HEDS Discussion Paper

No.12.12

A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes

Mónica Hernández Alava, Allan Wailoo, Fred Wolfe and Kaleb Michaud

A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes.

Mónica Hernández Alava[1] PhD, Allan Wailoo[1] PhD, Fred Wolfe[2] MD, Kaleb Michaud[2,3] PhD.

[1] *School of Health and Related Research, University of Sheffield, UK*

[2] *National Data Bank for Rheumatic Diseases, Wichita, US*

[3] *University of Nebraska Medical Center, Omaha, US*

Corresponding author:

Allan Wailoo,
Reader in Health Economics,
HEDS, ScHARR,
University of Sheffield
Regent Court,
30 Regent Street,
Sheffield
S1 4DA
Tel: 00 44 114 2220729
Email: a.j.wailoo@sheffield.ac.uk

Abstract

Background

Analysts often need to estimate health state utility values as a function of other outcome measures. Utility values like EQ-5D have several unusual characteristics that make standard statistical methods inappropriate. We have developed a bespoke approach based on mixture models to directly estimate EQ-5D. An indirect method, "response mapping", first estimates the level on each of the five dimensions of the EQ-5D descriptive system and then calculates the expected tariff score. These methods have never previously been compared.

Methods

We use a large observational database of patients diagnosed with Rheumatoid Arthritis (n=100,398 observations). Direct estimation of UK EQ-5D scores as a function of Health Assessment Questionnaire (HAQ), pain and age was performed using a limited dependent variable mixture model. Indirect modelling was undertaken using a set of generalized ordered probit models with expected tariff scores calculated mathematically. Linear regression was reported for comparison purposes.

Results

The linear model fits poorly, particularly at the extremes of the distribution. Both the bespoke mixture model and the generalized ordered probit approach offer improvements in fit over the entire range of EQ-5D. Mean average error is 10% and 5% lower compared to the linear model respectively. Root mean squared error is 3% and 2% lower. The mixture model demonstrates superior performance to the indirect method across almost the entire range of pain and HAQ.

Limitations

There is limited data from patients in the most extreme HAQ health states.

Conclusions

Modelling of EQ-5D from clinical measures is best performed directly using the bespoke mixture model. This substantially outperforms the indirect method in this example. Linear models are inappropriate, suffer from systematic bias and generate values outside the feasible range.

Acknowledgments

This study was funded was funded by the National Institute for Health and Clinical Excellence ("NICE") through its Decision Support Unit. The views, and any errors or omissions, expressed in this article are of the authors only.

Introduction

In economic evaluation, it is typical for analysts to estimate quality adjusted life years (QALYs) by administering a preference based health utility instrument to patients as part of a clinical study. Where no such instrument has been included in the clinical study, analysts regularly attempt to estimate the relationship between health utilities and some measure of outcome that has been included in the clinical studies by making use of other datasets. If other studies exist where patients have completed both a health utility instrument and the clinical outcome measure, then there exists the possibility of statistically estimating the relationship between the two. This process bridges the gap between the evidence required for the economic analysis and that available from the studies of clinical effectiveness and has variously been referred to as "mapping", "cross walking" and "transfer to utility"([1]). This is widely undertaken in economic evaluation. In a recent review of economic analyses submitted to the National Institute for Health and Clinical Excellence (NICE) in the UK, 22% were found to incorporate such approaches ([2]).

There are of course other reasons why analysts may wish to estimate health state utility values as a function of a range of different explanatory variables. For example, health utility instruments are increasingly accepted as performance measures in their own right and can be used to make comparisons between providers, interventions and conditions. There has therefore been a corresponding increase in such analyses.

However, health state utility data have several features that raise statistical challenges. They are right limited at 1 (full health), left limited at the worst health state and, in some cases, have gaps and multimodal distributions. Linear regression, whilst in widespread use ([3]), is not appropriate in this situation and leads to biased results. We have previously developed a

bespoke approach to direct modelling of EQ-5D data ([4],[5]) which reflects all of these characteristics and does not suffer from the systematically poor fit associated with other simple methods.

An alternative approach is an indirect method that has been referred to as "response mapping" ([6]). This approach has again been tested using the EQ-5D as the outcome of interest. Five separate equations are used to estimate the probability of being in each of the three levels for the different domains of health covered by EQ-5D. Expected tariff score values are then derived from these regressions as a separate second step. Whilst there is mixed evidence regarding the performance of this approach compared to linear regression, it does have intuitive appeal since it is more closely related to the actual data generation process for EQ-5D.

The direct approach based on bespoke mixture models and the indirect approaches have never previously been compared to each other. This paper provides that comparison utilising a very large dataset from patients with rheumatoid arthritis (RA) that includes the EQ-5D as a dependent variable. Section 2 describes the dataset and statistical methods. Section 3 provides results, followed by conclusions.

Methods

Statistical models

Direct models for EQ-5D tariff scores

We estimated two types of direct models. First, a simple linear regression with random effect (Model 1):

$$y_{it} = x'_{it}\beta + u_i + \varepsilon_{it}$$

where y_{it} represents the EQ-5D tariff score for individual *i* at time *t*. β is a (K×1) vector of coefficients, x'_{it} is a row vector of the within- and between-level covariates, ε_{it} is the within subject random variation assumed Independent and Identically Distributed (IID) N(0, σ_ε^2), u_i is an individual random error which is N(0, σ_u^2) and ε_{it} is independent of u_i. The linear model thus assumes conditional normality and it is this assumption that is unlikely to be appropriate given the distribution of EQ-5D.

The second approach is a modified version of the model described by Hernandez *et al.*([4]) (Model 2). The general approach is based on two innovations to reflect key features of the typical EQ-5D tariff distribution. First, EQ-5D is a limited dependent variable: values cannot exceed 1 (full health) or be lower than -0.594 (the "pits" state valuation) and there tends to be a mass of observations, at least at the upper extreme. Tobit type models were originally intended to deal with such limited dependent variables[7], though they are often used in a manner more applicable to censored dependent variables which is clearly not the case in relation to health state utilities. However, in the case of EQ-5D there is the additional feature that any health state less than full health scores a maximum of 0.883, that is, there is a substantial gap between full health and all other health states. Therefore, the following adaptation was made to the limited dependent variable distribution. y_{it} is assumed to be equal to 1 if the latent variable y_{it}^* is greater than 0.883 and equal to y_{it}^* otherwise. The distribution can be expressed as follows:

$$y_{it} = \begin{cases} 1 \; if \; y_{it}^* > 0.883 \\ \max\{y_{it}^*, -0.594\} \; otherwise \end{cases}$$

$$y_{it}^* = x'_{it}\beta + \varepsilon_{it}$$

$$\beta_0 = z'_i\alpha + u_i$$

where y_{it} represents the EQ-5D tariff score for individual *i* at time *t*, β is a vector of coefficients, which includes a random intercept β_0 which varies with individual characteristics z_i . The ε_{it}'s are the within-subject random variations, each assumed IID N(0, σ_ε^2), u_i is an individual random error N(0, σ_u^2) and the ε_{it}'s are independent of u_i.

This demonstrates that the EQ-5D value is a composite of the latent variable y_{it}^* and the probability of being either in excess of 0.883 or less than -0.594. Strictly speaking, the EQ-5D generates 243 discrete values across its range. However, all the gaps except for that between full health and 0.883 are relatively small. Therefore, our approach treats the remainder of the distribution as continuous.

The second innovation is to use the adjusted, limited dependent variable distribution in a mixture model. Such models combine a number of different component distributions to form a new density. Mixtures are an extremely flexible and convenient manner in which complex distributions (such as EQ-5D) can be analysed in a semi-parametric manner.

Classification of an observation into a particular component is modelled using a multinomial logit. Thus, the conditional probability of any observation belonging to class *c* can be written as:

$$P(C_{it} = c|w_{it}) = \frac{\exp(w_{it}'\partial_c)}{\sum_{s=1}^{P}\exp(w_{it}'\partial_s)}$$

where w_{it}' is a vector of variables that affect the probability of component membership, ∂_c is the vector of corresponding coefficients.

Indirect model for EQ-5D: response mapping

The third model (Model 3) that we estimate is derived from a set of five random effects generalised ordered probits, one for each dimension of EQ-5D. Each of these models predict for each observation the probability of selecting each level in that dimension. It has been found in the literature ([6]) that the standard ordered models (probits or logits) are not flexible enough as they assume the same coefficients for the explanatory variables across the different categories (parallel line assumption). This has lead researchers in this area to use a multinomial logit model instead ([6,8,9,10]). This relaxes the parallel line assumption but at the expense of ignoring the ordinal nature of the dependent variable. However, there exists a generalisation of the standard ordered probit model which relaxes the parallel line assumption while still taking into account the natural ordering in the dependent variable (see for example, Maddala, 1983)[11]. Let q_{it}^{s} denote a 3 point ordered discrete dependent variable for each of the five dimensions of EQ-5D, s={mobility, self care, usual activities, pain, anxiety and depression}. The conditional probabilities of observing the three outcomes, q_{it}^{s}, for each of the five s dimensions of EQ-5D can be written as:

$$\begin{gathered} P(q_{it}^{s} = 1|x_{it}, u_i^{s}) = 1 - \Phi(x_{it}\beta_1^{s} + u_i^{s}) \\ P(q_{it}^{s} = 2|x_{it}, u_i^{s}) = \Phi(x_{it}\beta_1^{s} + u_i^{s}) - \Phi(x_{it}\beta_2^{s} + u_i^{s}) \\ P(q_{it}^{s} = 3|x_{it}, u_i^{s}) = \Phi(x_{it}\beta_2^{s} + u_i^{s}) \end{gathered}$$

where x_{it} includes all variables and an intercept term and u_i^{s} is an IID normally distributed mean zero, variance σ_{us}^{2} individual error term.

Conditional on all q_{it}^{s} for each individual, EQ-5D, y_{it} , can be calculated by using the standard tariff values for the relevant question, in this case the UK tariff.

These models predict the individual probabilities for each of the dimension scores(q_{it}^{s}). The expected EQ-5D tariff score is calculated as the average of all the 243 possible combinations of the five EQ-5D dimensions, weighted by their corresponding estimated probabilities. Note that in this paper we calculate the expected values mathematically.

All models were estimated using maximum likelihood methods. The random effects regression and the random effects generalised ordered probits were estimated using STATA v11. The random effects generalised ordered probit was estimated using the REGOPROB module for STATA[12]. We programmed the rest of the analyses and data simulations using GAUSS v11 (Aptech systems Inc.) and used both local and global optimisation methods to ensure identification of the true maximum of the likelihood function of the direct model.

Models were refined and compared using a variety of different tools. Penalised likelihood measures (AIC and BIC) were used as a guide to the optimal model selection within each class of models. BIC in particular was used to guide the optimal number of components in the mixture model since there is considerable support for its use in this setting ([13,14]).
Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) are simple summary measures of fit used to compare across models, including by subsections of the data distribution. Both are relatively insensitive but have been widely used in the "mapping" literature and are therefore also reported here.

Monte Carlo simulation was also used to generate data from each of the three model types. This provides a further method for model comparison. It generates data values that can be used to assess the face validity of the data generating process implied by the model and allows comparisons with the observed data. Importantly, these simulated values are those that would be used in a patient level cost-effectiveness model and in RA many models do adopt this level of analysis([15,16,17]). The generation of non-feasible values, for example, is an important issue for analysts to consider, in addition to those of general model fit for the average. A thousand simulated values were produced for each model.

Dataset

Data were provided by the US National Data Bank for Rheumatic Diseases (NDB). The NDB is a not-for-profit rheumatic disease research databank in which patients complete detailed self-report questionnaires at 6 month intervals ([18]). Eligible patients in this study were those with RA who had completed a biannual survey for events occurring between July 1, 2002 and November 22, 2010.

At each assessment, demographic variables were recorded including sex, age, ethnic origin, education level, current marital status, medical history and total family income. Patients also completed the Health Assessment Questionnaire Disability Index (HAQ), including pain on a visual analogue scale (VAS) scored from 0-100 and EQ-5D, amongst other items. The HAQ is scored between 0 and 3 with higher scores representing greater degrees of functional disability. There is a de facto mandatory requirement for its inclusion in RA clinical trials and it is also widely used as the driver for many economic models ([15,16,17]). UK EQ-5D tariff values (or "index scores") were applied for this analysis to aid comparison with results from previous studies.

A total of 103,867 observations were included in the total dataset from 16,011 patients. Missing data occurred in 3,469 observations and were excluded in the statistical models. The size of the dataset dwarfs that which is typical of most "mapping" studies and provides a good exemplar in which to test competing methods because patients spanned the full range of HAQ, pain and EQ-5D values. Still, very few patients were observed in the most extreme severity HAQ health state; only 1244 observations (1.2%) from 528 patients had a HAQ exceeding 2.5, and just 152 observations (0.15%) from 64 patients had a HAQ of 3.

Figure 1a displays the distribution of the EQ-5D summary score, which demonstrates features typical of data from numerous different disease areas, that is, there is a mass of observations

at full health with two further distinct elements below. Figure 1b shows the distribution of responses within each of the five domains of the EQ-5D descriptive system. Only a small proportion of the respondents are at level three on any of the dimensions, though the greatest proportion is in the domain of pain and discomfort.

Results

The optimal linear regression specification included HAQ and HAQ^2, pain, gender, age and age^2 as explanatory variables. Age entered the model as the difference in age from the mean of the sample (62.82) divided by 10. Table 1 provides details.

A four component mixture model was selected as the optimal model([5]). Each of the components includes HAQ and HAQ^2, pain, age and age^2 as explanatory variables, though it can be seen that these are not always statistically significant and the magnitude of effect differs greatly between the components. Table 2 provides the coefficient values for each of the classes.

The first component of the mixture has HAQ and pain negatively related to EQ-5D ($p<0.000$). HAQ^2 is not significant. A positive relationship with age and age^2 is demonstrated but in the case of age^2 this is not statistically significant ($p=0.23$). For the second component, the coefficients for HAQ and HAQ^2 indicate that EQ-5D decreases, by increasing amounts, as HAQ worsens. The impact of pain on EQ-5D in this group is the most pronounced of all the classes. In component 3 HAQ is negatively associated with EQ-5D and is much greater in magnitude than the positive coefficient on HAQ^2. Pain is also negatively associated with EQ-5D. The final, 4th component shows no statistically significant relationship between EQ-5D

and either age or pain. HAQ is negatively related to EQ-5D ($p<0.05$). HAQ^2 is not statistically significant.

Results for the generalized ordered probit models are shown in Table 3. It is not possible to interpret the coefficients of these type of models directly as the effect differs across individuals and in general, the sign of the coefficient does not even determine the direction of the effect. We can however make some general statements about the effects on some of the probabilities. The conditional probability of being at level one, "no problems", decreases for variables with positive coefficients and the probability of being at level three, "severe problems", increases. Thus, for all five dimensions of EQ-5D, as pain increases, the probability of being at level one decreases and the probability of being at level three increases, *ceteris paribus*. The interpretation for HAQ is more complex due to the inclusion of the squared term. The probability of being at level 1 decreases as HAQ increases (greater functional disability) for all dimensions except the dimension of pain/discomfort. Here, once HAQ exceeds 1.875 the probability of being in level 1 begins to increase. The probability of being at level 3 increases was HAQ rises for the EQ-5D dimensions of "usual activities", "pain" and "depression/anxiety". This relationship also holds for "mobility" and "self-care" across most of the range of HAQ. However, the direction of the relationship reverses when HAQ is very low: below 0.5 for "mobility" and below 0.75 for "self-care". Note that the magnitude of these changes may be negligible.

Table 4 provides details of summary fit measures and this is supplemented by Figures 2a and 2b that show how the mean of the predicted EQ-5D values by HAQ and pain contrast with the mean of the observed data. Overall, model fit is substantially better using both the adjusted mixture model and the generalized probit models compared to a simple random

effects linear regression. MAE improves from 0.131 to 0.118 with the mixture model (a 10% improvement) and to 0.124 with the indirect modelling (5% improvement). RMSE is also improved and is lowest for the mixture model approach. Table 3 shows that there are substantial improvements in model fit relative to the linear model across the entire 0-3 range of HAQ. Improvements in MAE exceeding 11% are observed at both the highest and lowest ranges of functional disability when using the mixture model. There is also substantial improvement in the intermediate HAQ range. RMSE improves but since this is a less sensitive measure the proportional improvement is lower. At pain scores of zero the MAE reduces from 0.13 to 0.08, a 35% improvement. At pain scores exceeding 95, the MAE reduces from 0.23 to 0.18, a 22% improvement.

The response mapping approach also generates improvements over the linear model across the entire spectrum of functional disability, but the improvement is less than that observed for the mixture model method in the subsections presented in Table 3. The mixture model outperforms the generalised ordered probit model approach in all sections of the data as divided in Table 3 both in terms of MAE and RMSE. The improvement is greatest at low levels of disability, where the bulk of the data are observed.

Figure 2a shows that there is one section of the HAQ scale where this is not the case. When HAQ exceeds a value of approximately 2.5, the mean expected values from the generalised ordered probit model approach are closer to the observed data than the mixture approach. Figure 2b illustrates the mean fitted values as a function of pain. This provides a clearer demonstration of the very close fitting of the mixture model to the observed data and this is consistent across the entire pain range. The generalised ordered probit model flattens the function and as such does not fit well across large parts of the range, and is particularly poor

at the extremes. Where pain is zero, the MAE for the response mapping approach is 0.11 compared to 0.08 for the mixture model. For pain exceeding 95, the MAE for the response mapping approach is 0.20 versus 0.18 for the mixture model.

Simulated values

Figure 3 compares the distribution of the observed data from the NDB with that generated from the three different types of statistical models estimated. These simulations reflect individual level variability, as is obviously present in observed data. Figure 3b clearly demonstrates that the data generating process for EQ-5D is fundamentally different from the assumption of conditional normality which underpins the linear regression model. Here values are generated that fall outside the feasible range. This problem is particularly acute at the higher range of values but there are also a smaller number of values generated that fall below the minimum value of -0.594. Neither the mixture model nor the response mapping approaches can generate values outside the feasible range.

The key features of the EQ-5D are present in the simulated values from the mixture model approach. A mass of values at full health can be observed with a clear gap to the next set of values. A tri-modal distribution is evident with values for the remaining two elements of the distribution centred around 0.7 and 0.0. Simulated values from the response mapping approach reflect that this treats EQ-5D as fully discrete rather than continuous. Thus, the tri-modal distribution generated by the mixture model approach contrasts is repeated here but with "lumps" within the different sections of the distribution compared to the smooth results from the mixture model. The only substantial difference between the original data and the response mapping simulated results is that the latter obtains a lower proportion of the distribution at full health.

Conclusions

The EQ-5D is an instrument that demonstrates a number of statistical challenges that make simple off-the-shelf approaches to multivariate regression inappropriate. The poor performance of the linear regression has been observed in numerous other studies, including in RA[4] and is confirmed again here using a very large dataset. We have previously developed an approach to direct modelling of EQ-5D values that is based on a mixture of models derived from a bespoke distribution that reflects the fact that EQ-5D values are limited, in the statistical sense. This approach has been compared to linear and tobit models previously using a dataset comprising approximately 500 patients with RA. We have developed the approach and applied it to a very large dataset with more than 100,000 observations([5]).

In this paper, we have developed methods for response mapping by applying an approach that recognises the ordered nature of responses within each EQ-5D dimension. The generalised ordered probit has not previously been applied in the "mapping" field as far as we are aware. Our primary aim however is to compare the bespoke mixture model and the response mapping approaches. These direct and indirect methods are two fundamentally different approaches that have never previously been directly compared. Whilst the former directly estimates EQ-5D tariff scores, the latter uses a two stage approach: first estimating the probability of being on each level of the 5 separate dimensions of EQ-5D, and then estimating the expected value from each of the 243 possible combinations. Both of these approaches have merit because they have been designed to generate values that reflect the principal characteristics of the process by which EQ-5D data are generated. This ought to be an important consideration in the selection of any statistical model.

Most previous applications of the response mapping approach have used multinomial logit models, treating the data as nominal, and have used a simulation method to estimate these expected values. Here we demonstrate that the true ordered nature of the data can be reflected using generalized ordered probits. This modelling approach relaxes the parallel line assumption inherent in the ordered logit and probit models. There is not a requirement for simulation methods to estimate the expectations as these can be derived mathematically as we have done here.

The response mapping approach using this specification of generalised ordered probit models substantially outperforms the linear regression in this example. Previous evidence using multinomial logit models has been equivocal[6,8,9,10].

However, we also demonstrate that in this example dataset the better performing model is the bespoke mixture. Fit is vastly better than the linear model and substantially better than the response mapping approach across the entire range of pain, EQ-5D and HAQ with one exception: where HAQ exceeds 2.5 the response mapping approach is closer to the mean observed values. However, there are only 1% of patient observations at this extreme level of functional disability. Improvement in fit in the mixture model could be obtained by adding a greater number of components. However, this could potentially be a large increase due to the relatively small amount of data here. Adding more components will initially be more efficient where these are at other levels of functional disability. Furthermore, the credibility of data at this extreme is questionable. Certainly, patients would not be able to self-complete the forms if they were unable to do any of their daily activities of living, though the NDB does allow forms to be completed over the phone by interviewers or by the patient's assistance-provider.

Whilst this dataset was selected because it offers typical features of EQ-5D in which to compare methods, it may be warranted to complete further comparisons of the mixture modelling and response mapping approaches before definitive conclusions are reached. No such caution is required in the case of the linear model as there is now a wealth of evidence against its use.

Despite this caution, we provide some reasons why the response mapping approach may not perform as well as the bespoke mixture model. First, there are just three levels in each question in the EQ-5D. Therefore, the crudeness of the instrument means that it is quite possible that quite large errors can occur in the estimated values. Second, the correlations between the models for each of the five levels has not been investigated here. This, together with the potential gains from using more flexible functional forms for response mapping models are areas worthy of further investigation.

The response mapping approach does offer the potential advantage that weights from any country can be applied in the second stage rather than requiring the estimation of a new function as would be the case with all direct methods. The danger with that approach is that even where a good fit may be achieved with one set of weights, there is no guarantee that the method will perform well with a different set.

References

[1] Mortimer D, Segal L, Sturm J. Can we derive an 'exchange rate' between descriptive and preference-based outcome measures for stroke? Results from the transfer to utility (TTU) technique. *Health Qual Life Outcomes* 2009; 7:33.

[2] Kearns B, Ara R, Wailoo AJ. A review of the use of statistical regression models to inform cost effectiveness analyses within the NICE technology appraisals programme, Report by the NICE Decision Support Unit

[3] Brazier J, Yang Y, Tsuchiya A, Rowen, D. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010; **11**: 215–225.

[4] Hernández Alava M, Wailoo AJ, Ara R. Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Health State Utility Values. *Value Health* 2012; **15**: 550-561.

[5] Hernández Alava, M., Wailoo, A., Wolfe, F., and Michaud, K.(2012) The relationship between EQ-5D, HAQ and pain in patients with rheumatoid arthritis: further validation and development of the limited dependent variable, mixture model approach, HEDS Discussion Paper 12/10.Available at: http://www.shef.ac.uk/polopoly_fs/1.199216!/file/hedsdp1210.pdf

[6] Gray A, Rivero-Arias O, Clarke P. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. *Medical Decision Making* 2006; 26(1):18-29.

[7] Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958;26:24-36.

[8] Dakin, H, Gray, A., and Murray, D. (2012) Mapping analyses to estimate EQ-5D utilities and responses based on Oxford Knee Score, Quality of Life research, DOI 10.1007/s11136-012-0189-4

[9] Rivero Arias, O., Ouellet, M., Gray, A. et al. (2010) Mapping the Modified Rankin Scale (mRS) Measurement into the Generic EuroQol (EQ-5D) Health Outcome, Medical Decision Making, Vol. 30: 341–354)

[10] Pinedo Villanueva, R.A., Turner, D., Judge, A., et al. (2012) "Mapping the Oxford Hip Score onto the EQ-5D utility index", Quality of Life Research, DOI 10.1007/s11136-012-0174-y

[11] Maddala, G.S. (1983). Limited Dependent and Qualitative Variables in Econometrics, Cambridge.

[12] Boes, S., (2006). REGOPROB: Stata module to estimate random effects generalized ordered probit models, Statistical Software Components S456604, Boston College Department of Economics, revised 06 Sep 2006.

[13] McLachlan, G.J., and Peel, D. (2000) Finite Mixture Models, New York: Wiley.

[14] Muthén, B. and Muthén, L. (2008) Mplus. Statistical Analysis with Latent Variables. Users Guide. Los Angeles: Muthén and Muthén.

[15] Wailoo, A.J., Bansback, N., Brennan, A., et al. (2008) Biologic Drugs for Rheumatoid Arthritis in the Medicare Program: A Cost Effectiveness Analysis. *Arthritis and Rheumatism,* Vol.58:939-946.

[16] Tosh J, Brennan A, Wailoo A, Bansback N. (2011) "The Sheffield rheumatoid arthritis health economic model", Rheumatology, Vol.50 Suppl 4:iv26-iv31

[17] Chen, Y-F., Jobanputra, P., Barton, P., et al. (2006) A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. Health Technology Assessment, Vol.10.

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

Tables

Table 1: Results from linear regression model (Model 1)

	parameter	se	p-value
HAQ	-0.0790	0.0034	0.0000
HAQ^2	-0.0409	0.0014	0.0000
Pain/100	-0.0671	0.0086	0.0000
$Pain/100^2$	-0.3109	0.0090	0.0000
AgeM/10	0.0130	0.0008	0.0000
$(AgeM/10)^2$	0.0006	0.0004	0.1570
Male	-0.0422	0.0027	0.0000
Intercept	0.8879	0.0023	0.0000
σ_u	0.1100	0.0009	
σ_ε	0.1414	0.0003	

Note: AgeM = age-$\widehat{age}$

Table 2: Results from adjusted limited dependent variable mixture model (Model 2)

		Parameter	robust se	t-value	p-value
Explanatory variables within component 1	HAQ	-0.0898	0.0027	-32.9151	0.0000
	HAQ^2	0.0005	0.0009	0.5892	0.5557
	Pain/100	-0.0580	0.0023	-25.4275	0.0000
	^{2}AgeM/10	0.0049	0.0005	10.1656	0.0000
	$(AgeM/10)^2$	0.0003	0.0002	1.2111	0.2258
Explanatory variables within component 2	HAQ	0.0544	0.0301	1.8043	0.0712
	HAQ^2	-0.0509	0.0100	-5.1027	0.0000
	Pain/100	-0.3841	0.0225	-17.0781	0.0000
	AgeM/10	0.0291	0.0035	8.2411	0.0000
	$(AgeM/10)^2$	0.0023	0.0017	1.3532	0.1760
Explanatory variables within component 3	HAQ	-0.1415	0.0076	-18.5781	0.0000
	HAQ^2	0.0155	0.0027	5.7871	0.0000
	Pain/100	-0.0839	0.0089	-9.3978	0.0000
	AgeM/10	0.0037	0.0012	3.2078	0.0013
	$(AgeM/10)^2$	0.0007	0.0006	1.1702	0.2419
Explanatory variables within component 4	HAQ	-0.1958	0.0811	-2.4137	0.0158
	HAQ^2	0.0347	0.0246	1.4097	0.1586
	Pain/100	-0.0127	0.0693	-0.1839	0.8541
	AgeM/10	-0.0043	0.0058	-0.7417	0.4583
	$(AgeM/10)^2$	0.0002	0.0021	0.1106	0.9119
Random effects terms	Intercept1	0.8141	0.0013	629.4830	0.0000
	Intercept2	0.4266	0.0164	25.9934	0.0000
	Intercept3	0.3297	0.0081	40.6365	0.0000
	Intercept4	1.0220	0.0327	31.2430	0.0000
	Male	-0.0265	0.0013	-20.9092	0.0000
Variances for each component	Variance1	0.0025	0.0001	48.7842	0.0000
	Variance2	0.0240	0.0016	14.8595	0.0000
	Variance3	0.0022	0.0002	10.2405	0.0000
	Variance4	0.0044	0.0042	1.0374	0.2995
Random effects	Variance	0.0026	0.0001	46.2489	0.0000
Explanatory variables explaining the probability of component membership[1]	Intercept 1	-1.2746	0.0637	-20.0245	0.0000
	HAQ	0.2420	0.4424	0.5471	0.5843
	Pain/100	23.4673	0.5897	39.7970	0.0000
	$Pain/100^2$	-21.5513	0.6707	-32.1307	0.0000
	Intercept2	-6.6310	0.2597	-25.5366	0.0000
	HAQ	2.1936	0.4234	5.1808	0.0000
	Pain/100	18.3719	1.2220	15.0337	0.0000
	$Pain/100^2$	-13.8001	0.8071	-17.0981	0.0000
	Intercept3	-7.4768	0.2988	-25.0242	0.0000
	HAQ	1.0517	0.4344	2.4209	0.0155
	Pain/100	25.3396	1.1359	22.3075	0.0000
	$Pain/100^2$	-16.9622	0.7624	-22.2473	0.0000

[1] These probabilities are computed using component 4 as the reference. [2] AgeM = age-$\widehat{age}$

Table 3: Generalized Ordered Probit models for each EQ-5D Question (Model 3)

	Mobility		Self-care		Usual activities		Pain		Depression/Anxiety	
	Beta	**p**	**Beta**	**p**	**Beta**	**p**	**Beta**	**p**	**Beta**	**p**
Level 1										
HAQ	2.034	0.000	2.812	0.000	2.593	0.000	1.591	0.000	0.890	0.000
HAQ2	-0.151	0.000	-0.160	0.000	-0.346	0.000	-0.424	0.000	-0.080	0.000
Pain/100	1.503	0.000	0.738	0.000	1.672	0.000	5.363	0.000	1.069	0.000
[2]AgeM/10	0.010	0.320	-0.118	0.000	-0.074	0.000	0.017	0.083	-0.209	0.000
$(AgeM/10)^2$	0.005	0.286	-0.004	0.571	0.002	0.641	-0.022	0.000	-0.002	0.682
Sex	0.631	0.000	0.971	0.000	0.480	0.000	0.267	0.000	-0.023	0.497
constant	-2.521	0.000	-4.810	0.000	-2.698	0.000	-0.607	0.000	-1.854	0.000
Level 2										
HAQ	-0.993	0.002	-1.132	0.000	0.037	0.723	0.902	0.000	0.041	0.593
HAQ2	0.847	0.000	0.837	0.000	0.477	0.000	-0.014	0.587	0.209	0.000
Pain/100	0.550	0.001	0.143	0.255	0.865	0.000	4.653	0.000	0.976	0.000
AgeM/10	-0.098	0.006	0.042	0.124	-0.027	0.048	-0.072	0.000	-0.302	0.000
$(AgeM/10)^2$	0.039	0.035	0.041	0.010	0.019	0.014	-0.008	0.175	-0.006	0.494
Sex	0.472	0.000	0.565	0.000	0.364	0.000	0.174	0.000	0.018	0.728
constant	-5.801	0.000	-4.957	0.000	-4.445	0.000	-5.509	0.000	-4.214	0.000
[1]rho	0.527	0.000	0.535	0.000	0.421	0.000	0.420	0.000	0.662	0.000

[1] rho=$\sigma_u^2/(1+\sigma_u^2)$ [2] AgeM = age-$\widehat{age}$

Table 4: Comparison of Models 1, 2 and 3

	N		**Model 1**	**Model 2**	***% diff 2 vs 1***	**Model 3**	***% diff 3 vs 1***	***% diff 2 vs 3***
HAQ 0-1	54,086	MAE	0.0968	0.0854	*11.77%*	0.0906	*6.46%*	*5.68%*
		RMSE	0.1292	0.1215	*5.96%*	0.1250	*3.22%*	*2.83%*
HAQ 1-2	38,307	MAE	0.1571	0.1458	*7.17%*	0.1515	*3.53%*	*3.77%*
		RMSE	0.2061	0.2025	*1.75%*	0.2033	*1.39%*	*0.37%*
HAQ 2-3	8,005	MAE	0.2309	0.2052	*11.11%*	0.2130	*7.77%*	*3.63%*
		RMSE	0.2626	0.2520	*4.01%*	0.2543	*3.16%*	*0.88%*
Overall	100,398	MAE	0.1305	0.1180	*9.56%*	0.1236	*5.30%*	*4.50%*
		RMSE	0.1752	0.1693	*3.37%*	0.1713	*2.24%*	*1.16%*

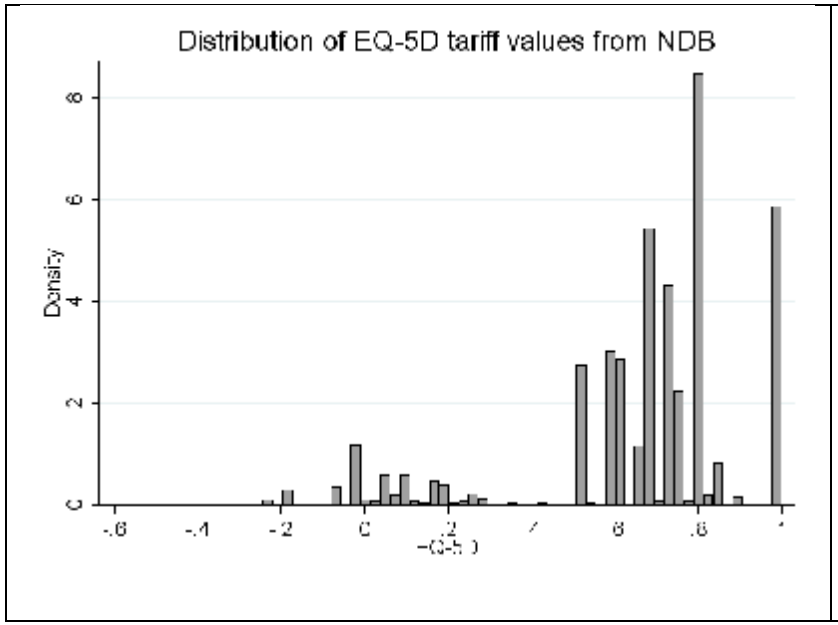
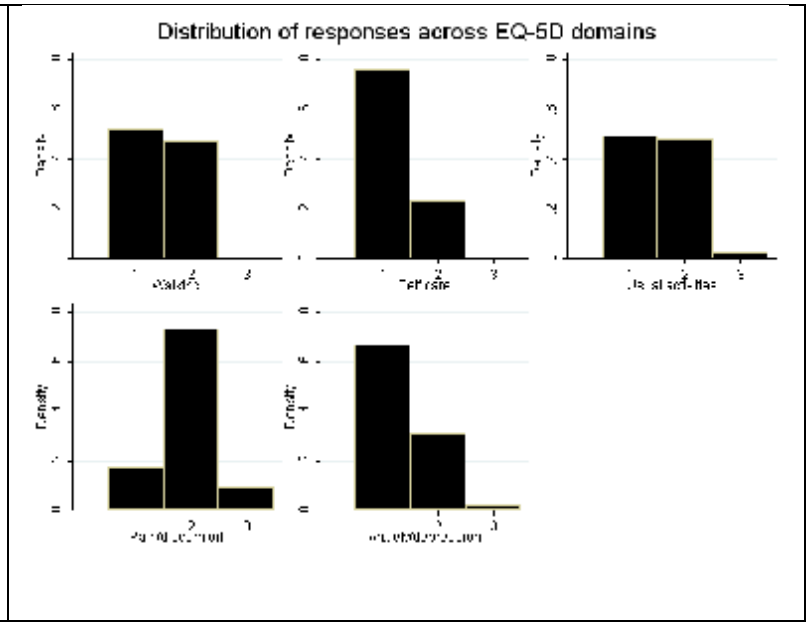
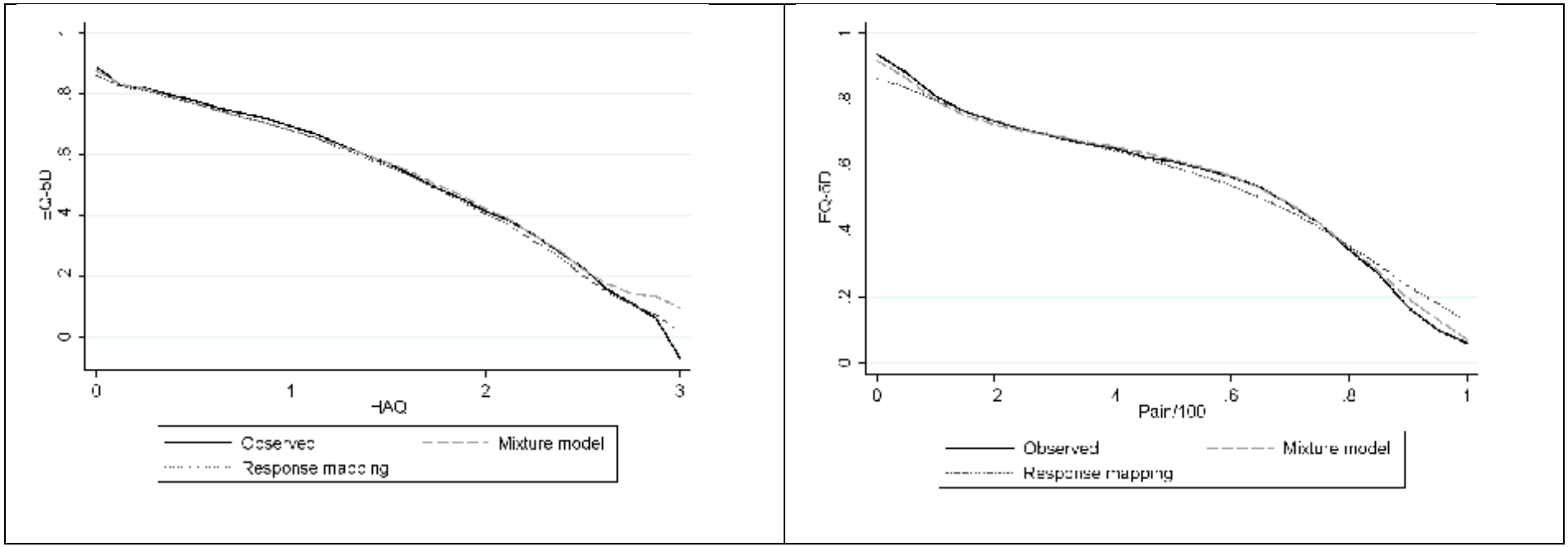

Distribution of EQ-5D tariff values from NDB
Density
0
2
4
6
8
-.6
-.4
-.2
0
.2
.4
.6
.8
1

Figure 1a)


Distribution of responses across EQ-5D domains

1b)



Figures 2a) Mean EQ-5D by mean HAQ: observed vs predicted **b) Mean EQ-5D by mean pain score: observed vs predicted**

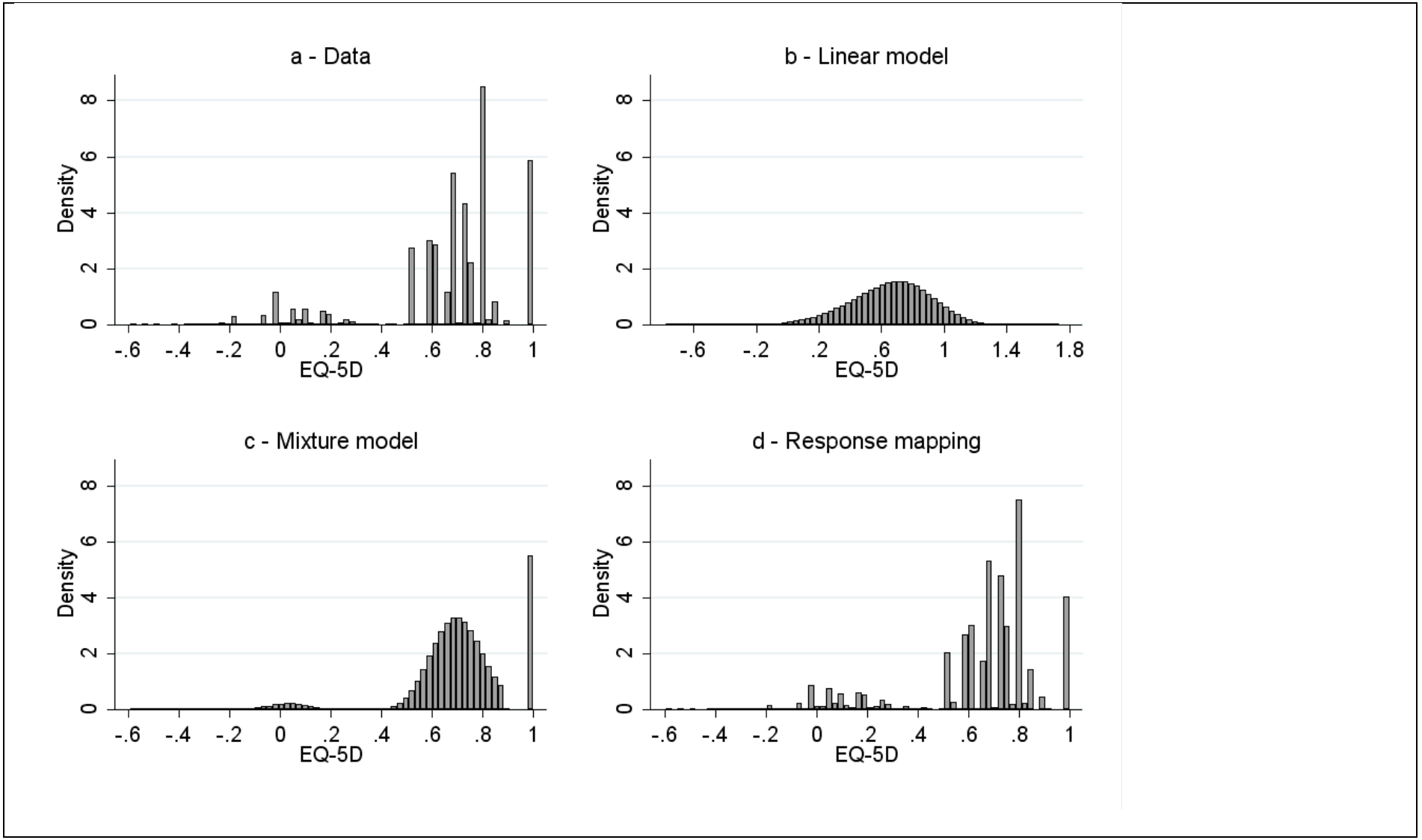


Figure 3a) Distribution of observed data and b) – d) simulated values from linear, mixture and response mapping models.