Can mapping algorithms based on raw scores overestimate QALYs gained by treatment? A comparison of mappings between the Roland-Morris Disability Questionnaire and the EQ-5D-3L based on raw and differenced score data

Jason Madan[1], PhD, Kamran A. Khan[1], MSc, Stavros Petrou[1], PhD, Sarah E. Lamb[2], DPhil

1   Warwick Clinical Trials Unit, Division of Health Sciences,

        Warwick Medical School, University of Warwick,

        Coventry CV4 7AL, UK

2   Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences

        Medical Science Division, University of Oxford

        Oxford OX3 7HE

Correspondence:

Jason Madan

Warwick Clinical Trials Unit, Division of Health Sciences,

Warwick Medical School, University of Warwick,

Coventry CV4 7AL, UK

Email: j.j.madan@warwick.ac.uk

Telephone: +44 (0)2476 151254

Running head: Mapping using raw and differenced data compared.

Abstract:

Introduction

Mapping algorithms are increasingly being used to predict health utility values based on responses or scores for non-preference based measures, thereby informing economic evaluations. We explored whether predictions in EQ-5D-3L health utility gains from mapping algorithms might differ if estimated using differenced versus raw scores, using the Roland Morris Disability Questionnaire (RMQ), a widely-used health status measure for low back pain, as an example.

Methods

We estimated algorithms mapping within-person changes in RMQ scores to changes in EQ-5D-3L health utilities using data from two clinical trials with repeated observations. We also estimated response mapping algorithms from these data to predict within-person changes in responses to each EQ-5D-3L dimension from changes in RMQ scores, using logistic regression models. Predicted health utility gains from these mappings were compared with predictions based on raw RMQ data.

Results

Using differenced scores reduced the predicted health utility gain from a unit decrease in RMQ score from 0.037 (standard error (SE) 0.001) to 0.020 (SE 0.002). Analysis of response mapping data suggests that use of differenced data reduces the predicted impact of reducing RMQ scores across EQ-5D-3L dimensions, and that patients can experience health utility gains on the EQ-5D-3L 'usual activity' dimension independent from improvements captured by the RMQ.

Conclusion

Mappings based on raw RMQ data overestimate the EQ-5D-3L health utility gains from interventions that reduce RMQ scores. Where possible, mapping algorithms should reflect within-person changes in health outcome and be estimated from datasets containing repeated observations if they are to be used to estimate incremental health utility gains.

Key Points for Decision Makers

The QALYs associated with changes in a non-preference based outcome measure can vary substantially if estimated using mapping algorithms based on differenced, rather than raw, scores.

Mappings should be estimated using differenced scores from repeated observations if they are to be used to estimate treatment-related incremental QALYs, to avoid the impact of confounders unrelated to treatment.

1. Introduction

Cost-utility analysis requires health outcomes to be measured on preference-based utility scales that reflect values assigned to all possible health states, including perfect health and death. It remains the preferred form of economic evaluation by public bodies such as the National Institute for Health and Care Excellence NICE in England and Wales (1), as it provides a single preference-based scale for assessing diverse health outcomes, and allows cost-effectiveness comparisons to made across clinical specialties. Many studies collect data on non-preference based measures of health-related quality of life or clinical symptoms, without also collecting data on preference-based outcome measures such as the EQ-5D-3L or SF-6D. Results from these studies can still inform cost-utility analyses if mapping algorithms are available to predict health utility values based on responses or scores for non-preference based measures. These mapping algorithms are typically constructed from datasets in which participants simultaneously report outcomes for preference-based and non-preference-based measures.(2) Such datasets are commonly cross-sectional, but examples exist where mappings have been derived from datasets with repeated observations on participants.(3, 4) This can increase the precision with which mapping coefficients are estimated, as long as the statistical methods used account for correlations between observations from the same participant.

The Roland Morris Disability Questionnaire (RMQ) is a commonly-used non preference based outcome measure for low back pain.(5) It consists of 24 items relating to a range of functions commonly affected by low back pain and disability, each with binary 'yes'/'no' options. The total number of positive responses is summed to form a score (from 0 to 24), and a low score is associated with less disability. We have previously developed mapping algorithms translating RMQ scores into EQ-5D-3L and SF-6D response and utility scores,

4

based on repeated observations from two randomised clinical trials, using a range of regression models to account for the properties of the distribution of utility scores and the relationships between repeated observations.(6) In this paper, we present an alternative approach for estimating mapping algorithms between the RMQ and the EQ-5D-3L, using individual patient differenced RMQ scores. We use the term 'differenced' to refer to the change in RMQ score, utility or EQ-5D-3L response between two observations on the same individual at two different time points. We compare the direct and response mappings derived from these data to mappings constructed using raw data (in this context, we refer to raw data as data that has not been processed or manipulated in any way, such as differencing), and explore the implications of our findings for economic evaluations of low back pain treatments and the construction of mapping algorithms based on non-preference based outcome measures in other clinical areas.

2. Data and methods

2.1 Data used for mapping estimation.

We previously used data from the Back Skills Training (BeST) trial (7) to develop a range of mapping algorithms from the RMQ to utility scores derived from the EQ-5D-3L.(6) BeST was a pragmatic, multicentre, randomised controlled trial of a cognitive behavioural intervention for low back pain combined with active management, compared with active management alone, which 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 algorithms we developed involved both direct mapping (in which RMQ scores were mapped to utility scores derived from EQ5D responses using UK tariffs)

and response mapping (in which RMQ scores were mapped to the actual responses to the EQ-5D-3L questionnaire), and versions were developed based on the total RMQ score as well as responses to individual RMQ items. Models were validated using data from the Back Pain Exercise and Manipulation (UK BEAM) trial.(8) The UK BEAM trial recruited and randomised 1334 participants presenting in primary care with low back pain to one of four interventions: manipulation, exercise, manipulation combined with exercise or best care in general practice. Among other criteria, individuals were eligible for inclusion if they were aged between 18 and 65 and had a score of four or more on the RMQ on the day of randomisation. Both the BeST and UK BEAM datasets included repeated observations of all outcome measures; the BeST trial at randomisation, and 3, 6 and 12 months post-randomisation, and the UK BEAM trial at randomisation, and 1, 3 and 12 months post-randomisation. Tables 1 and 2 summarise the data available on within-person changes to RMQ and EQ-5D-3L utility scores over the follow-up periods in the BEST and UK BEAM trials, respectively. In our previous work, we incorporated these repeated observations using robust standard error and hierarchical (random intercept and random coefficient) regression models; further details on these models are given in Khan et al. (6)

2.2 Description of mapping models

We extended our previous work by initially fitting an Ordinary Least Squares (OLS) regression model to the BeST dataset with change in EQ-5D-3L utility score as the dependent variable and change in RMQ score as the main explanatory variable. As in our previous work,(6) we included age and sex as covariates, and included all participants irrespective of trial arm allocation. The regression is therefore given by equation 1:

$$\Delta_{ij}^{U} = \alpha_j + \beta_{1j}\Delta_{ij}^{RMQ} + \beta_{2j}y_i + \beta_{3j}s_i + \beta_{4j}RMQ_{ij}^{B} + \beta_{4j}RMQ_{ij}^{B} * \Delta_{ij}^{RMQ} + \varepsilon_{ij} \qquad [1]$$

Here Δ_{ij}^{U} is the change in health utility score experienced by participant i during time interval j; Δ_{ij}^{RMQ} is the change in the RMQ score of participant i during that interval; y_i and s_i are the age and sex, respectively, of trial participant i, and RMQ_{ij}^{B} is the RMQ score at the start of interval j. $RMQ_{ij}^{B} * \Delta_{ij}^{RMQ}$ is a term allowing for interaction between RMQ_{ij}^{B} and Δ_{ij}^{RMQ}, and ε_{ij} is a normally-distributed random variable with mean zero. The time interval j refers to the time between successive observations, rather than the time from baseline to a given observation. This is to ensure that there is no overlap between time intervals, as such overlaps would induce correlations between differenced observations. Were no missing data present, there would be three time intervals, reflecting the follow-up intervals in the BeST trial: 0-3 months, 3-6 months, and 6-12 months (where 0 denotes the point of trial randomisation). Due to missing observations for some participants, there were three additional intervals in the differenced dataset: 0-6 months, 0-12 months and 3-12 months. Equation 1 therefore defines 6 independent regression models. We also fitted a seventh model in which the regression coefficients were assumed to be equal for each interval, so that we could test the assumption that the mapping relationship was stable over time during the follow-up period of the BeST trial. We used data from the UK BEAM trial for external validation of the models, and drew comparisons of their predictive power with models based on raw values rather than differenced scores.

Using the data from the BeST trial, we then fitted a range of response mapping models (9, 10) for each dimension of the EQ-5D-3L, based on differenced data. These models estimate the probability that an individual will report level 1,2 or 3 at a particular follow-up time point for each dimension of the EQ-5D, conditional on their response for that dimension at their most recent previous observation, and the difference between the RMQ scores reported at the current and most recent previous observation. This can be thought of as fitting five state-transition models, one for each EQ-5D-3L dimension, in which the states are the three

7

possible EQ-5D-3L dimension levels that can be reported. Multinomial regression models were used to estimate these probabilities, as a function of the change in RMQ score between observations, and any other variables that were found to be significant predictors, at the 5% significance level (this level is used to define significance throughout the paper), of change in utility score in the OLS regression (equation 1). These models are given by equation 2:

$$p_{ijk}^{ab} = \frac{\exp\left(\alpha_{jk}^{ab} + \beta_{jk}^{ab} x_{ij} + \varepsilon^{ab}{}_{ijk}\right)}{1 + \exp\left(\alpha_{jk}^{a2} + \beta_{jk}^{a2} x_{ij} + \varepsilon^{a2}{}_{ijk}\right) + \exp\left(\alpha_{jk}^{a3} + \beta_{jk}^{a3} x_{ij} + \varepsilon^{a3}{}_{ijk}\right)}; \alpha_{jk}^{a1} = \beta_{jk}^{a1} = E[\varepsilon^{a1}{}_{ijk}] = 0 \forall a, i, j, k$$

[2]

Here p_{ijk}^{ab} is the probability that participant i will provide response level b to item k of the EQ-5D-3L questionnaire at the end of interval j if they provided response level a at the start of that interval. For example, $p_{100,2,3}^{2,1}$ is the probability that, if participant 100 reported level 2 on the Usual Activities dimension of the EQ-5D-3L the second time they completed the questionnaire, they would report level 1 on this dimension the third time they completed the questionnaire. x_{ij} is the vector of explanatory variables identified as significant predictors of change in utility score in the previous model, including the change in RMQ over the interval j. α_{jk}^{ab} is the intercept term for the logistic regression predicting p_{ijk}^{ab} from x_{ij}, β_{jk}^{ab} is a vector of regression coefficients for the predictors, and $\varepsilon^{ab}{}_{ijk}$ is a normally distributed error term with zero mean. The restriction $\alpha_{jk}^{a1} = \beta_{jk}^{a1} = \varepsilon^{a1}{}_{ijk} = 0 \forall a, i, j, k$ ensures that p_{ijk}^{a1}, p_{ijk}^{a2}, and p_{ijk}^{a3} sum to 1, which they must always do since the levels 1, 2 and 3 are the only possible responses.

We compared this response mapping to a response mapping model based on raw RMQ scores, as given by equation 3:

$$p^b_{ijk} = \frac{\exp\left(\alpha^b_k + \beta^b_k x_i + \varepsilon^b_{ijk}\right)}{1 + \exp\left(\alpha^2_k + \beta^2_k x_i + \varepsilon^2_{ijk}\right) + \exp\left(\alpha^3_k + \beta^3_k x_i + \varepsilon^3_{ijk}\right)} ; \alpha^1_k = \beta^1_k = E[\varepsilon^1_{ijk}] = 0 \forall i, j, k$$

[3]

Here p^b_{ijk} is the probability that participant i will provide response level b to item k of the EQ-5D-3L questionnaire at the start of interval j and x_i is the vector of explanatory variables, including the RMQ score at the start of interval j, and any patient characteristics included in equation 2.

2.3 Illustrative example

Given the differences in structure between the models described by equations 2 and 3, a direct comparison of model coefficients is difficult to interpret. To aid this comparison, we carried out an illustrative hypothetical exercise to contrast, for the two models, the predicted changes in response levels that would result from an RMQ change that might typically be observed in a population with lower back pain. For this exercise, we chose to base our hypothetical cohort on the BEST participant population at baseline (n= 675). We then compared the predictions from the raw and differenced models described above as to the dimension-level responses that would be observed at follow-up in two hypothetical scenarios:

Scenario 1: The cohort experienced no change in disability from their back pain between the initial and follow-up observations ($\Delta^{RMQ}=0$).

The prediction of dimension scores using the raw scores response mapping model (i.e. equation (3)) is straightforward, since equation (3) predicts that EQ-5D-3L responses at follow-up will be the same as at the initial observation. This is not true for the differenced scores response mapping model, because of the intercept terms in equation 2, and the coefficient associated with baseline RMQ. For example, consider an individual in the hypothetical cohort who reported level 2 for mobility at the initial observation, and an RMQ score of 6 at both initial and follow-up observation. Using the notation of equation 2, this implies that a=2, j=1, and k=1, and the probabilities that they will report level 1, 2 or 3 for mobility at follow-up are given by:

$$p[mobility=1]=\frac{1}{1+\exp\left(\alpha_{1,1}^{2,2}+6\beta_{1,1}^{2,2}[2]\right)+\exp\left(\alpha_1^{2,3}+6\beta_{1,1}^{2,3}[2]\right)}$$

$$p[mobility=2]=\frac{\exp\left(\alpha_{1,1}^{2,2}+6\beta_1^{2,2}[2]\right)}{1+\exp\left(\alpha_{1,1}^{2,2}+6\beta_{1,1}^{2,2}[2]\right)+\exp\left(\alpha_{1,1}^{2,3}+6\beta_{1,1}^{2,3}[2]\right)}$$

$$p[mobility=3]=\frac{\exp\left(\alpha_1^{2,3}+6\beta_1^{2,3}[2]\right)}{1+\exp\left(\alpha_{1,1}^{2,2}+6\beta_{1,1}^{2,2}[2]\right)+\exp\left(\alpha_{1,1}^{2,3}+6\beta_{1,1}^{2,3}[2]\right)}$$

Here we define the second element of the β_1^b vector as the coefficient for baseline RMQ, which is why it appears in the equations above.

Scenario 2: The cohort experienced a moderate improvement in disability from their back pain between the initial and follow-up observation ($\Delta^{RMQ}=-2$).

We chose a two point reduction in RMQ to represent a moderate improvement in back pain disability as this is close to the improvement seen in the BEST population (the mean change across all participants in the study over its duration was -2.28).

The impact of this improvement on the individual described in scenario 1, as predicted by the raw score response mapping model, is given by equation 3:

$$p[mobility = 1] = \frac{1}{1 + \exp\left(\alpha_1^2 - 2\beta_1^2[1]\right) + \exp\left(\alpha_1^3 - 2\beta_1^3[1]\right)}$$

$$p[mobility = 2] = \frac{\exp\left(\alpha_1^2 - 2\beta_1^2[1]\right)}{1 + \exp\left(\alpha_1^2 - 2\beta_1^2[1]\right) + \exp\left(\alpha_1^3 - 2\beta_1^3[1]\right)}$$

$$p[mobility = 3] = \frac{\exp\left(\alpha_1^3 - 2\beta_1^2[1]\right)}{1 + \exp\left(\alpha_1^2 - 2\beta_1^2[1]\right) + \exp\left(\alpha_1^3 - 2\beta_1^3[1]\right)}$$

Here we define the first element of the β_1^b vector as the coefficient for change in RMQ, which is why it appears in the equations above.

Using the differenced scores response mapping model would result in the following predicted probabilities:

$$p[mobility = 1] = \frac{1}{1 + \exp\left(\alpha_{1,1}^{2,2} - 2\beta_{1,1}^{2,2}[1] + 6\beta_{1,1}^{2,2}[2]\right) + \exp\left(\alpha_{1,1}^{2,3} - 2\beta_{1,1}^{2,3}[1] + 6\beta_{1,1}^{2,3}[2]\right)}$$

$$p[mobility = 2] = \frac{\exp\left(\alpha_{1,1}^{2,2} - 2\beta_{1,1}^{2,2}[1] + 6\beta_{1,1}^{2,2}[2]\right)}{1 + \exp\left(\alpha_{1,1}^{2} - 2\beta_{1,1}^{2,2}[1] + 6\beta_{1,1}^{2,2}[2]\right) + \exp\left(\alpha_{1,1}^{2,3} - 2\beta_{1,1}^{2,3}[1] + 6\beta_{1,1}^{2,3}[2]\right)}$$

$$p[mobility = 1] = \frac{\exp\left(\alpha_{1,1}^{2,3} - 2\beta_{1,1}^{2,3}[1] + 6\beta_{1,1}^{2,3}[2]\right)}{1 + \exp\left(\alpha_{1,1}^{2,2} - 2\beta_{1,1}^{2,2}[1] + 6\beta_{1,1}^{2,2}[2]\right) + \exp\left(\alpha_{1,1}^{2,3} - 2\beta_{1,1}^{2,3}[1] + 6\beta_{1,1}^{2,3}[2]\right)}$$

The sole difference from scenario 1 is the addition of the $-2\beta_{jk}^{ab}[1]$ terms to alter the transition probabilities in light of the RMQ reduction in scenario 2.

We followed the following steps to generate our illustrative comparison:

Step 1: Read the baseline RMQ score and response for the mobility dimension for the 1st participant in the BEST dataset

Step 2: Estimate the predicted probability for this individual of reporting each possible level for the mobility dimension at follow-up, using the differenced model, assuming RMQ is unchanged (scenario 1).

Step 3: Repeat steps 1 and 2 for each participant in the BEST dataset.

Step 4: Calculate the mean predicted probability for each level across all BEST participants. This is interpreted as the predicted proportion of the hypothetical cohort reporting each level at follow up.

Step 5: Repeat steps 1-4 for all other EQ-5D-3L dimensions

Step 6: Repeat steps 1-5 assuming a 2-point reduction in RMQ score (scenario 2)

Step 7: Repeat step 6 using the raw score response mapping model.

Step 8: Calculate the proportions of the hypothetical cohort reporting each level for each dimension of the EQ-5D-3L. These will also be the predicted proportions from the raw score mapping model at follow-up under scenario 1.

Following these steps, we were able to compare and contrast the predicted change in these proportions from either model in either scenario, and illustrate how the raw and differenced score models would yield different EQ-5D response predictions for a treatment that yielded an improvement in back pain disability resulting in a 2-point reduction in RMQ.

All statistical analyses described above were performed in R (version 3.0.1).

3. Results

3.1. Direct mapping models

Tables 1 and 2 provide information on the number of observations available in the BeST and UK BEAM datasets at the different follow-up points and also present the mean changes in RMQ and EQ-5D-3L utility scores over the alternative time intervals. There were 701 patients recruited to the BeST trial, and three follow-up time points, giving a maximum of 2103 possible data observations, of which 1476 (70.2%) were actually collected. A differenced score was only calculated for an interval if the individual had provided both RMQ and EQ-5D-3L data at both the start and the end of the interval. If an individual had failed to provide responses at 3 months for RMQ and/or EQ-5D, but had done so at 0, 6 and 12 months, they would contribute an observation for the 0-6 and 6-12 month intervals only. Table 2 gives the coefficients for OLS regression models predicting the change in utility score between successive observations of BeST trial participants. The first column in table 2 presents the results from fitting an OLS regression model using all 1476 observations. The predicted decrease (increase) in EQ-5D-3L health utilities from a 1-point increase (decrease)

in RMQ score from this model is 0.020 (p<0.01). Age and sex were not significantly associated with changes in EQ-5D-3L utility score, but the baseline RMQ co-efficient of -0.003 was significant (p=0.001) . This co-efficient, together with the intercept, determines the predicted change in utility if RMQ does not change between observations. The intercept gives the predicted utility score change if the RMQ score at the start of an interval is 9 (the baseline mean for the BEST dataset), and does not change. The intercept is negative, suggesting that utility declines if the RMQ score does not change, although the term is not significant. However, the coefficient for RMQ[B] is negative, implying that the more severe the condition initially (higher RMQ is equivalent to greater disability), the greater the utility loss associated with no improvement in back pain disability. The interaction term was not significant, suggesting that the change in utility per unit change in RMQ is independent of baseline RMQ.

Table 2 also includes results from fitting separate OLS regressions for each time interval. The slope coefficient of 0.20 derived from fitting the single OLS regression across time intervals lies within each of the slope coefficient confidence intervals obtained by fitting separate OLS regressions for each time interval, which is consistent with the assumption that the coefficient mapping changes in RMQ with changes in utility is stable over time. We also explored a reduced form of this model in which the specific time interval between observations was included as a six-level factor (taking values from 1 to 6); none of these levels were significantly associated with the change in EQ-5D-3L utility score. Values for other coefficients were broadly consistent across intervals, with differences that were consistent with sampling variation. The exception was the interaction term, which was statistically significant for the 3-6 month and 3-12 month intervals.

Figure 1 illustrates how the relationship between RMQ score change and EQ-5D-3L utility

score change predicted by the model described above, which involves regressing

differenced utility score on differenced RMQ score (the 'OLS differenced score model'),

compares with predictions from algorithms based on raw score data. In our previous work

we found that a beta regression multi-level model fitted to baseline and follow-up

observations was the strongest-performing model.(6) Figure 1 compares the results of this

model to the OLS model presented in table 1. The predicted impact of a unit change in RMQ

score on health utility is approximately 50% lower when based on differenced rather than

raw score data as in our previous work (0.019 vs 0.037). However, this is not a direct

comparison of mapping algorithms derived from changes in RMQ scores with mapping

algorithms derived from RMQ scores at a single time point, since the beta regression

algorithm draws on repeated observations and uses a different regression method to the

OLS intervals model. To allow for direct comparison, Figure 1 also includes predictions from

an OLS model fitted to baseline data alone. This shows that the adoption of beta rather than

linear regression and the inclusion of repeated observations have little impact on the

discrepancy, especially for low-to-moderate changes in RMQ score. Figure 1 further depicts

the actual relationship between differenced RMQ and utility observed in the BEST trial, from

which the improvement in fit from modelling differenced data directly can be clearly seen.

Table 2 presents the results from external validation of the three models illustrated in Figure

1, using separate data from the UK BEAM trial. Model fit is assessed by calculating root

mean square error (RMSE) for predicted changes in utility scores between time-points in the

UK BEAM trial. The OLS differenced scores model results in a 0.02 reduction in RMSE

compared with the raw score OLS model.

3.2 Response mapping models

Table 3 provides fitted values from the response mapping state transition models predicting

changes in reported levels in each EQ-5D-3Ldimension between observations. Based on the

results of fitting the direct mapping differenced model, only RMQ^{B} and Δ_{ij}^{RMQ} were included

as explanatory covariates for the response mapping model. Probabilities are not given for

transitions between levels one ('no problems') and three ('severe or extreme problems') as

there were too few such transitions (< 5) for their estimation. For the same reason,

probabilities are not given for transitions to or from level three for the EQ-5D-3L'mobility' and

'self care' dimensions. The results suggest that there is a non-trivial possibility of changing

level even when the RMQ score remains unchanged, the probability of which will depend on

this RMQ score, since it is included as a covariate. Table 3 presents these probabilities

assuming RMQ values of 9 (the mean at baseline in the BEST dataset), as well as 5 and 12

(the interquartile values in the BEST dataset). This probability is greatest for those reporting

level 1 on the EQ-5D-3L pain dimension, who have a 97% chance of reporting level 2 on that

dimension at the next observation if their RMQ score remains unchanged at 9. By contrast,

93% of those reporting level 1 on the EQ-5D-3L self-care dimension continue to report that

level if their RMQ score remains unchanged at 9.

The odds ratios presented in the final column of table 3 can be used to estimate the change

in probabilities of each transition associated with a given change in RMQ score. For

example, each 1-point increase in RMQ (i.e. worsening in back pain disability is associated

with a 20% increase in the odds of reporting some mobility problems (level 2) for someone

who had not reported any mobility problems before their RMQ increased. The impact of a

reduction in RMQ is equal in magnitude, but opposite in sign. Overall, an increase in RMQ

score (representing an increase in back pain related disability) is associated with an

increased change of reporting worsening health on all EQ-5D-3L dimensions. There is a

suggestion, however, that anxiety is less influenced by RMQ score changes than other EQ-

5D-3Ldimensions, as the mean odds ratios for this dimension are lower than for the others.

In order to determine whether specific EQ-5D-3L dimensions were driving the discrepancy illustrated in Figure 1, it is necessary to compare the predictions of this model from those from a standard response mapping model fitted to baseline (point of randomisation) raw score/level data only. Table 4 lists the coefficients for such a model. One obvious difference is that the model presented in table 4 does not permit changes in level when the RMQ score remains unchanged. Further comparison of the coefficients presented in tables 3 and 4 is of limited value, given the differences in structure between the two response mapping models. For this reason, we constructed our illustrative model, the results of which are presented in table 5. For the illustrative example, the raw score response mapping model predicts EQ-5D-3L utility scores to be more sensitive to RMQ score changes than the differenced score model for all dimensions, with the greatest discrepancy for anxiety/ depression and pain. Estimates are also provided of movements predicted by the differenced score response mapping model in EQ-5D-3L responses in the absence of any change in RMQ score. The predicted increase in the proportion reporting 'no problems' for usual activities, given no change in RMQ score, is 9%. The equivalent change in proportions for other EQ-5D-3L dimensions is 5% or less.

4. Discussion

The development and use of mapping algorithms has become increasingly common, particularly following the publication, in 2008, of NICE methods guidance endorsing the use of such algorithms when directly measured EQ-5D-3L utilities are unavailable.(3) Detailed methods guidance has recently been published on the development of mapping algorithms for use in economic evaluations.(11, 12) This guidance covers a wide range of issues such as the comparability of populations for estimation and validation, the inclusion of covariates, and the choice of statistical models. No guidance is provided, however, on the use of cross-

17

sectional versus longitudinal data, or the appropriate statistical techniques for analysing the latter. Our work is the first, to our knowledge, to compare mapping algorithms based on within-person changes in scores with those based on raw data. We also present a novel state-transition approach to response mapping between changes in a widely used clinical score for disability due to low back pain (13) and changes in EQ-5D-3L dimension levels.

4.1 Why using differenced data alters the mapping algorithm

We observed that using within-person changes in RMQ scores reduces the predicted EQ-5D-3L health utility decrement of a unit increase in RMQ score by approximately 50%. This suggests that factors, such as comorbidities, may independently increase the disability associated with low back pain in those with lower health utilities. Such factors will lead to potential over-estimation of the health utility benefits from treatments that alleviate low back pain by mappings based on raw data. For example, it may be that those with anxiety or depression unrelated to their back pain tend to experience greater functional impairment from low back pain. Alleviating this functional impairment may reduce raw RMQ scores without necessarily improving mental wellbeing. Examination of relationships at the level of EQ-5D-3L dimensions can provide insights into which of these reasons are most relevant in a specific example. Results reported in table 5 are consistent with the mechanism suggested above, but suggest that analysis of cross-sectional data over-estimates the impact of changing RMQ scores across all EQ-5D-3L dimensions. The differenced score response mapping model also predicts how EQ-5D-3L responses might change even if the RMQ score does not. For individuals, the predicted response at follow-up, conditional on RMQ remaining constant, may involve improvement or worsening, depending on the initial response. However, the net predicted effect for the BeST population is for minimal change in responses if the RMQ score does not change. The possible exception is 'usual activities', where the predicted increase in those reporting 'no problems' would be 9% if the RMQ score remained unchanged. This is consistent with NICE guidelines on back pain, which

recommend that patients are encouraged to '… continue with normal activities as far as possible'. (14)

4.2 Study limitations

While our example illustrates the value of basing mapping algorithms on changes in scores for source measures, it does have limitations. We were only able to include a limited number of covariates in our mapping algorithms (age and sex). We felt that the influence of these covariates on predicted utility was too weak to justify their inclusion in the final version presented here, although we accept that expert judgement has a role to play in this decision, and versions of the response mapping algorithms with these covariates included are available from the authors on request. It is possible that inclusion of additional covariates in the baseline mapping algorithm would eliminate some of the discrepancy with the differences-based algorithm.

We did not find strong evidence to suggest that the relationship between differenced RMQ and utility scores varied over the duration of follow-up in BEST (12 months), suggesting that the differenced model could be applied to RMQ data collected over any time interval up to 12 months. Analysis of additional data would allow us to further test this conclusion, and explore whether the relationship is stable for intervals longer than 12 months. It may be that this stability occurs because the changes in back pain symptoms at different times produce similar effects on health-related quality of life. In other conditions, the impact of clinical symptoms on the dimensions of health-related quality of life might vary over the life history of the illness, so that changes that appear of similar magnitude in a clinical measure might produce different utilities at different stages in the disease. Further work applying the differencing approach in other disease areas would allow this to be tested.

Our study included 1476 observations, which is greater than the median sample size (1167) in the mapping studies identified in the recent review by Dakin et al. of studies mapping between clinical or health-related quality of life measures and the EQ-5D.(3) However, response mapping models tend to need larger samples than direct utility mapping models for reliable estimation, since they contain more parameters. Our state-transition response mapping model has more parameters still. While we had sufficient observations to estimate most transitions, we could not estimate transitions between levels 1 and 3 for any of the dimensions. A larger sample size would have allowed us to be more definitive in our comparison with the baseline response mapping model, and provide more accurate parameter estimates for certain transitions such as those from level 3 in the usual activities dimension.

We have used OLS regression to develop our mapping algorithms, and the limitations of this model when predicting health utility data have been extensively documented in the literature.(2) While limitations such as ceiling effects and multi-modality are less prominent when modelling changes in health utility, it is still possible that a more sophisticated model would improve the accuracy of the change-in-scores model. It would also allow us to relax the assumption that a unit change is RMQ has the same implications for health utility independent from baseline RMQ. However, given the improvements in model accuracy seen in our previous work, it is very unlikely that such models would change the nature of our conclusions.

We validated our algorithm using an independent dataset generated by the UK BEAM trial. A comparison of the demographic characteristics of participants in the UK BEAM and BEST trial has been previously published [6]. The UK BEAM trial was chosen because of similarities in participant characteristics and setting with the BEST trial (patients with low

back pain presenting in a UK general practice setting, 12 month follow up, similar gender balance (BEST 60% vs UK BEAM 56%), similar rates of loss to follow up (BEST 28% vs UK BEAM 24% at 12 months)). The pragmatic nature of both studies, and the primary care setting, meant that both populations were broadly representative of the general lower back pain population. However, there were some differences in participant demographics – the mean age of UK BEAM participants was lower (43 vs 54 years), and while the median RMQ was identical in both studies (8), the minimum RMQ score in the UK BEAM trial was higher (4 vs 0). Also, the interventions in each trial were qualitatively different, as BEST involved a psychological therapy (CBT), whereas UK BEAM involved physical therapies (exercise and spinal manipulation). Despite these differences, UK BEAM provides useful validation for our analysis, although further validation with other datasets would provide additional reassurance.

4.3 Implications for future mapping studies

Despite these caveats, our results provide a useful illustration of the potential impact of using within-person differences between observations, rather than the raw scores, when developing mapping algorithms. The former gives the impact of a change in a non-preference based health-related quality of life or clinical measure on health utility for an individual, whilst the latter generates the predicted difference in health utilities, at given point in time, between two individuals with different clinical scores. These are clearly different processes, and the appropriate choice depends on the use to which the algorithm will be put. For economic evaluations aiming to inform decisions around the adoption of new treatments, it is often changes in health utility which are relevant to decision-makers. Our findings suggest that, in such cases, mapping algorithms should reflect within-person changes in health outcome and be developed using longitudinal data wherever possible.

Data Availability Statement

This study uses data from two published clinical trials. For access to these data, please contact the corresponding authors of the relevant publications. The models used to analyse these data and generate the results reported in this study are available from the corresponding author on request.

Compliance with Ethical Standards.

Author contributions

Study concept and design: Madan, Khan, Petrou; acquisition of data: Petrou, Lamb; all authors participated in analysis and interpretation of the data and preparation of the manuscript.

References

1. National Institute for Health and Care Excellence, (NICE). Guide to the methods of technology appraisal. National Institute for Health and Care Excellence (NICE). 2013.

2. Longworth L, Rowen D. NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values. 2011.

3. Dakin H. Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. Health and Quality of Life Outcomes. 2013; 11: 151.

4. Brazier J, 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. The European Journal of Health Economics. 2010; 11: 215-25.

5. Kopec JA. Measuring functional outcomes in persons with back pain: a review of back-specific questionnaires. Spine. 2000; 25: 3110-14.

6. Khan KA, Madan J, Petrou S, et al. Mapping between the Roland Morris Questionnaire and generic preference-based measures. Value in Health. 2014; 17: 686-95.

7. Lamb SE, Lall R Fau - Hansen Z, Hansen Z Fau - Castelnuovo 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.

8. Russell I, Underwood M, Brealey S, 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: 1381.

9. Gray AM, Rivero-Arias O, Clarke PM. Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping. Med Decis Making. 2006; 26: 18-29.

10. Rivero-Arias O, Ouellet M, Gray A, et al. Mapping the Modified Rankin Scale (mRS) Measurement into the Generic EuroQol (EQ-5D) Health Outcome. Med Decis Making. 2010; 30: 341-54.

11.   Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value in Health. 2013; 16: 202-10.

12.   Chuang L-H, Whitehead SJ. Mapping for economic evaluation. British medical bulletin. 2011; 101: 1-15.

13.   Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine. 2011; 36: S54-S68.

14.   Savigny P, Watson P, Underwood M. Early management of persistent non-specific low back pain: summary of NICE guidance. BMJ. 2009; 338.

Table 1: Summary of between-observation data from the BEST trial

Interval between follow-ups (months)	All Intervals	0*-3months	3-6months	6-12months	0*-6months	0*-12months	3-12months
Number of observations over this interval	1476	488	445	439	61	26	17
Mean (SD) Δ^{RMQ}	-0.743 (3.731)	-1.715 (4.022)	-0.245 (3.389)	-0.009 (3.195)	-1.852 (4.892)	-1.077 (4.214)	-0.294 (4.41)
Mean (SD) Δ^{U}	0.010 (0.243)	0.028 (0.257)	0.004 (0.236)	-0.003 (0.231)	0.038 (0.250)	-0.013 (0.229)	-0.041 (0.299)

*O denotes point of randomisation into the BeST trial. SD = standard deviation. Δ^{RMQ} = change in RMQ between observations. Δ^{U} = change in EQ-5D utility between observations.

Table 2: Summary of coefficients from OLS regression models predicting change in EQ-5D-3L health utility score between follow-up observations in the BEST trial, with results from validation using data from the UK BEAM trial.

Results from fitting models to BeST Data							
Interval between follow-ups	All intervals	0-3 months	3-6 months	6-12 months	0-6 months	0-12 months	3-12 months
Number of observations over this interval	1476	488	445	439	61	26	17
Intercept Coefficient (se)	-0.019 (0.026)	0.041 (0.048)	-0.034 (0.045)	-0.004 (0.046)	-0.098 (0.120)	-0.094 (0.281)	-0.284 (0.209)
Δ^{RMQ} Coefficient (se)	-0.020* (0.002)	-0.020 * (0.003)	-0.026* (0.003)	-0.020* (0.004)	-0.008 (0.007)	-0.008 (0.0132)	-0.034 (0.016)
Age Coefficient (se)	0.000 (0.000)	-0.001 (0.001)	0.000 (0.001)	0.000 (0.000)	0.002 (0.002)	0.001 (0.004)	-0.002 (0.004)
Sex Coefficient (se)	0.011 (0.012)	0.026 (0.023)	0.011 (0.022)	0.042 (0.022)	0.060 (0.066)	0.006 (0.119)	0.255 (0.137)
RMQ [B] Coefficient (se)	-0.003* (0.001)	-0.001 (0.003)	-0.002 (0.002)	0.006 * (0.002)	0.005 (0.008)	-0.005 (0.011)	-0.011 (0.012)
RMQ[B] x Δ^{RMQ} Coefficient (se)	0.000 (0.000)	0.000 (0.000)	-0.002* (0.001)	0.000 (0.001)	0.000 (0.001)	0.001 (0.003)	-0.005* (0.002)
Results from validating models using UK BEAM data							
Interval between follow-ups	All intervals	0-1 months	1-3 months	3-12 months	0-3 months	0-12 months	1-12 months
Number of observations over this interval	2130	739	668	610	42	19	52
Mean (SD) Δ^{RMQ}	-1.212* (3.867)	-1.984* (3.803)	-1.186* (3.793)	-0.285 (3.754)	-2.357* (3.570)	-2.579 (4.694)	-0.024 (4.276)
Mean (SD) Δ^{U}	0.025* (0.230)	0.045* (0.238)	0.035* (0.204)	0.000 (0.229)	0.030 (0.286)	0.011 (0.279)	-0.074 (0.315)
RMSE: Differenced score model	0.213	0.227	0.187	0.209	0.254	0.235	0.291
RMSE: Beta regression model	0.214	0.228	0.186	0.212	0.247	0.253	0.304
RMSE: Raw score model	0.215	0.231	0.189	0.211	0.248	0.246	0.301

* denotes values significantly different from 0 at the 5% significance level. Δ^{RMQ} = change in RMQ between observations. Δ^{U} = change in EQ-5D utility

between observations. RMQ^B = RMQ score at start of interval. Intercept = predicted Δ^{U} when $\Delta^{RMQ} = 0$ and RMQ^B = 9

Table 3: Results from fitting the state transition response mapping model to differenced data from the BEST trial.

EQ-5D-3L Dimension	Transition	Number observed in dataset	Probability of transition if RMQ unchanged* Estimate (95% CI)			Reference Transition* (No. observed in brackets)	Change in odds ratio (relative to reference transition) per unit change in RMQ Estimate (95% CI)
			RMQ = 5	RMQ = 9	RMQ = 12		
Mobility	1 -> 2	585	0.18 (0.14,0.23)	0.39 (0.32,0.47)	0.59 (0.52,0.67)	1 -> 1 (124)	1.32 (1.23,1.41)
	2 -> 2	574	0.64 (0.59,0.69)	0.82 (0.78,0.85)	0.9 (0.88,0.92)	2 -> 1 (189)	1.30 (1.23,1.37)
Self care	1 -> 2	1145	0.03 (0.02,0.04)	0.07 (0.05,0.09)	0.13 (0.1,0.16)	1 -> 1 (75)	1.32 (1.23,1.37)
	2 -> 2	162	0.48 (0.39,0.57)	0.63 (0.55,0.71)	0.73 (0.66,0.8)	2 -> 1 (84)	1.25 (1.10,1.27)
Usual activities	1 -> 2	120	0.29 (0.21,0.38)	0.53 (0.43,0.63)	0.71 (0.62,0.79)	1 -> 1 (379)	1.28 (1.18,1.39)
	2 -> 2	664	0.66 (0.6,0.71)	0.85 (0.82,0.87)	0.91 (0.88,0.91)	2 -> 1 (237)	1.31 (1.24,1.39)
	2 -> 3	29	0.01 (0,0.01)	0.02 (0.01,0.02)	0.03 (0.02,0.05)	2 -> 1 (237)	1.80 (1.59,2.03)

Category	Transition	N				Transition (N)	
	3 -> 3	11	0.23 (0.07,0.56)	0.25 (0.07,0.59)	0.27 (0.08,0.61)	3 -> 2 (32)	1.22 (0.95,1.58)
Pain	1 -> 2	43	0.82 (0.35,0.97)	0.97 (0.78,1)	0.99 (0.93,1)	1 -> 1 (41)	1.50 (1.14,1.97)
	2 -> 2	989	0.93 (0.89,0.95)	0.90 (0.90,0.90)	0.8 (0.8,0.8)	2 -> 1 (101)	1.67 (1.48,1.89)
	2 -> 3	97	0.03 (0.03,0.03)	0.09 (0.09,0.10)	0.1 (0.1,0.2)	2 -> 1 (101)	2.16 (1.88,2.48)
	3 -> 3	110	0.31 (0.23,0.4)	0.44 (0.34,0.55)	0.55 (0.45,0.65)	3 -> 2 (92)	1.21 (1.11,1.31)
Anxiety or depression	1 -> 2	162	0.23 (0.2,0.28)	0.24 (0.20,0.28)	0.24 (0.2,0.28)	1 -> 1 654	1.08 (1.04,1.12)
	2 -> 2	383	0.61 (0.57,0.65)	0.69 (0.66,0.71)	0.73 (0.71,0.74)	2 -> 1 (170)	1.12 (1.06,1.18)
	2 -> 3	37	0.04 (0.03,0.05)	0.06 (0.05,0.08)	0.07 (0.06,0.1)	2 -> 1 (170)	1.24 (1.12,1.37)
	3 -> 3	34	0.51 (0.35,0.67)	0.54 (0.37,0.69)	0.55 (0.38,0.71)	3 -> 2 (34)	1.06 (0.95,1.20)

*The odds of each transition in the table are compared with the odds for the reference transition to calculate an odds ratio which is then adjusted for change in RMQ using the change in odds ratio reported in the next column.

Table 4: Results from a response mapping model relating EQ-5D-3L dimensions to raw RMQ scores using baseline observations from the BEST trial only

EQ-5D-3L Dimension	Level	Probability of reporting this level if RMQ = 9	Change in odds (compared to level 1) per unit change in RMQ Mean (95% CI)
Mobility	1	0.35 (0.31,0.40)	NA
	2	0.65 (0.60,0.69)	1.28 (1.22,1.34)
	3	NA	NA
Self care	1	0.86 (0.83,0.89)	NA
	2	0.14 (0.11,0.17)	1.33 (1.26,1.4)
	3	NA	NA
Usual activities	1	0.16 (0.14,0.19)	NA
	2	0.84 (0.81,0.86)	1.31 (1.23,1.39)
	3	0.02 (0.01,0.03)	1.69 (1.52,1.87)
Pain	1	0.02 (0.01,0.03)	NA
	2	0.86 (0.85,0.87)	1.22 (1.05,1.42)
	3	0.12 (0.11,0.12)	1.6 (1.37,1.87)
Anxiety or depression	1	0.51 (0.47,0.53)	NA
	2	0.46 (0.43,0.49)	1.13 (1.09,1.18)
	3	0.03 (0.02,0.05)	1.36 (1.25,1.47)

Table 5: Comparison of predictions from response mapping models based on raw vs. differenced score data from the BEST trial

EQ-5D-3L Dimension	Level	Probability of reporting this level at baseline	Probability of reporting this level at follow-up if $\Delta_{RMQ}=0$		Probability of reporting this level at follow-up if $\Delta_{RMQ}=-2$		Predicted incremental impact of a 2-point reduction in RMQ	
			Raw score	Differenced score	Raw score	Differenced score	Raw score	Differenced score
Mobility	1	0.41	0.41	0.41	0.50	0.48	0.09	0.07
	2	0.59	0.59	0.59	0.50	0.52	-0.09	-0.07
	3	NA	NA	NA	NA	NA	NA	NA
Self care	1	0.81	0.81	0.79	0.87	0.83	0.06	0.04
	2	0.19	0.19	0.21	0.13	0.17	-0.06	-0.04
	3	NA	NA	NA	NA	NA	NA	NA
Usual activities	1	0.22	0.22	0.31	0.30	0.39	0.08	0.08
	2	0.74	0.74	0.66	0.67	0.59	-0.07	-0.07
	3	0.04	0.04	0.03	0.03	0.02	-0.01	-0.01
Pain	1	0.03	0.03	0.03	0.04	0.07	0.01	0.04
	2	0.80	0.80	0.75	0.84	0.75	0.04	0.01
	3	0.17	0.17	0.22	0.12	0.17	-0.05	-0.05
Anxiety or depression	1	0.52	0.52	0.52	0.58	0.55	0.06	0.03
	2	0.44	0.44	0.43	0.39	0.41	-0.05	-0.02
	3	0.05	0.05	0.06	0.03	0.05	-0.02	-0.01

Raw score results are derived from the response mapping model presented in table 3. Differenced score results are derived from the state transition mapping model presented in table 4.

Figure Legends

Figure 1: Relationship between changes in RMQ and health utility as predicted by models fitted to BEST raw and differenced data

Figure 1