The
University
Of
Sheffield.

This is a repository copy of *Tails from the Peak District: adjusted censored mixture models of EQ-5D health state utility values*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/74602/

Monograph:
Hernández Alava, M., Wailoo, A.J. and Ara, R. (2010) Tails from the Peak District: adjusted censored mixture models of EQ-5D health state utility values. Discussion Paper. HEDS Discussion Paper 10/08 .

HEDS Discussion Paper 10/08

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

HEDS Discussion Paper 10/08

Tails from the Peak District: Adjusted Censored Mixture Models of EQ-5D Health State Utility Values

Monica Hernandez Alava, Allan J. Wailoo, Roberta Ara
Health Economics & Decision Science,
School of Health and Related Research, University of Sheffield

July 2010

ABSTRACT: Health state utility data generated using the EQ-5D instrument are typically right bounded at one with a substantial gap to the next set of observations, left bounded by some negative value, and are multi modal. These features present challenges to the estimation of the effect of clinical and socioeconomic characteristics on health utilities. We present an adjusted censored model and then use this in a flexible, mixture modelling framework to address these issues. We demonstrate superior performance of this model compared to linear regression and Tobit censored regression using a dataset from repeated observations of patients with rheumatoid arthritis. We find that three latent classes are appropriate in estimating EQ-5D from function, pain and sociodemographic factors. Analysis of utility data should apply methods that recognise the distributional features of the data.

KEYWORDS: Mixture models, Latent Class model, Censored regression, EQ-5D, Mapping

I

1 Introduction

The Quality Adjusted Life Year (QALY) has become one of the most widely used outcome measures in economic evaluations. The QALY is useful to healthcare decision makers seeking to apply a consistent approach across a broad range of disease areas, treatments and patients and is required by several international bodies, such as the National Institute for Health and Clinical Excellence (NICE) in England and Wales (NICE (2008)). Typically, QALYs are generated from patient completion of a survey instrument that provides a generic description of health in terms of symptoms and impact on functioning, to which standardised, preference based scoring systems can then be applied. Instruments such as the EQ-5D, SF6D and Health Utilities Index (HUI) are in widespread use. Other approaches to generating QALYs include the use of disease specific instruments which have similar preference based scoring systems or direct valuation of health states by patients themselves. It is well documented that the use of different approaches or instruments results in different estimates of health state utilities, and therefore, ultimately, different estimates of cost effectiveness and decisions. As a consequence, some decision makers express a clear preference for the use of a particular approach. In England and Wales for example, NICE recommends the EQ-5D (NICE (2008)).

Clinical studies, and in particular randomised controlled trials, used to estimate the treatment effect of a health technology often do not include any preference based outcome measures. Furthermore, even where such outcomes are included, they may not be relevant to the setting for the economic evaluation. There may therefore be a gap between the data available from the clinical studies and the requirements of the economic evaluation.

In some situations it may be possible to bridge this gap by estimating the relationship between a clinical measure(s) and a preference based measure where both have been included in an external dataset. This then provides a statistical link between the treatment effect observed in the clinical studies using clinical outcome

measures and a preference based measure that can be used to estimate QALYs in the economic evaluation. The fitting of a statistical model for this purpose has been referred to in previous literature as "mapping" or "crosswalking", borrowing terminology from psychometrics. There are other situations where one may wish to fit a statistical model to health utility data, for example to explore the impact of socioeconomic factors or treatments directly.

A recent review of 30 studies (Brazier et al. (2009)) indicates that the statistical models used tend to be relatively simplistic. Simple linear models dominate, with limited use of Tobit or similar models for dealing with censored data. However, health state utility data tend to exhibit features which may call for more flexible statistical models. In addition to upper bounding at full health (1), utility data are also left bounded at the worst imaginable health state, have gaps between values and tend to have distinct bi or tri-modal distributions. Furthermore, both clinical trials and observational studies typically include multiple observations from each individual. Statistical models used to estimate health state utilities ought to reflect these data characteristics in order to avoid biased estimates. This is now well recognised in relation to health related costs where the use of generalized linear multilevel (sometimes called random effects or hierarchical) models have been discussed to deal with left boundedness, skewness and the clustered nature of the data (Hernandez Alava and Wailoo (2010), Thompson et al. (2006)).

In this paper we develop an adjusted censored regression to address the right censoring and the gap between full health and intermediate health states that is a feature of EQ-5D. We then develop flexible, random effects mixture (or latent class) models to account for the other key features of the distribution of EQ-5D values. Section 2 provides a detailed account of the statistical issues to be addressed and outlines how these have been considered in the literature to date. Then in section 3 we describe the data and methods. This first covers a description of an example dataset from a trial of patients with rheumatoid arthritis (RA) and then describes

3

the statistical models to be applied. Results are provided in section 4. Section 5 provides a discussion of the results and the implications of the models estimated here.

2 Background

2.1 The typical distribution of EQ-5D

The EQ-5D questionnaire asks respondents to describe their health in five domains (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) each of which can be at one of three severity levels (no problems/some or moderate problems/extreme problems). 243 combinations can be described in this way. A score can be assigned to each of these states based on analysis of preference data obtained from approximately 3,000 members of the general public in the UK Dolan et al. (1995).

Figure 1 displays the distribution of health state utilities derived using UK preference weights for the widely used EQ-5D instrument from 11 studies in different clinical areas. Several features should be noted. First, there is often a mass of observations at 1, full health, which is the maximum value feasible for health utility. There is then a relatively large gap before the next observations which begin at 0.883. 0.883 is the highest utility score that can be generated using the UK regression model for scoring EQ-5D reported by Dolan et al. (1995) and applies to the health state 11211, that is, where the patient indicates that the only reduction from full health is by having "some problems" with their usual activities. The score of 1, that is, full health, is not capable of being generated by the Dolan model. Thus, there is no connection between the value for full health and the values for all other health states accounting for this first large gap in observed EQ-5D scores.

Second, for each of the examples, the distribution of values are bi or tri-modal with each of the separate components of the distribution centred around 0.7 and

0.2 approximately. Their precise location, the degree of kurtosis and skew in each of these components varies according to the specific setting and, in particular, the severity of the condition. The lower section of the distribution is a consequence of the "N3" term in the scoring model which assigns a large utility decrement to any health state that includes extreme problems in any dimension. Furthermore, these characteristics are not limited to health utilities generated via the UK EQ-5D scoring model. The US scoring of EQ-5D for example also demonstrates a large gap between full health and the next set of values at 0.86, Shaw et al. (2005) and apparently separate components of the distribution below this level. Huang et al. (2008) presents a histogram of US EQ-5D valuations from patients with HIV that demonstrates this gap as well as the multi modal characteristic of the distribution.

The process of building a statistical model which uses EQ-5D as the dependent variable should recognise these features in order to avoid biased estimates.

2.2   Existing literature

Brazier et al. (2009) identified 30 studies which examine the relationship between health outcome measures that are not preference based and generic preference-based measures. Of these, half used the EQ-5D as the dependent variable, with variants of the Health Utilities Index (HUI) in 8 studies and SF-6D in 5 studies. AQoL, QWB and I5D were included in the remainder with some studies including more than one measure.

The authors of the review reported that the vast majority of the included studies fitted straightforward linear models. There was very limited use of censored regression models or any attempt to deal with the non normal characteristics of the distribution of values. Typically scant attention is paid to model diagnostics. The R^2 statistic was seen to dominate model selection in these studies, notwithstanding the fact that R^2 cannot be used to compare models with different dependent vari-

Figure I: Distribution of EQ-5D from II studies in different clinical areas

Asthma (n=2,935)

Chest pain (n=679)

Cronic obstructive pulmonary disease (n=185)

Clodronate (n=320)

Hormone replacement therapy (n=755)

Irritable bowel syndrome (n=374)

Lower back pain (n=500)

Leg Ulcers (n=233)

Leg reconstruction (n=92)

Osteoporosis (n=221)

Varicose veins (n=887)
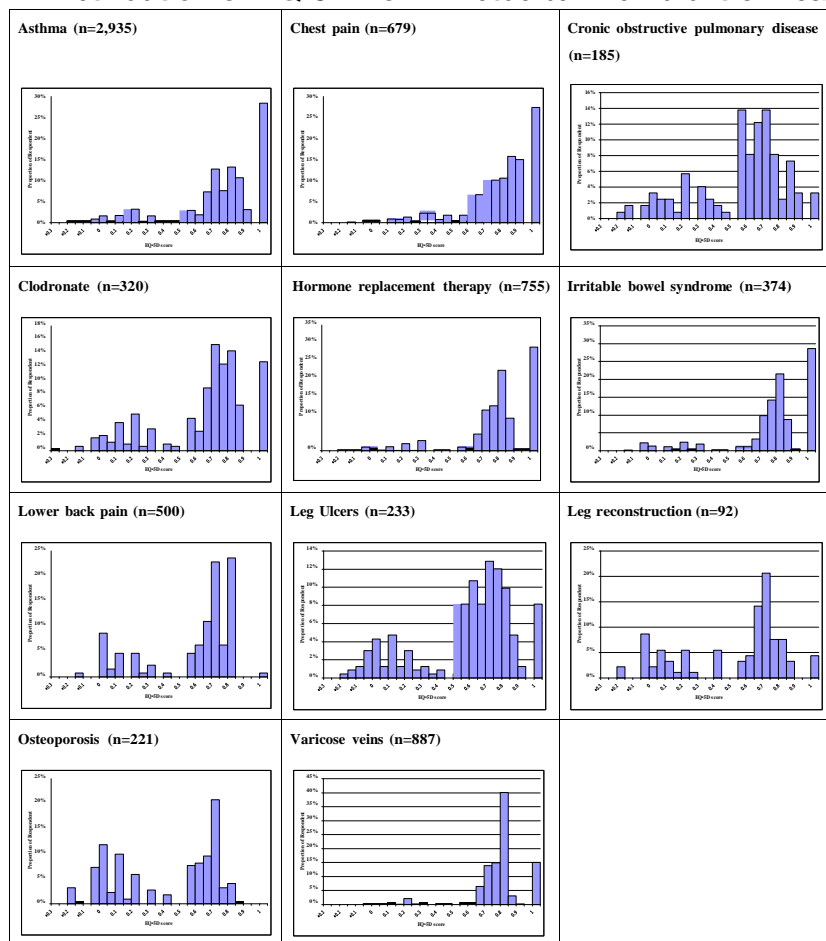
ables and a raft of other criticisms of this approach that are well documented in the econometrics literature, see for example Charemza and Deadman (1992).

Several other authors have considered the use of Tobit and other censored regression models for dealing with the bounded nature of health utility data. These studies were not included in the Brazier et al. (2009) review because they do not attempt to model the relationship between health preference data and other outcome measures but are instead interested in the impact of the determinants of health or socioeconomic factors. Austin et al. (2000) conducted a simulation study to compare linear regression models to Tobit censored models and found Tobit performed better in the presence of censoring. Austin (2002) also compared linear regression with the Tobit model and variants of the Tobit: the symmetrically trimmed least squares and censored least absolute deviations (CLAD) models using data on the HUI from a large Canadian population health survey. Huang et al. (2008) found that the CLAD model performed poorly in a study using the EQ-5D (US scores) as the dependent variable based on data from HIV infected patients. They reported consistently better performance from latent class models and two-part models where a log transformation is used in the second part. The focus of this study was solely on methods to address the mass of observations at one and for this reason they include only two latent classes. However, it is unclear that a two class model is capable of overcoming this challenge unless either the underlying distribution for the classes is itself suitable for censored data or if one class has a zero variance in which case the model is equivalent to the two part model. Similarly, Li and Fu (2009) applied a two part model to US EQ-5D data. Specifically they explored how the second part of the model, for those individuals that do not score one, may be approached in a number of ways.

Pullenayegum et al. (2010) compared the performance of Tobit, CLAD, linear regression, two-part and latent class models when modelling data that are constrained

by one. As with Huang et al. (2008), only two latent classes were included, both with normal distributions. They reported that both Tobit and CLAD models yield biased estimates when the variable of interest is not able to exceed one rather than simply being censored by the measurement instrument at one. Linear regression, latent class model and the two part estimators were reported as unbiased.

No studies have developed methods to deal with the numerous challenges that are presented by the distribution of EQ-5D data. Of particular note is that the use of latent class models has been limited to addressing the issue of upper censoring but, as previously mentioned, it is not clear that the general framework provided by this approach has been fully exploited to date. A general problem with modelling of EQ-5D utility values to date is that the fit of the models is poor at the extremes of the distribution. Specifically, models tend to underpredict at the upper extreme of the EQ-5D scale and overpredict over much of the remainder of the scale but particularly at the lower end, Brazier et al. (2009), Rowen et al. (2009). Therefore, as Crott and Briggs (2010) note, there is currently no agreement of the best method to use. Studies have not applied the same criteria for judging the appopriateness of models and there are differences in the characteristics of the datasets used in these studies that may influence the findings.

3 Data and Methods

3.1 The Rheumatoid Arthritis dataset

Rheumatoid Arthritis (RA) is the most common form of inflammatory arthritis with prevalence estimated at 0.8% of the population, Symmons et al. (2002), and incidence between 1.5 per 10,000 for males and 3.6 per 10,000 for females in the UK, Cooper (2000). In recent years, treatment of this condition has been vastly altered by the development of so-called biologic drugs. Whilst proven to be clinically efficacious, their relatively high cost make them obvious candidates for cost

effectiveness analyses. Thus, there has been an explosion in the number of such studies undertaken in this field.

Almost all these economic evaluations are based on models which define health states or profiles in terms of the Health Assessment Questionnaire Disability Index (HAQ) (NICE (2009), Wailoo et al. (2008), Chen et al (2006), Kobelt et al. (2002)). This is a validated clinical outcome measure that focuses on the functional capacity of a patient and there is a de facto mandatory requirement for its inclusion in RA clinical trials, Aletaha et al. (2008). The HAQ covers 8 categories of functioning (dressing, rising, eating, walking, hygiene, reach, grip, and usual activities) and is scored from 0 (no disability) to 3 (completely disabled) in steps of 0.125, although the scale is generally treated as fully continuous. Since preference based measures of outcome that would allow direct modelling of QALYs were not included in many of the clinical trials in this area, nor in the studies that inform methods of extrapolation beyond the trials, it becomes necessary to statistically model the relationship between HAQ and a preference based measure of health related quality of life in order to estimate QALYs.

A number of such models have been reported, many of which are used in economic evaluations. These cover a range of preference based instruments including EQ-5D (Bansback et al. (2009), Hawthorne et al. (2000), Hurst et al. (1997), Lindgren et al (2009), Marra et al. (2007), Wailoo et al. (2008)), SF6D (Bansback et al. (2009), Marra et al. (2007),Wailoo et al. (2008)), HUI2 and HUI3 (Bansback et al. (2005), Marra et al. (2007)). Almost all are simple linear regression models. In general these studies only consider HAQ as a covariate, although some studies include age (Marra et al. (2007), Wailoo et al. (2008)), gender (Bansback et al. (2005), Wailoo et al. (2008), Lindgren et al (2009)), or other clinical measures such as pain (Hurst et al. (1997)), disease activity (Hurst et al. (1997), Lindgren et al (2009)), or disease duration (Wailoo et al. (2008)).

Most consider only a linear relationship between HAQ and health utility. Malottki

Figure 2: Histogram of EQ-5D from patients with rheumatoid arthritis used m current study
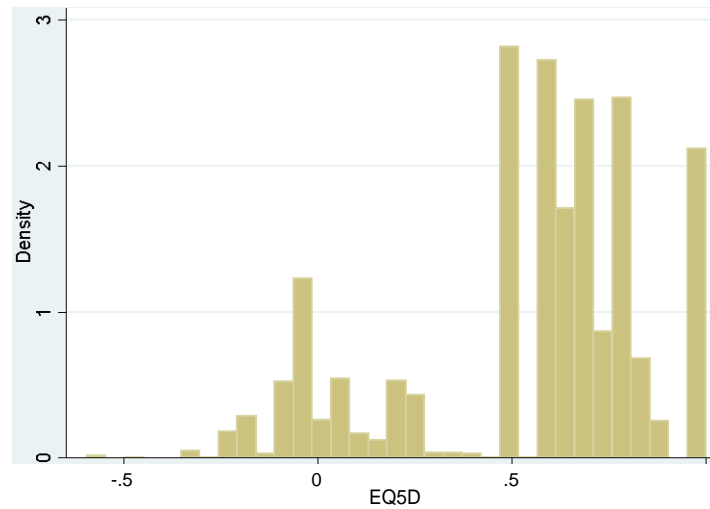
et al. (2009) include a quadratic term for HAQ making the relationship to EQ-5D non-linear. Wailoo et al. (2008) use a non linear regression with logistic function to constrain the predicted values of the feasible range of the EQ-5D/SF6D instruments. Bansback et al. (2009) models the individual domains of HAQ as explanatory variables rather than the HAQ summary measure and is therefore a quite different approach to the other publications.

Here we use a new, rich dataset to estimate the relationship between EQ-5D and HAQ as well as other relevant explanatory variables. The data come from the Combination Anti-Rheumatic Drugs in Early Rheumatoid Arthritis (CARDERA) Trial (Choy et al. (2008)) comprising 467 patients randomised to receive four different drug treatment strategies. Patient outcomes were assessed at baseline, 6, 12, 18 and 24 months.

Whilst this was a patient population with recently (within 2 years) diagnosed RA, the spread of patients across the entire feasible ranges of both HAQ and EQ-5D make this a dataset well suited for the current problem. As is typical with most studies, particularly clinical trials which tend to recruit relatively healthier patients than in clinical practice, there is a relative paucity of observations at the most

10

extreme level of functional limitation, that is a HAQ score of 3.

The data shown in Figure 2 clearly demonstrate the typical pattern in EQ-5D data. There is a mass point at 1 (full health) and a clear gap until the next set of values. There are at least two other groupings – one between 0.5 and 0.85 which has a left skew and another centred around zero. The data span the entire range of feasible EQ-5D values.

3.2   Models

We estimated models within four broad classes.

Model 1:   Random effects linear regression.   We use a standard linear regression with random coefficients to reflect the fact that each patient provides values at several timepoints during the study. The estimated model for y_{it} (EQ-5D for individual i -level 2 or between level units, and time period t - level 1 or within level units) can be written as:

$$y_{it} - x_{it}\beta_i + E_{it}$$
$$\beta_{ki} - {}_i a_k + u_{ki}$$

where β_i is a (k × 1) vector of random coefficients β_{ki}, x_{it} is a row vector of level 1 covariates, $_i$ is a row vector of level 2 covariates, E_{it} is IID $N(0, 0^2)$, u_{ki} is an element of the (k × 1) vector u_k which is $N(0, D)$ and E_{it} is independent of all the u_{ki}. In principle, this general specification allows all the coefficients in the vector β_i to be random. The majority of applications where a random effects linear regression model is used allow only the intercept to vary randomly.

Model 2:   Random effects Tobit model. The linear regression assumption is problematic because it implies that values outside the EQ-5D lower and upper boundaries can be generated by the model. The Tobit model takes into account

not only that our dependent variable is censored at one but also that there may be a substantial concentration of observations at the censored point (full health)[1]. A latent variable, y_{it}^*, is defined with a conditional normal distribution. This latent variable is artificially censored at one in our case, turning the usual regression model into a model with a discrete element at the censored point and a continuous model elsewhere. The top panel of Figure 3 shows the differences between the distributions implied by the linear regression and the Tobit models and how the Tobit model is able to generate a concentration of observations at the tail of the distribution. This is essentially an ad hoc modification of the previous linear regression model 1 in order to account for these features. In many applications, the latent variable y_{it}^* is given some meaning. For example, when modelling the number of hours worked using a Tobit model, the latent variable is often thought of as the "desired hours of work" which may be negative. However, the derivation of the original Tobit model (Tobin, 1958) does not require any interpretation of the latent variable. The random effects Tobit model can be written as:

$$y_{it} = \min\{y_{it}^*, 1\}$$

$$\hat{y}_{it} = x_{it}\beta_i + \varepsilon_{it}$$

$$\beta_{ki} = {}_ia_k + u_{ki}$$

Model 3: Random effects Adjusted Censored Model (ACM). Another key feature of EQ-5D data is that it is not feasible to generate values between 0.833 and 1. The EQ-5D instrument is relatively crude and may be insufficiently sensitive to detect minor departures from full health. Departures from full health are scored as equivalent to full health unless they are sufficient to reduce patient quality of life on at least one of the five dimensions in the EQ-5D instrument. Therefore, the

---

[1] It does not deal with the lower limiting value although it can easily be modified to do so. In our dataset, we have very few observations at the bottom end of the distribution so we make no attempt to deal with this issue.

standard Tobit model is not sufficient since it deals with the upper censoring but not with the gap between one and the next feasible value. We modify the Tobit model so that the concentration of observations at 1 is accompanied by a gap to the next set of observations at 0.883. The peak at 1 is therefore comprised both of the gap to the left and the censoring to the right. the model can be written as follows:

$$
y_{it} = \begin{cases} 1 & \text{if } y^*_{it} > 0.0 \\ y^*_{it} & \text{otherwise} \end{cases}
$$

$$
y^*_{it} = x_{it}\beta_i + E_{it}
$$

$$
\beta_{ki} = {}_i a_k + u_{ki}
$$

An example of a distribution generated by this type of model can be seen in the lower panel of Figure 3.

Model 4: Random effects Adjusted Censored Mixture Model (ACMM). The last feature of the EQ-5D data that none of the previous models deal with is the multi-modality of the distribution and possible departures from normality across the rest of the distribution. This feature can be the result of unobserved heterogeneity in the form of latent classes. Intuitively, the population may be made up of several groups, or "latent classes", with potentially different relationships to the dependent variable. These models can be estimated by using mixtures (McLachlan and Peel (2000)). Mixtures are very flexible and can well accommodate the statistical challenges posed by typical EQ-5D distributions.

Conditional on an observation belonging to class C_{it}, the model becomes:

$$y_{it\backslash C_{it}} = \begin{cases} 1 & \text{if } y^*_{it\backslash C_{it}} > 0 \\ y^*_{it\backslash C_{it}} & \text{otherwise} \end{cases}$$

$$y^*_{it\backslash C_{it}} = x_{it}\beta_{ic} + \varepsilon_{itc}$$

$$\beta_{kic} = a_{kc} + u_{ki}$$

We assume a multinomial logit model for the probability of latent class membership:

$$P\left(C_{it} = c\backslash w_{it}\right) = \frac{\exp\left(w^t \theta_c\right)}{\sum_{s=1}^{P} \exp\left(w^t_{it}\theta_s\right)}$$

where w^t_{it} is a vector of variables that affect the probability of class membership, θ_c is the vector of corresponding coefficients and P is the number of classes. Note that it is possible when estimating this model (and also a general latent class model) to find that the mean of one of the classes is one, irrespective of the values of the covariates, and its variance tends to zero. In this case, if the optimal number of classes is two, the resultant model is analogous to the two part or hurdle model.

Whilst we apply adjusted censored normal distributions for each of the latent classes, as the mean of the distribution moves away from the censoring point and/or the variance decreases, so the distribution tends to that of the normal.

Some judgment must be used in determining the appropriate number of latent classes since the usual likelihood ratio test cannot be used to test nested latent class models. Some of the parameters (the variances of the latent classes) are on the boundary of the parameter space which distorts the distribution of the statistic and thus the usual test cannot be applied. The Bayesian Information Criteria (BIC) is a recommended, good indicator of the appropriate number of classes as well as plots of the likelihood values for models with different classes to identify a flattening of the likelihood values. This indicates where the addition of further latent classes does not

improve the likelihood substantially (Nylund et al. (2007)). It is worth noting that as the number of classes is increased, the model can be viewed as semiparametric, a midpoint between a fully parametric model with a single latent class (or mixture component) and a nonparametric model in the case where the number of components equals the sample size. If the aim is to achieve the best fit possible, increasing the number of classes based only on BIC for example might be a good idea although in this case the classes might loose their meaning. If the aim is however to fit a model where the classes have a substantive meaning so that it can be used for out of sample predictions then a compromise between the BIC and consideration of the size and differences between the latent classes is needed to prevent the inclusion of latent classes with a very small size including perhaps only a small number of outliers. Figure 3 compares an example distribution generated by this model to models 1 to 3.

All these models can be estimated using maximum likelihood. Robust standard errors using a sandwich estimator are used for all the models to protect against non-normality. All analyses were undertaken using the Mplus programme, Muth€n and Muth€n (2008), except for those based on adjusted censored models which were programmed in GAUSS, GAUSS (2008). The problems of estimating mixture models are well documented in the literature due to multiple optima of the likelihood function. Using only one run of the usual local optimisation algorithms typically leads to finding only a local maximum. To overcome this problem, Mplus uses a large set of random starting values for a few iterations of an Expectation Maximisation algorithm before selecting a few promising values to optimise fully. The parameter values that achieve the highest likelihood are then selected as the global maximum. For the adjusted censored models we used a global optimisation algorithm, simulated annealing (Corana (1987); Goffe et al. (1994)) to obtain a starting value near the final solution. A stopping rule was applied to ensure that the function was in the vicinity of a global maximum. We then switched to a local maximisation algorithm

for the final optimisation stages (see Hernandez Alava (2002) for the first application of simulated annealing to the optimisation of mixture models that we are aware of). Scripts of the codes and details of results for all models that were run are available from the authors on request.

There is a tendency for models to be compared in terms of their goodness of fit. Typically R^2 or adjusted R^2 are used but there are various other measures of "error" that are widely reported and are used to choose between models, for example, the Mean Error (ME), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Such summary measures provide an indication of the magnitude of the difference between the observed and predicted values but are of limited value for model comparisons (see Charemza and Deadman (1992) for a discussion of the dangers of using $R^2/\bar{R}^2$). Particularly in the situation where we wish to use such estimates in decision models that cover a large spectrum of disease, that is, predicting across a wide EQ-5D range, it is essential that no systematic bias is introduced and that predictions lie in the feasible range. Summary measures of model prediction do not provide a basis for choosing between models in this situation.

In what follows, we compare models using a range of plots and criteria. Model fit is described using penalised likelihood criteria (Akaike Information Criteria (AIC) and BIC).

4   Results

We initially compared linear and Tobit models using only HAQ and patient demographic variables age and sex (age is centred at its sample mean and measured in ten year units). We found that the inclusion of HAQ^2, thereby allowing a non linear relationship with EQ-5D was warranted. Similarly, the inclusion of age^2 improved the models. In all models we found a positive association between EQ-5D and age, conditional on HAQ. This is consistent with previous literature both in RA (Marra

Figure 3: Illustrative histograms of possible model distributions
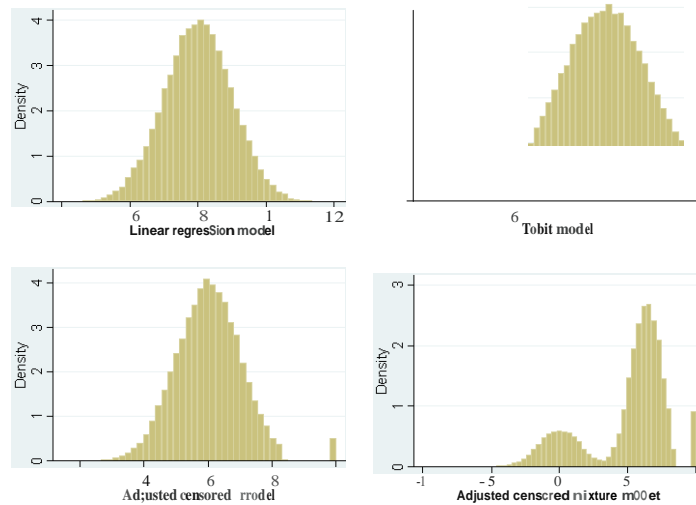
et al. (2007), Wailoo et al. (2008)) and other conditions (see for example Goldsmith et al. (2010)). We included the follow up times as separate covariates to test for any effect of disease duration but these were grossly insignificant and not retained in any of the models. The preferred models are a random effects linear regression and a Tobit model with two independent random effects for the intercept and the coefficient of HAQ. In identifying the preferred specification of the models, we considered a range of options. These included a standard Tobit with no random effects, one, two and three random effects (for the intercept and the coefficients of HAQ and HAQ^2), both considered as either independent or correlated, and the inclusion of an inflation factor for the censoring point. The preferred Tobit specification includes two independent random effects and achieved the lowest information criteria compared to the alternative Tobit specifications. The linear regression has a lower AIC (-624.0 vs 32.1) and BIC (-579.2 vs 99.4) than the Tobit model. This is due to the fact that the observations include a substantial peak at EQ-5D scores of about 0.5, thereby pulling down the estimates for the linear regression and causing the observations around one to have less influence. Importantly, within the sample the linear regression does not predict values exceeding unit y.

17

The inclusion of pain measured on a Visual Analogue Scale (VAS) vastly improved the models, results for which are shown in Table 1. Pain is one of the most heavily weighted items in the EQ-5D valuation regression but does not feature in the HAQ summary score. Previous studies focus on function but the trial data provided an opportunity to include this missing dimension of quality of life in the analysis. For the linear regression the AIC reduced to -1058.2 and the BIC reduced to -1007.7. With the inclusion of pain, the preferred specification of the Tobit model required only one random effect, a random intercept. The AIC reduced to -345.6 and the BIC reduced to -295.1. Thus, the apparent heterogeneity in the coefficient of HAQ may be explained by the omission of the pain covariate. The linear regression still outperforms the Tobit model in terms of model fit. In addition to AIC and BIC, the ME, MAE and RMSE are approximately equivalent on the utility scale. However, there are other aspects of the performance of the linear regression that warrant consideration. Most importantly, the predicted values can exceed one. Within the current sample this is not the case, as with the models excluding pain, but when predicting out of sample, unfeasible predictions will be generated. Furthermore, the predicted values from the linear regression do not reflect the characteristics of the underlying data in other ways. In particular, the mass of observations at one but also the gaps and multi-modality of the distribution are not well reflected by this model as demonstrated in Figure 4 by the histogram of the predicted values.

The RE Tobit model addresses the issue of censoring and Figure 4 illustrates that there is a resultant peak in values close to one. The RE adjusted censored model develops this further to address the gap between full health and all intermediate health states. Table 1 shows that model fit, measured by the information criteria, is improved substantially compared to the unadjusted RE Tobit model. More importantly, the distribution of predicted values illustrate that the model captures the key feature of the data at the top of the distribution. The difference is slight when considering the predicted values but there is a more pronounced rise in the

density at higher values of HAQ. Note however that the distribution of expected values will not itself demonstrate the peak at one with the subsequent gap since this plot smooths out such peaks. This is a feature of the underlying adjusted censored model, not the resultant expected values. The predicted values are averages across individuals so although no individual will have a value between 1 and 0.883, the predicted values can lie in this range.

We considered a variety of mixture models. The models varied in terms of the underlying distributions that comprised the mixture. We considered mixtures of normal, Tobit and adjusted censored models. It was considered that the models that included three latent classes were preferred. We chose three latent classes considering mainly the BIC balanced against consideration of the size of the latent classes. There is a danger in the inclusion of an excessive number of latent classes where a class may include only a small number of outliers. This phenomenon is clearly observed when we move from a model with three latent classes to a model with four. The model for the four latent classes essentially splits one of the latent classes of the three class model, the class at the highest levels of HAQ. One of these two classes is quite substantial in size but the other class is not. In fact we find that only 13 observations out of the 2003 in our sample are most likely to be in this class. In addition, the increase in the likelihood value is not as substantial as when the classes are increased from two to three also signalling that this latent class is not substantial enough and might just contain a few outliers. Initially the model with three latent classes contained a HAQ^2 term in all the classes. However, it was found that the estimated coefficients for HAQ and HAQ^2 in latent class 1 were insignificant but very highly correlated and therefore only a linear term was needed in this class.

Table 2 demonstrates that this further development produces a model that vastly outperforms the non mixture models. The information criteria are both substantially lower than the other models. The average errors are smaller on all measures and, although the scale of EQ-5D can mask these differences, we see an approximate 4%

improvement in MAE and 1% in RMSE compared to the standard RE linear model.

The model identifies three separate latent classes which demonstrate different relationships between EQ-5D, function and pain. Table 3 shows the summary characteristics of the observations when grouped by the latent class with the most probable membership. The groups are clearly different in terms of functional disability and pain. Class 1 has a mean HAQ of 2.26 indicating substantial disability, and a high pain score mean of 72.1. This reflects the element of the data distribution at the bottom of the EQ-5D scale. Class 2 has the least disability as measured by HAQ and the least amount of pain. The third class has moderate pain (mean 33.3) and functional disability (mean 1.24 HAQ). Classes 2 and 3 are both required in order to reflect the concentration of observations at one and the peak in the skewed distribution below this. It is the mixture of two latent classes with different means and variances that reproduces this skew.

Table 2 shows that for latent class 1, only a linear term on HAQ is needed and the coefficient for pain is both large and significant. Latent class 2 has a non significant coefficient for pain. Both HAQ and HAQ^2 are significant and relatively large, the latter negatively so. This results in a U- shaped relationship between HAQ and the latent variable, where the latter is predicted to increase once the HAQ functional ability deteriorates beyond around 2. The reason for this is that the model is trying to incorporate the small number of outliers that occur at the highest HAQ value where we see very few observations. For the third latent class, both HAQ and pain covariates are significant. The quadratic HAQ coefficients result in a deterioration in the latent variable as function decreases, but at a decreasing rate within the feasible range. The size of the coefficients and their implications cannot be judged directly since this is a highly nonlinear model. Table 4 presents examples of the predicted values for selected combinations of covariates. For a female of average age in the sample and both HAQ and pain of zero the predicted EQ-5D is 0.94. We see that the expected values for each class range from 0.34 for class 1 to 0.98 for class

2. The predicted probabilities of latent class membership are very different. This individual has a zero probability of being in class 1 and a high predicted probability (0.77) of being in class 2, the class with the highest expected EQ-5D value. Thus the predicted EQ-5D value lies in between that of class 2 and class 3. As pain increases, the predicted EQ-5D value goes down, but the relationship is not linear (unlike the linear regression model which implies that a change in pain will always lead to the same change in the predicted EQ-5D regardles of the levels of pain). The expected values for each of the classes decrease and the probabilities change dramatically. When pain is very high (93) the probability of class 3 membership is 0.91. Also as HAQ increases, we see a decrease in the predicted EQ-5D. The impact of gender is relatively small and varies in magnitude depending on the values of other covariates. Males have consistently lower predicted values of EQ-5D. The differences are much smaller than the linear regression which estimates a constant difference of 0.05.

The improvement in model fit gained from the mixture model approach is very noticeable at the lower end of the HAQ scale, that is, where patients have the least functional disability. A recognised issue in modelling health utility data arises in the relatively poor model fit at the extremes of the health profile. Figure 5 illustrates that the mixture model fits extremely well between HAQ scores of 0 to 1, unlike all the other non-mixture models which underpredict at HAQ of 0 (no functional disability) and systematically overpredict at HAQ scores between 0.3 and 1. Interestingly, we found that a mixture of normal distributions (not reported here) consistently underpredicts at HAQ between 0.15 and 0.9 suggesting that it is the combination of mixture modelling and the ACM that is required. Table 5 presents the mean error, MAE and RMSE of all the models at three different intervals of HAQ. The ACMM always outperforms the other models in terms of mean error and MAE. It also outperforms the other models in terms of RMSE between HAQ scores of 0 to 1 and 1 to 2. The RMSE of the ACMM is only worse between HAQ scores of 2 to 3. Further investigation identified that this lower predictive ability of

Figure 4: Histograms of predicted values
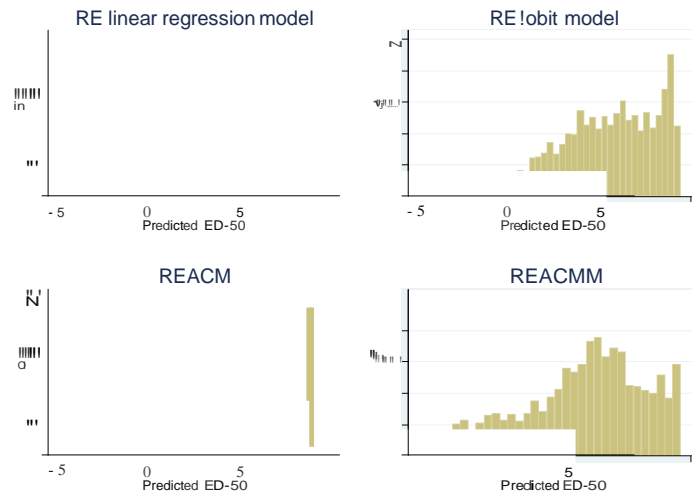
the RE ACMM, as judged by the RMSE, is only for the last 3 HAQ scores - 2.75, 2.875 and 3. At these levels of functional disability the dataset has an extremely low number of observations which appear very different and may be outliers. Up to and including a HAQ score of 2.625, the ACMM still outperforms the linear model with a RMSE for the interval (2-2.625) of 0.2425 compared to a RMSE for the linear model of 0.2427. The ACMM is more flexible than the other models and is therefore sensitive to these extreme observations at the tails of the distribution.

Figure 5: Observed and predicted mean EQ-5D by HAQ (0 to1)
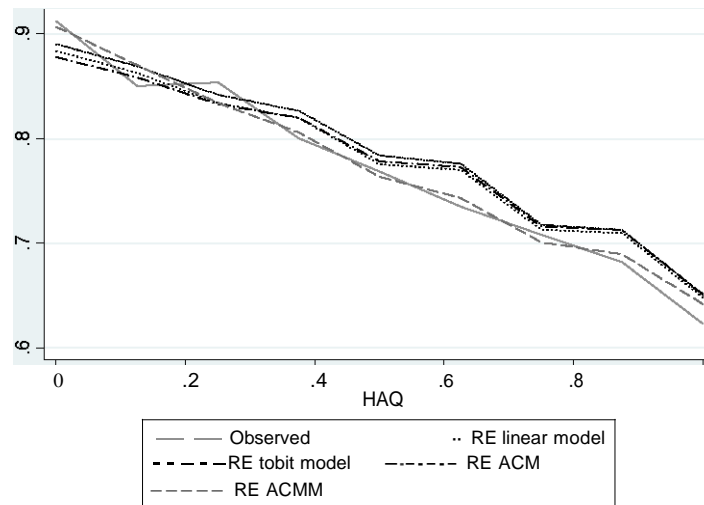
Table 1: Parameter estimates (se) for linear, Tobit and adjusted censored models of EQ-5D

		RE Linear Model	RE Tobit	REAdjusted Censored Model
Within	HAQ	-0.084 (0.020)	-0.165 (0.025)	-0.115 (0.023)
level	HAQ^2	-0.045 (0.008)	-0.022 (0.009)	-0.036 (0.009)
	Vaspain/100	-0.478 (0.027)	-0.499 (0.028)	-0.484 (0.028)
	(J2$_{E:}$	0.028 (0.001)	0.032 (0.002)	0.030 (0.002)
Between	Intercept	0.941 (0.012)	1.013 (0.017)	0.967 (0.015)
level	$\dfrac{\text{Age-54.32}}{10}$	0.019 (0.005)	0.018 (0.005)	0.019 (0.005)
	(Age; 4.32)^2	0.005 (0.003)	0.007 (0.003)	0.006 (0.003)
	Male	-0.046 (0.013)	-0.047 (0.014)	-0.047 (0.014)
	(J2$_u$	0.010 (0.002)	0.012 (0.002)	0.011 (0.002)
AIC		-1058.17	-345.55	-558.84
BIC		-1007.75	-295.13	-508.42
ME (sd)		0.0003(0.194)	$-0.0001(0.193)$	0.0005(0.194)
MAE (sd)		0.1505(0.122)	0.1508(0.121)	0.1508(0.121)
RMSE		0.1935	0.1934	0.1935

AIC - Akaike Information Criteria, BIC - Bayesian Information Criteria,

ME + Mean Error, MAE - Mean Absolute Error, RMSE+ Root Mean Squared Error

Table 2: Parameter estimates (se) for ACMM model of EQ-5D

Within	Latent class 1	HAQ	-0.0062 (0.015)
level		HAQ^2	$-$ $(-)$
		VASpain/100	-0.0295 (0.030)
		θ^2_1	0.0015 (0.002)
	Latent class 2	HAQ	-0.0245 (0.044)
		HAQ^2	0.006 (0.019)
		VASpain/100	-0.0105 (0.134)
		θ^2_2	0.0006 (0.001)
	Latent class 3	HAQ	-0.0160 (0.013)
		HAQ^2	0.0025 (0.005)
		VASpain/100	-0.0056 (0.018)
		θ^2_3	0.000 (.000)
Between	Latent class 1	Intercept	0.04 (0.037)
level	Latent class 2	Intercept	0.0990 (.025)
	Latent class 3	Intercept	0.06 (0.011)
	All classes	$\frac{Age-54.32}{10}$	0.0007 (0.002)
		$\frac{Age-54.32}{10}^2$	0.0004 (0.001)
		Male	-0.0012 (0.006)
		θ^2_u	0.0002 (0.000)
Within	Latent class 1	Intercept	-5.0201 (0.423)
level		HAQ	2.6 (0.178)
categorical		VASpain/100	5.0179 (0.433)
latent	Latent class 2	Intercept	2.020 (0.312)
variables		HAQ	0.045 (0.214)
		VASpain/100	-11.066 (4.227)
AIC		-2051.011	
BIC		-1911.005	
ME (sd)		-0.0000 (0.0192)	
MAE (sd) [% improvement]		0.014 (0.012) [4%]	
RMSE [% improvement]		0.0192 [1%]	

AIC - Akaike Information Criteria, BIC - Bayesian Information Criteria,

ME - Mean Error, MAE - Mean Absolute Error, RMSE- Root Mean Squared Error

Table 3: Patient characteristics by most
likely latent class (std. dev.)

Variable	Class 1	Class 2	Class 3
HAQ	2026(00 5)	00 7(0046)	1024(0060)
VASpain	7201(1406)	00(205)	0 (1906)
Age (yrs)	5405(1204)	5 0 (110)	5402(1206)
Male	0025(0044)	00 9(0049)	00 2(0046)

Table 4: Predicted EQ-5D values for the RE ACMM.

		Male, age 54						
HAQ	VASpain	EQ5D	$EQ5D_1$	$EQ5D_2$	$EQ5D_2$	P_1	P_2	P_3
0	0	0.9430	0.3426	0.9814	0.8165	0.0005	0.7687	0.2308
0	52	0.7724	0.1890	0.9520	0.7818	0.0174	0.0052	0.9773
0	93	0.6942	0.0679	0.9153	0.7561	0.0899	0.0000	0.9100
1	0	0.7716	0.2806	0.8257	0.6704	0.0044	0.6626	0.3331
1	52	0.5884	0.1270	0.7633	0.6410	0.1032	0.0029	0.8940
1	93	0.3790	0.0058	0.7168	0.6179	0.3903	0.0000	0.6097
2	0	0.6711	0.2185	0.7760	0.5848	0.0374	0.5227	0.4399
2	52	0.3461	0.0649	0.7166	0.5555	0.4272	0.0011	0.5717
2	93	0.0582	-0.0562	0.6721	0.5324	0.8057	0.0000	0.1943
2.5	0	0.6299	0.1875	0.8094	0.5607	0.1014	0.4302	0.4684
2.5	52	0.2056	0.0339	0.7479	0.5314	0.6551	0.0005	0.3444
2.5	93	-0.0357	-0.0872	0.7022	0.5083	0.9134	0.0000	0.0866
		Female, age 54						
HAQ	VASpain	EQ5D	$EQ5D_1$	$EQ5D_2$	$EQ5D_2$	P_1	P_2	P_3
0	0	0.9360	0.3307	0.9766	0.8021	0.0005	0.7687	0.2308
0	52	0.7590	0.1771	0.9430	0.7683	0.0174	0.0052	0.9773
0	93	0.6814	0.0560	0.9033	0.7432	0.0899	0.0000	0.9100
1	0	0.7583	0.2686	0.8118	0.6584	0.0044	0.6626	0.3331
1	52	0.5764	0.1150	0.7502	0.6291	0.1032	0.0029	0.8940
1	93	0.3671	-0.0061	0.7043	0.6060	0.3903	0.0000	0.6097
2	0	0.6584	0.2066	0.7627	0.5729	0.0374	0.5227	0.4399
2	52	0.3342	0.0530	0.7041	0.5436	0.4272	0.0011	0.5717
2	93	0.0463	-0.0681	0.6600	0.5205	0.8057	0.0000	0.1943
2.5	0	0.6171	0.1756	0.7957	0.5488	0.1014	0.4302	0.4684
2.5	52	0.1937	0.0220	0.7350	0.5195	0.6551	0.0005	0.3444
2.5	93	-0.0476	-0.0991	0.6898	0.4964	0.9134	0.0000	0.0866

EQ5D predicted EQ-5D, $EQ5D_s$ predicted EQ-5D for class s
P_s probability of class s membership

Table 5: Measures of accuracy of predictions by HAQ interval.

| | HAQ scores [0, l) (n − 75) | | | |
	RE linear model	RE Tobit	RE ACM	RE ACMM
ME	-0.0055	-0.0ll3	-0.0053	-0.0025
MAE	0.l032	0.l034	0.l038	0.0955
RMSE	0.l366	0.l363	0.l367	0.l328
	HAQ scores (l,2) (n − 91)			
	RE linear model	RE Tobit	RE ACM	RE ACM
ME	0.0040	0.0089	0.0048	0.00l9
MAE	0.l670	0.l675	0.l67l	0.l603
RMSE	0.2l09	0.2l08	0.2l09	0.2096
	HAQ scores (2,3) (n − 54)			
	RE linear model	RE Tobit	RE ACM	RE ACM
ME	0.0032	0.00ll	0.0022	-0.00l2
MAE	0.2l0l	0.2l02	0.2l02	0.2057
RMSE	0.2447	0.2449	0.2447	0.2467

ME - Mean Error, MAE - Mean Absolute Error, RMSE- Root Mean Squared Error

5   Discussion

Health related quality of life data typically exhibit distributional properties that raise numerous statistical challenges. In some respects these are more complex than those that arise in relation to healthcare costs, where there has been substantial attention given to the development and application of flexible statistical models to deal with issues such as repeated measures, skewness and left censoring (Hernandez Alava and Wailoo (20l0)).

Whilst standard models and methods for dealing with censoring have been applied when modelling health state utilities, these offer only limited and partial solutions to these challenges. A common feature arising from the limitations of such models is poor fit at the extremes of the distribution. The commonly used Tobit and other censored regression models offer a method for dealing with the upper bound of full health in health utility data. We have developed a censored regression model that provides a more appropriate method to reflect the gap between full health and intermediate health states, a particular feature that results from the approach used

to generate the EQ-5D tariff. Only limited consideration has been given to the use of latent class mixture models in this field to date. The small number of studies that have considered the approach have focussed on the issue of ceiling effects, Huang et al. (2008), Pullenayegum et al. (2010).

When considering measures of model fit we use penalised likelihood measures. When we compared the models using data from a clinical trial of patients with rheumatoid arthritis, we found that the linear regression outperforms either of the censored regression models. This is because our dataset exhibits a relatively large peak of observations around an EQ-5D value of 0.5. This pulls the estimates from the linear regression down and diminishes the importance of the values at high EQ-5D scores, where differences betwen the EQ-5D and censored models will be more profound. The linear regression does not predict unfeasible values within the observed dataset but may do so in a different sample where patients exhibit different characteristics to those included in this trial. This is of critical importance when considering the intended use of such models in cost effectiveness analyses. Here it is typical to simulate patients with varying characteristics and over long time periods. These models can be expected to cover a wide range of functional disability, pain and other relevant patient characteristics such that it is likely that implausible predicted values will be generated. In this situation the analyst may need to artificially censor the predicted values themselves. Thus, there are clear dangers with a reliance on model fit alone in model selection in this situation. The adjusted censored model improves both model fit and provides a more faithful reflection of the underlying data compared to standard censored models such as the Tobit model.

This paper develops the approach further by employing mixture models. Mixture models offer a highly flexible tool that can be used to deal with the remaining distributional challenges. Each of the mixture models we considered offered vastly superior performance compared to standard linear regression and censored regression models. In particular, we demonstrate that this model predicts accurately at the

highest EQ-5D and maintains this high degree of predictive accuracy across the EQ-5D range. Only where data is sparse does this accuracy decrease. This is the first application of censored mixture models in this area that we are aware of.

Our preferred model specification identifies three latent classes clearly distinguished by the relative role of functional disability and pain in determining EQ-5D utility values. The model is formed as a mixture of adjusted censored distributions.

Our findings also have specific implications for cost effectiveness modelling of interventions in rheumatoid arthritis. First, it is clear that estimates of health state utilities are improved by the inclusion of pain and other patient level covariates rather than functional ability alone. Since pain is not a feature of the HAQ score but is a heavily weighted component of the EQ-5D tariff the finding is perhaps not surprising. Standard composite outcome measures used in RA clinical trials and observational datsets, such as the American College of Rheumatology (ACR) response criteria, include VAS pain as one of the components. Where treatment effects are observed on pain as well as function, appropriate statistical models to estimate health state utilities become critical to avoid biased estimates of cost effectiveness. In order for researchers to make use of the estimates in the proposed mixture models and the associated uncertainty, the full variance covariance matrix is available from the following website (www.sheffield.ac.uk/scharr/sections/heds/dps-2010.html) and the expression for the predicted values in the appendix.

There are several potential limitations. The approach treats the values from the EQ-5D tariff as if they were data rather than estimates. The uncertainty in the original regression work reported by Dolan et al. (1995) is ignored. Furthermore, the challenges that arise from the distributional characteristics of the EQ-5D tariff may be avoidable. The method by which EQ-5D values are generated is based on simple linear regression model that itself does not perform well and does not apply a consistent approach to the values assigned to full health versus all other intermediate health states. The application of more flexible models to generate

the EQ-5D tariff may result in less statistical challenges for analysts that need to estimate the relationship with clinical and sociodemographic variables. We note that health state utilities generated from the EQ-5D in different populations exhibit the same distributional features as the data presented here, see for example Huang et al. (2008). Nevertheless, there may also be value in examining the performance of the mixture of adjusted censored features of distributions for health utility data generated from instruments other than the EQ-5D.

The dataset on which the analyses are based is from a group of patients with early RA at the point of entry to the study. Despite the fact that the estimates span two years of follow up and we found no evidence of a time trend, it may be the case that patients with more established disease do not exhibit the same relationships between EQ-5D, pain and function. For example, there may be a greater degree of adaptation to functional decline in later disease that is reflected to a greater degree in HAQ than EQ-5D. In addition, the dataset includes few observations at the most extreme level of functional disability, with outliers perhaps exerting a strong influence on the model. Therefore, further refinement of the ACMM in an additional dataset of established disease may be useful.

In summary, it is clear that the mixture modelling approach provides a general framework which can reflect the specific distributional characteristics of health utility data. When combined with the adjusted censored distribution it is possible to obtain a flexible model that vastly outperforms standard linear regression and censored regression approaches.

References

Aletaha, D., Landewe, R., Karonitsch, T., et al. (2008) Reporting disease activity in clinical trials of patients with rheumatoid arthritis: EULAR/ACR collaborative, Annals fo the Rheumatic Diseases, Vol.67:1360-1364.

Austin, P.C. (2002) A comparison of methods for analyzing health-related quality of life measures, Value in Health., Vol.5:329-337.

Austin, P.C., Escobar, M. and Kopec, J.A. (2000) The use of the Tobit model for analyzing measures of health status, Quality of Life Research, Vol.9:901-910.

Bansback, N.J., Brennan, A., and Ghatnekar, O. (2005) Cost effectiveness of adalimumab in the treatment of patients with moderate to severe rheumatoid arthritis in Sweden, Annals of the Rheumatic Diseases, Vol.64:995-1002

Bansback, N.J., Marra, C., Tsuchiya, A., et al. (2007) Using the Health Assessment Questionnaire to Estimate Preference-Based Single Indices in Patients With Rheumatoid Arthritis, Arthritis and Rheumatism (Arthritis Care and Research), Vol.57:963-971.

Brazier, JE., Yang, Y., Tsuchiya, A. and Rowen, D.L. (2009) A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. European Journal of Health Economics, Vol.11:215–225.

Charemza, W.W. and Deadman, D.F. (1997) New Directions in Econometric Practice: General to Specific Modelling, Cointegration and Vector Autoregression, Edward Elgar: Aldershot.

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.

Choy, E.H.S., Smith, C.M., Farewell, V., et al. (2008) Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis, Annals of the Rheumatic Diseases, Vol.67:656-663.

Cooper, N. (2000) Economic burden of rheumatoid arthritis: a systematic review, Rheumatology, Vol.39:28-33.

Corana, A., Marchesi, C., Martini, C., and Ridella, S. (1987) Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm, ACM Transactions on Mathematical Software, Vol.13:262-280.

Crott, R. and Briggs, A. (2010) Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences, European Journal of Health Economics, Vol.11:427-434.

Dolan, P., Gudex, C., Kind, P., and Williams, A. (1995) A Social Tariff for EuroQol: Results from a UK Population Survey. University of York, Centre for health Economics, Discussion Paper 138.

GAUSS (2008) Mathematical and Statistical System v9.0. Aptech Systems Inc.

Goldsmith, K.A., Dyer, M.T., Buxton, M.J., and Sharples, L.D. (2010) Mapping of the EQ-5D index from clinical outcome measures and demographic variables in patients with coronary heart disease, Health and Quality of Life Outcomes, Vol.8:13

Goffe, N.L., Ferrier, G.D., and Rogers, J. (1994) Global optimization of statistical functions with simulated annealing, Journal of Econometrics, Vol.60:65-99.

Hawthorne, G., Buchbinder, R., and Defina, J. (2000) Functional Status and Health-related Quality of Life Assessment in Patients with Rheumatoid Arthritis, Monash University Centre for Health Program Evaluation, Working Paper 116.

Hernandez Alava, M. (2002) Growth dynamics: an empirical investigation of output growth using international data. PhD thesis Department of Economics. University of Leicester.

Hernandez Alava, M. and Wailoo, A.J. (2010) Multilevel modelling of cost data: an application to thrombolysis and primary angioplasty in the UK NHS.A multi-level modelling approach to analysis of patient costs under managed care. Health Economics and Decision Science, University of Sheffield Discussion paper 10/06

Huang, I., Frangakis, C., Atinson, M.J., et al. (2008) Addressing ceiling effects in health status measures: A comparison of techniques applied to measures for people with HIV disease, Health Services Research, Vol.43:327-339

Hurst, N.P., Kind, P., Ruta, D., et al. (1997) Measuring health related quality of life in rheumatoid arthritis: validity , responsiveness and reliability of EuroQol, British Journal of Rheumatology, Vol. 36:551-559.

Kobelt, G., Jonsson, L., Lindgren, P. et al. (2002) Modeling the Progression of Rheumatoid Arthritis. A Two-Country Model to Estimate Costs and Consequences of Rheumatoid Arthritis, Arthritis and Rheumatism, Vol.46: 2310-2319.

Li, L. and Fu, A.Z. (2009) Some methodological issues with the analysis of preference based EQ5D index score, Health Service Outcomes Research Methods, Vol.9:162-176.

Lindgren, P., Geborek, P., and Kobelt, G. (2009) Modeling the cost-effectiveness of treatment of rheumatoid arthritis with rituximab using registry data from Southern Sweden, International Journal of Health Technology Assessment, Vol.25:181-189.

Malottki, K., Barton, P., Tsourapas, A., et al. (2009) Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor: a systematic review and economic evaluation, NICE Technology Assesment Report, available at http://www.nice.org.uk/nicemedia/live/12135/46677/46677.pdf (accesed 14th July 2010)

Marra, C.A., Marion, S.A., Guh, D.P., et al. (2007) Not all "quality adjusted life years" are equal, Journal of Clinical Epidemiology, Vol.60:616-624.

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

Muthén, B. and Muthén, L. (2008) Mplus User's Guide. Los Angeles: Muthén and Muthén.

National Institute for Health and Clinical Excellence (2008) Guide to the Methods of Technology Appraisal

National Institute for Health and Clinical Excellence (2009) Rheumatoid Arthritis: The Management of Rheumatoid Arthritis in Adults. Clinical Guideline, http://www.nice.org.uk/Guidance/CG/Wave13/6

Nylund. K. L., Asparouhov, T. and Muthen, B. O. (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. Structural Equation Modelling, Vol:14, 535-569.

Pullenayegum, E.M., TArride, J., Xie, F., et al. (2010) Analysis of health utility data when some subjects attain the upper bound of 1: Are Tobit and CLAD models appropriate?, Value in Health, Vol.13: 487-494

Rowen, D., Brazier, J., and Roberts, J. (2009) "Mapping SF-36 onto the EQ-5D index: how reliable is the relationship?", Health and Quality of Life Outcomes 2009, 7:27

Shaw, J.W., Johnson, J.A., and Coons, S.J. (2005) US valuation of the EQ5D health states: Development and testing of the D1 valuation model, Medical Care, Vol.43: 203- 219.

Symmons, D., Turner, G., Webb, R., et al. (2002) The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century, Rheumatology, Vol.41:793-800.

Thompson, S. G., Nixon, R. and Grieve, R. (2006). Addressing the issues that arise in analysing multicentre cost data, with application to a multinational study, Journal of Health Economics, Vol.25:1015-1028.

Tobin, J. (1958) Estimation of relationships for limited dependent variables, Econometrica, Vol.26:24-36.

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.

Appendix

EQ-5D predictions using the random effects ACMM.

This appendix derives the expression for the prediction of EQ-5D using the ACMM model with a random intercept. Conditional on an observation belonging to class C_{it}, the random effects ACMM model can be written as:

$$y_{it\backslash C_{it}} = \begin{cases} 1 & \text{if } y^*_{it\backslash C_{it}} > 0 \\ y^*_{it\backslash C_{it}} & \text{otherwise} \end{cases}$$

$$y^*_{it\backslash C_{it}} = x_{it}\beta_{ic} + \varepsilon_{itc}$$

$$\beta_{lic} = z_i a_c + u_i$$

where β_i is a $(k \times 1)$ vector of coefficients β_{ki}, x_{it} is a row vector of level 1 covariates, z_i is a row vector of level 2 covariates, ε_{it} is IID $N(0, \sigma^2)$, u_i is IID $N(0, \sigma_u^2)$ and ε_{it} is independent of u_i. A multinomial logit model for the probability of latent class membership is assumed as follows:

$$P(C_{it} = c \backslash w_{it}) = \frac{\exp(w^t_{it}\delta_c)}{\sum_{s=1}^{P} \exp(w^t_{it}\delta_s)}$$

where w^t_{it} is a vector of variables that affect the probability of class membership, δ_c is the vector of corresponding coefficients and P is the number of classes ($P = $ in the final model used in our analysis).

The EQ-5D prediction based on this model, $E\left(y_{it}\backslash x_{it}, z_i, w^t_i\right)$, is calculated using the following expression:

$$E\left(y_{it}\backslash x_{it}, z_i, w^t_i\right) = E_{u_i}\left[E\left(y_{it}\backslash x_{it}, z_i, u_i\right)\right]$$

$$= \frac{1}{\sigma_u} \phi\left(\frac{u_i}{\sigma_u}\right) E\left(y_{it}\backslash x_{it}, z_i, w^t_i, u_i\right) du_i$$

where $\phi()$ is the standard normal density function and

$$E\left(y_{it}\backslash x_{it}, z_i, w^t_i, u_i\right) = \sum_{c=1}^{C} \frac{\exp\left(w^t_i \delta_c\right)}{\sum_{s=1}^{C} \exp\left(w^t_i \delta_s\right)} \left\{ \left[1 - \Phi\left(\frac{W - x_{it}\beta_c - z_i a - u_i}{\sigma_c}\right)\right] + \right.$$

$$\left. \Phi\left(\frac{W - x_{it}\beta_c - z_i a - u_i}{\sigma_c}\right)\left[x_{it}\beta_c + z_i a + u_i - \sigma \frac{\phi\left(\frac{fl - x_{it}\beta_c - z_i a - u_i}{a_{c;c}}\right)}{\Phi\left(\frac{fl - x_{it}\beta_c - z_i a - u_i}{a_{c;c}}\right)}\right] \right\}$$

In the last equation $\Phi()$ is the standard cumulative normal and $W = 0$.

33