 (-0.036,0.091)	0.359 (0.330, 0.387)	0.007 (-0.058 , 0.029)	0.039 (-0.006, 0.083)	0.131 (0.088, 0.175)	-0.107 (-0.175 , -0.022)
24	0	NA	0.037 (-0.019, 0.093)	0.004 (-0.137, 0.145)	-0.010 (-0.071,0.056)	0.339 (0.309, 0.369)	-0.015 (-0.191,-0.009)	0.003 (-0.044, 0.050)	0.101 (0.056, 0.147)	-0.191 (-0.191,-0.186)

Table 10: Impact of reducing RMQ score from 9 to 7 (the mean BeST study treatment effect) as estimated by each Bayesian response mapping model.

		Change in probability of response associated with a change in RMQ score from 9 to 7 (Median and 95% credible interval)		
Dimension	Level	Model using baseline data	Multilevel model using all data	Transition model
Mobility	1	0.15 (0.13,0.18)	0.22 (0.19,0.26)	0.24 (0.19,0.29)
	2	-0.15 (-0.18,-0.13)	-0.22 (-0.26,-0.19)	-0.24 (-0.29,-0.19)
	3	0 .00(0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)
Self-Care	1	0.05 (0.04,0.06)	0.03 (0.02,0.04)	0.03 (0.01,0.06)
	2	-0.05 (-0.06,-0.04)	-0.03 (-0.04,-0.02)	-0.03 (-0.06,-0.01)
	3	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.01)
Usual Activities	1	0.08 (0.07,0.10)	0.09 (0.07,0.11)	0.33 (0.28,0.38)
	2	-0.07 (-0.09,-0.06)	-0.09 (-0.11,-0.07)	-0.32 (-0.37,-0.27)
	3	-0.01 (-0.01,0.00)	0.00 (-0.01,0.00)	-0.01 (-0.02,0.00)
Pain	1	0.01 (0.00,0.01)	0.00 (0.00,0.01)	0.07 (0.04,0.10)
	2	0.04 (0.03,0.06)	0.02 (0.01,0.04)	-0.02 (-0.06,0.02)
	3	-0.05 (-0.06,-0.04)	-0.03 (-0.04,-0.02)	-0.05 (-0.08,-0.02)
Anxiety/Depression	1	0.08 (0.07,0.10)	0.13 (0.11,0.16)	0.16 (0.12,0.20)
	2	-0.08 (-0.10,-0.06)	-0.12 (-0.15,-0.10)	-0.15 (-0.2,-0.10)
	3	-0.01 (-0.01,0.00)	0.00 (-0.01,0.00)	-0.01 (-0.03,0.00)
Associated Utility change		0.051 (0.045,0.057)	0.046 (0.041,0.052)	0.078 (0.062,0.094)

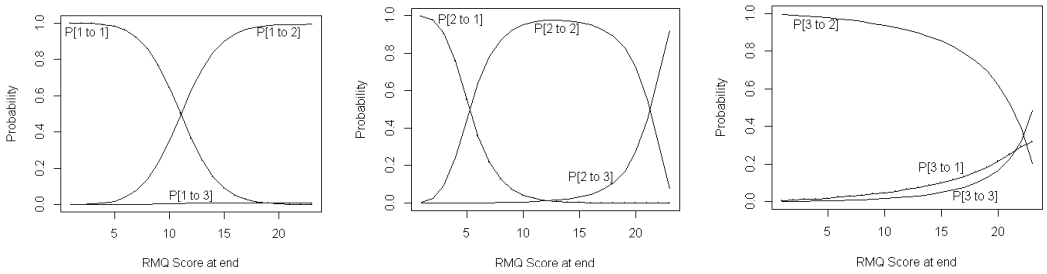


Fig 3.1 Mobility

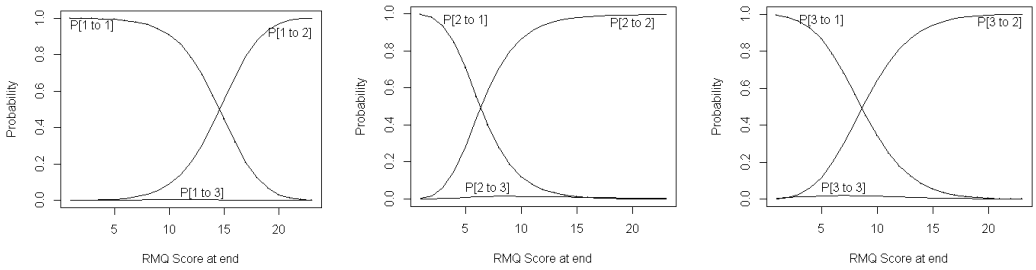


Fig 3.2: Self care

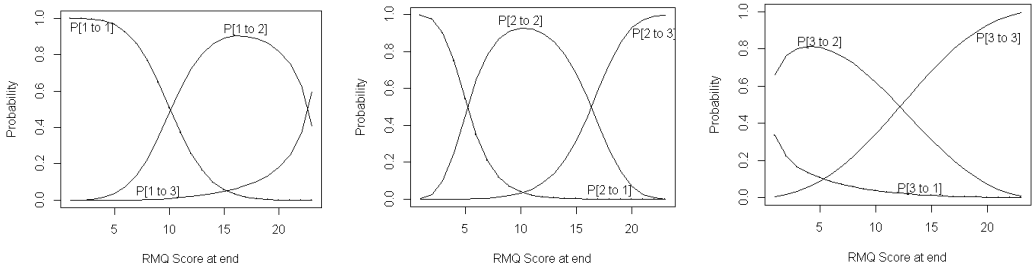


Fig 3.3 Usual activities

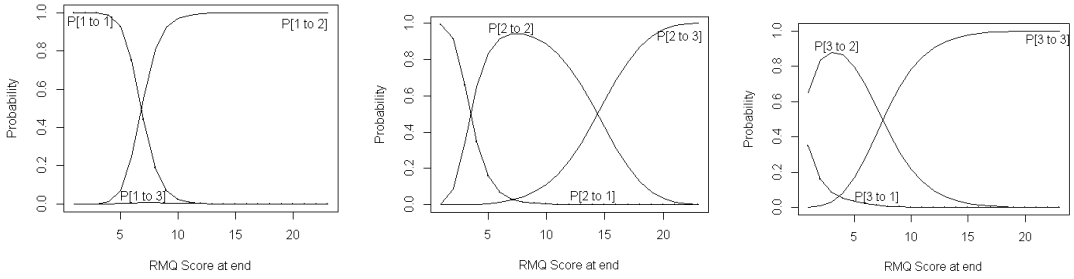


Fig 3.4 Pain

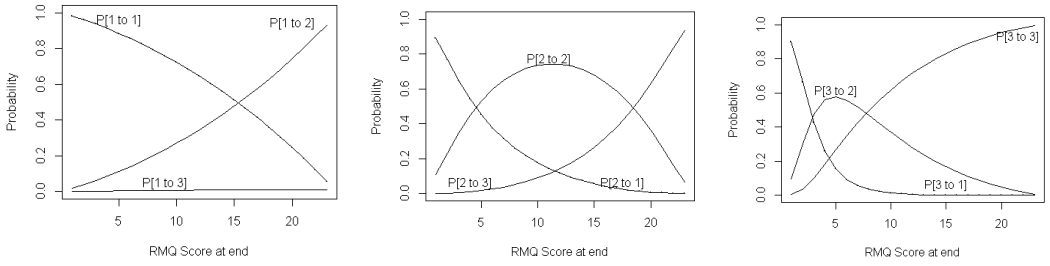


Fig 3.5 Anxiety / Depression

Figure 2: Relationship between change in RMQ and change in EQ-5D item response. Probabilities are calculated assuming a starting RMQ of 9.