Capturing valuation study sampling uncertainty in the estimation of health state utility values using the EQ-5D-3L

Abstract

Objectives

Utility scores associated with preference based health-related quality of life instruments such as the EQ-5D-3L are reported as point estimates. In this study we develop methods for capturing the uncertainty associated with the valuation study of the UK EQ-5D-3L that arises from the variability inherent in the underlying data, which is tacitly ignored by point estimates. We derive a new tariff which properly accounts for this and assigns a specific closed-form distribution to the utility of each of the 243 health states of the EQ-5D-3L.

Methods

Using the UK EQ-5D-3L valuation study we use a Bayesian approach to obtain the posterior distributions of the derived utility scores. We construct a hierarchical model which accounts for model-misspecification and the responses of the survey participants to obtain MCMC samples from the posteriors. The posterior distributions are approximated by mixtures of Normal distributions under the Kullback–Leibler (KL) divergence as the criterion for the assessment of the approximation. We consider the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to estimate the parameters of the mixture distributions.

Results

We derive an MCMC sample of total size 4,000×243. No evidence of non-convergence is found. Our model is robust to changes in priors and starting values. The posterior utility distributions of the EQ-5D-3L states are summarised as three-component mixtures of Normal distributions and the corresponding KL divergence values are low.

Conclusions

Our method accounts for layers of uncertainty in valuation studies which is otherwise ignored. Our techniques can be applied to other instruments and countries' populations.

Keywords

Health state utility, uncertainty quantification, Bayesian methods, mixture of Normal distributions, economic evaluation.

Highlights

- Guidelines for health technology assessments typically require that uncertainty be accounted for in economic evaluations, but the parameter uncertainty of the regression model used in the valuation study of the health instrument is often tacitly ignored.
- We consider the UK valuation study of the EQ-5D-3L and construct a Bayesian model which accounts for layers of uncertainty which would otherwise be disregarded, and we derive closed-form utility distributions.
- The derived tariff can be used by researchers in economic evaluations, as it allows analysts to directly sample a utility value from its corresponding distribution, which reflects the associated uncertainty of the utility score.

Introduction

Utility scores derived from preference based health related quality of life (HRQoL) instruments are typically used to represent the value associated with health states in health economic evaluations. HQRL can be measured by health state instruments, such as the popular generic instrument EQ-5D-3L [1], which measures health related quality of life for 5 dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression), where each dimension is measured in terms of severity, with three levels. Specifically, each feasible combination of answers given by an individual to an instrument's questions corresponds to a certain health state, which is perceived to have a utility value reflecting the individual's HRQoL.

HRQoL instruments typically describe a large number of unique health states, based on the number of domains and levels of severity measured. The EQ-5D-3L has 3^5=243 unique health states , a combination of the three response levels of the five dimensions of the instrument, while the five-level version, the EQ-5D-5L [2] defines 5^5=3,125 health states. Health states utility values are then estimated using techniques including Time-Trade Off (TTO), Standard Gamble (SG) or Discrete Choice Experiments (DCE). In a typical valuation study [3] a sample of the general population is asked to perform a valuation task where a subset of the possible health states is directly assessed. For example, in the UK valuation study of the EQ-5D-3L [4] 2,997 respondents valued only 42 of 243 states. This gives a sample of directly valued states. Using these directly valued states, estimates of the values of states not directly valued sample are generated using frequentist regression techniques, generating a tariff of all states which is then presented as corresponding point estimates. Valuation studies using these methods have been conducted around the world in order to derive tariffs of utility scores corresponding to all attainable states of a health instrument, such as the EQ-5D-3L [4]. These tariffs are subsequently used in economic evaluations in order to estimate quality-adjusted life-years (QALYs), which are a compound measure of HRQoL and quantity of life lived.

These point estimates of mean health state utilities, although helpful to obtain, they do not contain information about the within-state variability; we believe that the information provided by a point estimate is not as useful as the description of a full probability distribution would be. Furthermore, while a point estimate can in general be considered more useful in cases when the data come from a symmetrical distribution such as the Normal distribution, in our case there are signs of skewness and sometimes of multimodality in directly valued states that warrant further analysis. For instance, Figure 1 illustrates the kernel density plot [5] of the utility scores which were assigned to state *11133* by participants of the UK valuation study of the EQ-5D-3L. Here we seek to understand the shape of the distribution of each health state index in order to acquire informative knowledge about the dispersion of state valuations by the population.

Spiegelhalter et al [6] , Baio [7] , Grieve et al [8] , and O'Hagan and Stevens [9] among many advocate the use of Bayesian methods as a useful tool in the area of health economic evaluation. Previous work on the USA EQ-5D-3L value set in a Bayesian context has demonstrated the importance of quantifying the uncertainty of the utility values [10]. Some researchers have worked on similar health instruments, such as the SF-6D [11], or on the EQ-5D-3L and EQ-5D-5L, with a focus on using the available valuation data in a more efficient way to produce value sets which are subject to less parameter uncertainty [12-16]. Nevertheless, the utility distributions of the 243 unique EQ-5D-3L states have not been quantified as explicitly specified probability distributions.

The aim of this study is to estimate the mean (over the general population) utility assigned to each state and derive an approach that captures the uncertainty arising from the sampling variation of the UK EQ-5D-3L scores by constructing a Bayesian model which assigns appropriate probability distributions to each of the EQ-5D-3L health states instead of point estimates. This allows researchers to model each EQ-5D-3L state (e.g. in sensitivity analyses) by using specified closed form probability functions without the need of making any further assumptions.

We build a hierarchical Bayesian model which accounts for the parameter uncertainty in the regression coefficients and also the uncertainty due to functional model-misspecification. Specifically, with the valuation study valuing directly only a subset of the total number of health states, the regression model is nonsaturated, meaning that the total number of regression parameters estimated is less than the total number of states valued by the respondents. Only a saturated model, which estimates as many regression parameters as the number of states valued, can be expected to fit the mean model correctly. Hence, consideration must be given to parameter uncertainty which exists due to using a nonsaturated functional form of the model.

The derived MCMC samples are then considered in order to approximate the posterior distributions as three-component mixtures of Normal distributions. Numerical optimisation is required for this task; we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm to derive the new tariff. Notably, although in this article we exhibit our techniques by using the EQ-5D-3L and the UK dataset, our methods can be applied to other health questionnaires and countries' datasets too.

Methods

UK EQ-5D-3L Valuation Study

The MVH group conducted the UK valuation study of the EQ-5D-3L by interviewing 3235 individuals[17-18]. The state of "perfect health" was assumed to have a utility of 1 and the state of "death" was assumed to have a utility of 0. Each respondent valued 12 other health states which are a combination of different levels of the five dimensions of the instrument, and in total across all respondents 42 health states were directly valued. Health states can be abbreviated by the use of a five-digit number where each digit corresponds to the severity level of each of the EQ-5D-3L dimensions. For example, *11113* is the abbreviation for the health state corresponding to no problems at all for the first four dimensions and severe problems for the fifth dimension. For valuations of states considered worse than death by the respondents, a transformation was used [19] so that all resulting scores lie between -1 and 1. After the application of stringent exclusion criteria [17], 2997 individuals contributed data to the valuation study. The aforementioned data of the MVH project can be obtained from the UK Data Service [20].

The regression model which was used for the computation of the frequentist tariff comprised 12 dummy variables: $ALL, N3, M2, M3, S2, S3, U2, U3, P2, P3, A2, A3$; it is also known as the "$N3$ model". The variable ALL takes value of 1 for any state other than that of "perfect health", whereas the variable $N3$ takes value of 1 if the state has at least one dimension in the third level of severity. $M2$ and $M3$ are dummy variables indicating whether mobility was at level 2 or 3 (respectively), and similarly for self-care ($S2, S3$), usual activities ($U2, U3$), pain/discomfort ($P2, P3$), and anxiety/depression ($A2, A3$). We take the core principles of the MVH approach into consideration, and we extend it in a Bayesian setting whilst also accounting for model-misspecification.

Specifying our Bayesian model

For the purposes of specifying our model, we define function $h(p, q)$, the codomain of which is $\{1,2, \dots ,242\}$, so that $h(p, q)$ equals the index of the EQ-5D-3L health corresponding to the q-th EQ-5D-3L state ($q = 1, \dots ,12$) which was evaluated by the p-th survey respondent ($p = 1, \dots ,2997$). The actual range of this function consists of the specific 42 indexes which were chosen to valued in the MVH project.

Our model can be written as:

$$1 - y_{pq} \sim N\left(\mu_{h(p,q)\cancel{pq}}, \sigma_{\varepsilon}\cancel{\sigma^{2}_{\varepsilon}}\right),$$

$$\mu_{h(p,q)\cancel{pq}} = \boldsymbol{X}^{T}_{h(p,q)\cancel{pq}}\boldsymbol{\beta} + \omega_p + \xi_{h(p,q)\cancel{\{q-th\ state\ valued\ by\ the\ p-th\ respondent\}}},$$

$$\begin{cases} \xi_{h(p,q)\cancel{\{q-th\ state\ valued\ by\ the\ p-th\ respondent\}}} = 0, & for\ \ "perfect\ health" \\ \xi_{h(p,q)\cancel{\{q-th\ state\ valued\ by\ the\ p-th\ respondent\}}} \sim N\left(0, \sigma_{\xi}\cancel{\sigma^{2}_{\xi}}\right), & otherwise \end{cases}$$

$$\omega_p \sim N(0, \sigma_{\omega}\cancel{\sigma^{2}_{\omega}}),$$

where $y_{pq}$is the utility value of the q-th EQ-5D-3L state ($q = 1, \dots, 12$) which was evaluated by the p-th survey respondent ($p = 1, \dots, 2997$), whereas ω_p is a subject-specific random effect and $\boldsymbol{X}^{T}_{pq}$ is a horizontal vector consisting of the following entries, which are defined as in the case of the $N3$ model: $ALL_{h(p,q)\cancel{pq}}$, $N3_{h(p,q)\cancel{pq}}$, $M2_{h(p,q)\cancel{pq}}$, $M3_{h(p,q)\cancel{pq}}$, $S2_{h(p,q)\cancel{pq}}$, $S3_{h(p,q)\cancel{pq}}$, $U2_{h(p,q)\cancel{pq}}$, $U3_{h(p,q)\cancel{pq}}$, $P2_{h(p,q)\cancel{pq}}$, $P3_{h(p,q)\cancel{pq}}$, $A2_{h(p,q)\cancel{pq}}$, $A3_{h(p,q)\cancel{pq}}$. It should also be stated that, throughout this paper, the parameters of the Normal distribution are its mean and standard deviation. Notably, ξ_s is related to the s-th distinct EQ-5D-3L health state, and it is the term which accounts for functional model-misspecification~~; each "$\{q - th\ state\ valued\ by\ the\ p - th\ respondent\}$" corresponds to a specific value of s (i.e. one of those 42 states which were valued by the respondents)~~. The importance of accounting for model-misspecification is discussed by Pullenayegum et al [10].

The computation of the utility score u_s of the s-th distinct EQ-5D-3L health state ($s = 1, \dots, 243$) is done as follows: $u_s = 1 - \boldsymbol{X}^{T}_{s}\boldsymbol{\beta} - \xi^{new}_{s}$, where the elements of the horizontal vector $\boldsymbol{X}^{T}_{s} = (ALL_s,\ N3_s, M2_s, M3_s, S2_s, S3_s, U2_s, U3_s, P2_s, P3_s, A2_s, A3_s)$ are the dummy variables as defined in the case of the $N3$ model. Furthermore, $\xi^{new}_{s} \sim N\left(0, \sigma_{\xi}\cancel{\sigma^{2}_{\xi}}\right)$, so that the mean of the utilities is $\boldsymbol{X}^{T}_{s}\boldsymbol{\beta}$, but with a ~~variance~~ standard deviation that better reflects the true uncertainty in the data.

In Bayesian analysis, many of the distributions which we attempt to compute are not analytically tractable. However, we can simulate the random variable and obtain a sample of values originating from that variable, using Markov Chain Monte Carlo (MCMC) techniques [21-24]. Using JAGS [25], MCMC simulations are obtained from the posterior

The following priors are used:

$$\beta_d \sim N(0{,}10), for\ d = 0{,}1,$$

$$\beta_d \sim N(0{,}1), for\ d = 2{,}3, \ldots ,11,$$

$$\sigma_\xi \sim U(0{,}1),$$

$$\sigma_\varepsilon \sim U(0{,}1),$$

$$\sigma_\omega \sim U(0{,}1).$$

Here, we use “minimally informative” uniform priors for the dispersion parameters, which, while stabilising the inference within a reasonable range of values, does not induce overly-strong reliance on prior assumptions. Furthermore, the β coefficients are centred around 0 to encode the assumption that initially we do not know the sign of each of these coefficients, even though in the absence of any logical inconsistencies each one of them will have a positive sign. The prior ~~variance~~ standard deviation of the β 's associated with all indicator variables except ALL and $N3$ is chosen to be 1 to reflect the prior uncertainty about the coefficients which are related to one dimension of the EQ-5D-3L. Conversely, the prior ~~variance~~ standard deviation of the coefficients of the ALL and $N3$ variables, which are related to multiple EQ-5D-3L dimensions is chosen to be 10 because of the underlying wider uncertainty. The values of the β coefficients are used to compute the EQ-5D-3L utility scores deterministically; since the values of the utility scores are expected to be between -1 and 1, the assigned prior distributions for the β coefficients do not provide strong prior information. Moreover, our priors are in agreement with those used in other research work related to EQ-5D-3L modelling [10].

Sensitivity analysis of our Bayesian model

We also use different priors to do sensitivity analysis and examine the robustness of the model. A different choice of priors for the dispersion σ parameters is as follows:

$$log\left(\sigma_{\xi}\right) \sim N(0,1000\cancel{10^6}),$$

$$log(\sigma_{\varepsilon}) \sim N(0,1000\cancel{10^6}),$$

$$log(\sigma_{\omega}) \sim N(0,1000\cancel{10^6}).$$

This time, Normal distributions are assigned to the natural logarithms of the σ parameters, where the corresponding standard deviations 1000~~10^6~~ are quite large. Alternatively, we also consider the following priors:

$$1/\sigma_{\xi} \sim Gamma(0.001,0.001),$$

$$1/\sigma_{\varepsilon} \sim Gamma(0.001,0.001),$$

$$1/\sigma_{\omega} \sim Gamma(0.001,0.001).$$

The motivation behind this is that the prior should be similar to the improper distribution 1/σ~Gamma(0,0), but the prior 1/σ~Gamma(0.001,0.001) actually favours small values of the standard deviation σ.

In terms of alternative priors for the β coefficients, we use different Normal distributions with even larger standard deviations, as follows:

$$\beta_d \sim N(0,1000\cancel{10^6}), for\ d = 0,1, \dots ,11.$$

These distributions are even less-informative compared to the original choice of prior distributions.

Approximation of the posterior distributions as mixtures of Normal distributions

An MCMC sample of size C can be obtained, for the posterior utility u_s of the s-th EQ-5D-3L state, coming from its posterior distribution $f_{u_s}(\cdot)$, the parametric form of which is not known directly, because our model is not a conjugate or simplistic one. We aspire to approximate these derived distributions with suitable parametric distributions $\widehat{f_{u_s}}(\cdot)$, and, due to evidence of multi-modality, it is reasonable to approximate them using mixtures of distributions which are expected to be capable of approximating the target distributions well. Specifically, the use of an algorithm for the approximation of multi-modal distributions with the use of standard probability distributions, such as normal, is a suitable solution to the problem of approximating the mixtures [26]. These distributions are approximated as mixtures of normal probability functions with Z finite components:

$$\widehat{f_{u_s}}(\cdot) = \sum_{z=1}^{Z} w_z \, N(\cdot \mid a_z, b_z).$$

The weights w_z's of the components as well as the components' means a_z's, and their corresponding standard deviations b_z's will have to be estimated separately. There is a trade-off between the quality of the approximation, which is increased by having further components, and the complexity of the algorithm as adding further components is associated with practical inconveniences, it brings further computational intensity as well as difficulties for the algorithm to reach convergence. Similar work of Schmidli et al [27] argued that a parsimonious and convenient approximation is a three-component $Z = 3$ mixture which satisfactory approximates the target distribution.

In order to attain a close approximation of the real distribution the Kullback–Leibler (KL) divergence is regarded, which is considered the standard measure in inference problems [28]. The

KL divergence between the target distribution $f_{u_s}(x)$, and the approximate (mixture) distribution $\widehat{f_{u_s}}(x)$ is defined as:

$$KL\left(f_{u_s}(x), \widehat{f_{u_s}}(x)\right) = \int log\left(f_{u_s}(x)\right) f_{u_s}(x)\, dx - \int log\left(\widehat{f_{u_s}}(x)\right) f_{u_s}(x)\, dx.$$

The lower the KL divergence (between the proposed and the actual distribution), the better the approximation. The ideal approximation (which theoretically is 0) is derived by selecting such weights and hyper-parameters (using numerical optimisation) to have a maximum in the second right term, i.e.: $\int log\left(\widehat{f_{u_s}}(x)\right) f_{u_s}(x)\, dx$. Using the MCMC sample $u_s^{(1)}, u_s^{(2)}, \dots, u_s^{(C)}$ generated from the posterior distribution of u_s, we can deduce the Monte-Carlo estimate of this integral as: $\frac{1}{C}\sum_{c=1}^{C} log\left(\widehat{f_{u_s}}\left(u_s^{(C)}\right)\right)$. Moreover, this is the same as the mean log-likelihood of the parameters of the mixture $\widehat{f_{u_s}}(\cdot)$ given the observed MCMC sample. Hence, this implies that in order to minimise the KL divergence, we have to maximise this log-likelihood.

The problem of finding a good approximate distribution is simplified to deriving the maximum likelihood estimates, because KL divergence is optimal when the weights and the hyper-parameters are equal to the maximum likelihood estimates. Having a multivariate case, numerical optimisation is required for the successful estimation of these parameters. We consider the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, which was proposed by Broyden [29], Fletcher [30], Goldfarb [31], and Shanno [32] independently; it is the most efficient of the quasi-Newton methods [33].

Note: ~~Regarding the~~Impact of correlation between states

In other words, once we obtain our MCMC sample of utility values of size $C \times 243$, we will use it and consider the BFGS algorithm to eventually summarise each health state utility distribution as a mixture of Normal distributions. Given the way our Bayesian model was defined, one would expect the existence of some correlation between different EQ-5D-3L state utilities; indeed, at the c-th iteration of MCMC, all the state utilities u_s are actually computed by using the same c-th instances of the β coefficients. Nevertheless, the use of the subsequently derived mixture distributions means that, in a sense, utilities of different health states are treated like they are independent. If no strong dependence exists here, though, then there is not much difference between using values sampled from the mixture distributions and those from the joint posterior of utility values.

The impact of this correlation can be examined by the consideration of a hypothetical two-arms randomised trial which runs for one year, at the end of which the EQ-5D-3L is administered. Let $\boldsymbol{\theta}_{\boldsymbol{j}}^{\boldsymbol{T}} = \left(\theta_{1j}, \dots, \theta_{Sj}\right)$ capture the probabilities of individuals of the j-th group of the trial (for $j = 1{,}2$) falling into each of the EQ-5D-3L health states, where is the probability of an individual in group failing into health state s (for $s = 1, \dots, 243$). Then, $\boldsymbol{\Delta}_{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\boldsymbol{1}} - \boldsymbol{\theta}_{\boldsymbol{2}}$ and let $\bar{\boldsymbol{u}}^{\boldsymbol{T}} = (\bar{u}_1, \dots, \bar{u}_S)$, where $\bar{u}_s$ is the mean utility for the s-th EQ-5D-3L state. Moreover, Pullenayegum et al [10] report that if Δ_e is the difference in mean QALYs between the groups, then we have: $\Delta_e = \boldsymbol{\Delta}_{\boldsymbol{\theta}}^{\boldsymbol{T}} \cdot \bar{\boldsymbol{u}}$. Furthermore, for the variance of Δ_e given $\boldsymbol{\Delta}_{\boldsymbol{\theta}}$ we have $var(\Delta_e | \boldsymbol{\Delta}_{\boldsymbol{\theta}}) = \boldsymbol{\Delta}_{\boldsymbol{\theta}}^{\boldsymbol{T}} \cdot var(\bar{\boldsymbol{u}}) \cdot \boldsymbol{\Delta}_{\boldsymbol{\theta}}$, where $var(\bar{\boldsymbol{u}})$ ~~can be~~is computed by using the matrix of posterior utility values of size $C \times 243$. Specifically, the aforementioned impact of the corelation ~~can be~~is examined by comparing VAR$_1$ and VAR$_2$ where VAR$_1$= $\boldsymbol{\Delta}_{\boldsymbol{\theta}}^{\boldsymbol{T}} \cdot var(\bar{\boldsymbol{u}}) \cdot \boldsymbol{\Delta}_{\boldsymbol{\theta}}$ and VAR$_1$= $\boldsymbol{\Delta}_{\boldsymbol{\theta}}^{\boldsymbol{T}} \cdot diag\left(var(\bar{\boldsymbol{u}})\right) \cdot \boldsymbol{\Delta}_{\boldsymbol{\theta}}$, while $diag\left(var(\bar{\boldsymbol{u}})\right)$ is a diagonal matrix with the same dimensions and diagonal entries as matrix $var(\bar{\boldsymbol{u}})$.

Since the probability vectors $\boldsymbol{\theta}_{\boldsymbol{1}}$ and $\boldsymbol{\theta}_{\boldsymbol{2}}$ will vary from one trial to another, 1,000,000 pairs of $\boldsymbol{\theta}_{\boldsymbol{1}}$ and $\boldsymbol{\theta}_{\boldsymbol{2}}$ are simulated from the Dirichlet distribution, where the concentration parameters are 1/243,…,1/243. Thus, we obtain 1,000,000 pairs of VAR$_1$ and VAR$_2$.

Note: dealing with utility values different than $[-1,1]$

We have already stated that the utility scores are assumed to be bounded by -1 and 1. As we derive "unbounded" distributions, when sampling from them, it is theoretically possible to sample a value from the tails of those distributions which falls outside of the interval $[-1,1]$. If that occurs, we merely convert the sampled value to the corresponding endpoint of the interval $[-1,1]$, before any further use of this value in an economic evaluation. Furthermore, the specifications of our model imply that for the case of the state of "perfect health", its utility is assumed to always be equal to 1. Thus, ultimately, we end up dealing with utility values which lie in $[-1,1]$.

Results

Derivation of the posterior distributions

The MCMC algorithm is run using *R* and *JAGS*. Initially, the Raftery and Lewis's diagnostic [34] is considered from a pilot MCMC run: the number of iterations required in order to estimate the 25th permille to within an accuracy of +/- 0.005 with probability 0.95 is less than 4,000. We run two MCMC chains; the first 4,000 iterations of each chain are discarded as burn-in, and then a further $C = 4{,}000$ iterations are used in total for making inference on the posterior distributions of the parameters of interest. In the Appendix, Table 1 provides a summary of posterior statistics of the β's. In order to assess potential lack of convergence of the MCMC run, we calculate the Geweke statistic [35], the Gelman and Rubin statistic [36], and the effective sample size. The aforementioned table illustrates for each of the β's the corresponding p-value of the Geweke statistic, the Gelman and Rubin statistic (also known as $\hat{R}$) and the effective sample size (the values are rounded to the nearest 100) which is calculated using the *R2jags* package [37]. Large p-values of the Geweke statistic suggest no evidence of non-convergence; no evidence was found in favour of non-convergence at $a = 5\%$. Regarding the Gelman and Rubin statistic, if the value of the statistic is large (as a rule of thumb if it is greater than 1.1), then this suggests that there is no-convergence; all the computed $\hat{R}$ values are very small and thus there is no evidence of non-convergence. The higher the decay of the autocorrelation with the increased number of simulations used, the higher the effective sample size; the values of the computed effective sample sizes are high and thus there is no suggestion of having autocorrelation issues. In overall, the aforementioned diagnostics did not find evidence of non-convergence; thus the assumption is that the 4,000 iterations are considered sufficient for the model to converge.

Moreover, the results of the derived MCMC samples are also robust to changes in priors using the alternative prior distributions which were previously stated. When different starting values for the MCMC method or different priors are used, in the end we still obtain similar results for the posterior statistics. In other words, the parameters are estimated precisely enough that the inferences of this study are not sensitive to the starting MCMC values or the particular choice of prior distributions.

The mean squared prediction error (MSE) and the mean absolute error (MAE) of the observed versus the posterior means were calculated as 0.0021 and 0.0369 in respect. The predictive performance of the model can be reviewed by considering leave-one-out-cross-validation. Specifically, each of the 42 health states which were evaluated by the MVH survey participants is removed in turn, and the model is used without the data from the that state in order to predict the mean utility for that specific health state. The leave-one-out-cross-validation MSE and MAE are 0.0046 and 0.0538 in respect. The observed means utilities and the corresponding predicted means and 99% credible intervals which were derived under leave-one-out-cross-validation are shown in Figure 2. The predictive performance of the model is quite satisfactory; all but one of the observed means fall within their 99% credible intervals derived under leave-one-out-cross-validation, and even the single observed mean which does not fall within the corresponding 99% credible interval is just slightly above the interval's upper endpoint.

Derivation of the mixtures of Normal distributions

Each posterior distribution is described as a mixture of normal distributions. The *MASS* package [38] was used in order to apply the BFGS algorithm and to estimate the parameters of the mixture components; relevant *R* code for the case of estimating the parameters of a three-component mixture of Normals can be found in the Appendix. For instance, Figure 3 illustrates the kernel density plot of state *31113* for the MCMC simulations, and the superimposed probability density functions of its approximation as a one, three, and five components mixture of normal density functions. It can be seen that the KL divergence value is small, and that a satisfactory approximation has been achieved. Figure 4 provides a visual inspection for the improvement of the approximation (in terms of the KL divergence value) of the distribution of state *31113* for the case of one, three, and five components. We observe that fewer than three components may not provide a totally satisfactory approximation to the original distribution. On average, the inclusion of three components reduces the KL values by approximately 40% and the highest KL valued that is observed is only 0.00346. However, if more than three components are used, then the algorithm allows for a minimally better approximation, but the computational complexity increases.

Given the objective to use enough components to obtain good approximations without having too many of them, the decision was to use three components. This is in agreement with the conclusions of Schmidli et al [27]. For instance, the three-component mixture of state *31113* is: 0.27864 N(0.04702,0.06881) + 0.52051 N(0.08374,0.04617) + 0.20085 N(0.16007,0.05238), where the parameters of each Normal distribution are its mean and standard deviation. The table with the three-component Normal distributions of all of the utility scores of the UK EQ-5D-3L is given in the appendix. Moreover, the appendix contains the Kernel density plots for the MCMC simulations of all relevant health states and their corresponding probability density functions of their approximation as three-component mixtures of Normals.

Note: Reviewing VAR_1 and VAR_2

~~Since the probability vectors $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ will vary from one trial to another, 1,000,000 pairs of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are simulated from the Dirichlet distribution, where the concentration parameters are 1/243,…,1/243. Thus, we obtain 1,000,000 pairs of VAR_1 and VAR_2.~~ The summary statistics of VAR_2/VAR_1 and VAR_1-VAR_2, as well as their corresponding kernel density plots are provided in the Appendix. From the summary statistics and the density plots we can see that the values of VAR_1 and VAR_2 are similar for most of the cases. Hence, whilst there are a few times where VAR_1 and VAR_2 are not very similar, these could be extreme examples that we are unlikely to encounter in practice, and for the vast majority of the times we can see that three-component mixtures produce similar results compared to those produced by the full joint distribution of utilities.

Discussion

We have developed a Bayesian hierarchical model to obtain a new tariff which fully captures the uncertainty from sample variation of the UK EQ-5D-3L scores. Specifically, the MCMC samples of the posterior utility distributions were approximated as three-component mixtures of Normal distributions by regarding the KL divergence and by applying the Broyden-Fletcher-Goldfarb-Shanno algorithm. Moreover, the model was found to be robust to changes in priors and initial values. The derived tariff can be used in economic evaluations.

The method we developed here can be applied to valuations for preference-based instruments such as EQ-5D or SF-6D conducted using TTO or SG methods. In TTO exercises we can deterministically conclude for each evaluation-survey participant the exact utility value which they assigned to each of the states they were asked to evaluate. This is also the case for SG exercises. However, for DCE's in general this information is not known at the individual level of each subject of the evaluation-survey.

Researchers have two subsequent options about how to sample from the utility distributions of the health states of the EQ-5D-3L. The first option is to sample directly from our derived MCMC samples of the posterior distributions of our Bayesian model. In fact, these MCMC samples of total size 4,000 × 243 could be incorporated in an R package so that researchers can have direct access to them for their research, if they do not want to re-run the model. The second option is to avoid the use of the MCMC samples and to use the approximate mixture distributions. Clearly, in terms of accessibility it is easier to describe the EQ-5D-3L health state utilities by presenting the derived utility distributions, and then sample from them (if needed), instead of presenting a large number of value sets. Another advantage of this option is that, theoretically, there is no upper limit to how many times a researcher can choose to sample values from the derived mixture distributions, which is useful if the objective is to use more values than the fixed number of published values. Although there are multiple advantages of using the mixtures of Normal distributions, the reader should be reminded that the mixture distributions remain an approximation. In any case, we recommend that the researcher uses the approaches described in this article, instead of a tariff which was derived under frequentist techniques, and then sample utility scores either from the derived MCMC samples or from the corresponding mixture distributions.

In fact, the approximate distributions are appealing as all the associated KL divergence values are actually less than or equal to 0.00346. Specifically, on average the KL divergence values decrease by approximately 40% when we use three components, whereas the corresponding KL divergence decrease is minimal when adding further components. Similarly to the conclusions of Schmidli et al [27], we recommend using three component mixtures as they seem to provide a good approximation while still keeping the complexity of the algorithm relatively low compared to the increasing computational complexity of having more components.

Our model respects the core principles of the framework of the UK EQ-5D-3L valuation study and extends it in Bayesian setting where further layers of uncertainty are also taken into consideration. Some assumptions of the UK EQ-5D-3L valuation study (as well as the valuation studies of other instruments and populations around the world) could be considered imperfect and debated. These include the use of TTO as an appropriate method for eliciting utilities, the concept of cardinal utility and its application for health states. Nevertheless, although not all the concepts of the UK EQ-5D-3L are ideal, the derived MVH tariff has been broadly used in economic evaluations. Thus, our model improves our knowledge about the EQ-5D-3L state utilities by accounting for the uncertainty due to the variability inherent in the data.

Guidelines for health technology assessments in the UK request that uncertainty be accounted for in economic evaluations [39]. Whilst some researchers might use bootstrap to account for sampling uncertainty or they make further probabilistic assumptions for the distribution of the benefits, when conducting sensitivity analysis, we actually provide a certain closed-form distribution for the utility value of each health state. By making this new tariff available to researchers, these distributions could be used as a known reference point which eases the situation when having to deal with different assumptions made in separate economic evaluations. Unlike the point estimates derived by frequentist methods, the distributions we have derived provide a direct representation of the perception of the associated uncertainty of the utility score of each particular health state, making it possible to directly sample a utility value from its corresponding distribution. Moreover, in addition to accounting for the uncertainty related to parameter estimates for regression coefficients, our model also accounts for functional model-misspecification, the importance of which has been highlighted by Pullenayegum et al [10]. In overall, by using our tariff in order to properly account for further levels of uncertainty, which would otherwise be tacitly ignored by other approaches, the inference of economic evaluations can be made on a certain level of certainty. Moreover, in some cases the conclusions of economic evaluations based on our approach could be potentially different compared to the conclusions had these important levels of uncertainty been improperly disregarded and not taken into account.

Pullenayegum et al [10] have argued the importance of quantifying the uncertainty of the utility values of the USA EQ-5D-3L value set in a Bayesian perspective. Some research work is focused on using the data available from other countries for the valuation of the EQ-5D-3L or similar health instruments in a more efficient way in order to produce value sets which are subject to less parameter uncertainty [12-16]; for instance they found evidence that incorporating spatial correlation among health states improves predictive accuracy. In our case we have demonstrated that most of the times our derived posterior distributions perform similarly to the full joint distribution of utilities and thus that it is acceptable to use the three-component mixtures of Normals. Furthermore, whilst we used ξ_s^{new} for the computation of every state utility u_s, Shams and Pullenayegum [15] did that only for the states which were not evaluated in the US evaluation study of EQ-5D-3L, and they used ξ_s for those states that were captured in the evaluation study, which resulted in having less uncertainty. Others have attempted to use a bootstrap approach for dealing with the underlying uncertainty [40]. Nevertheless, our principal objective was to focus on reporting the utility distributions of the distinct UK EQ-5D-3L states as explicitly specified probability distributions, whilst accounting for the underlying uncertainty. Moreover, whilst researchers are welcome to use the approximate distributions which we derived, these remain approximations, and so researchers anyway also have the option to run the full Bayesian model and use the $4{,}000 \times 243$ MCMC values of posterior utilities, instead.

Notably, the need to propagate uncertainty in the value set is not unique to the UK valuation study, nor to the EQ-5D-3L, which was considered due to its widespread use. We have used the UK valuation study as an example to illustrate how to use our methods to derive a new tariff because of their courtesy to share the data and because of their pioneering research in this field. Nevertheless, our techniques can be applied to the valuation studies of other countries and health instruments. For instance, our techniques can be applied to the EQ-5D-5L; the process needed to would be similar to that described in this paper, but this time vectors $\boldsymbol{X}_s^T$ and $\boldsymbol{\beta}$ would have more entries because the EQ-5D-5L has five response levels for each of its dimensions while EQ-5D-3L only has three response levels. Furthermore, our methods are also relevant in the area of mapping across health instruments and in the context of model-based economic evaluations.

Ara and Wailoo [41] note that uncertainty around health state utility values is usually underreported, whereas frequently only mean values are used in decision analytic models. Probability distributions can be assigned to utility scores in the context of probabilistic sensitivity analysis but the choice of the parameters of the distributions is made by considering trial-sampling variation whereas in such a case the parameter uncertainty of the regression model of the valuation study is ignored. However, our derived distributions account for multiple layers of uncertainty and they allow researchers to use them without the need to make other assumptions on the forms of such distributions. Therefore, we support the extension of the methods of this study in model-based economic evaluations in order to properly assess the impact of this approach. Future work should focus on the application of our derived distributions in model- or trial-based economic evaluations in order to properly assess the impact of our approach.

Conclusions

In conclusion, we have derived a new tariff for the UK EQ-5D-3L, which represents the utility score of each health state by a three-mixture Normal distribution. This was achieved by using a Bayesian hierarchical model which also accounts for mode-misspecification, whereas numerical optimisation was used for the approximation of the posterior distributions. Unlike its point-estimates counterparts, this tariff propagates the uncertainty due to the variability inherent in the data, as we now have closed-form distributions. Thus, we recommend the use of this tariff in economic evaluations. We believe that this approach should also be used for the valuation studies of other instruments and countries, as well as in the context of model-based economic evaluations.

Acknowledgments

(See Title Page for Acknowledgements)

References

1. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53-72.
2. EQ-5D-5L | About [Internet]. EQ-5D. Available from: https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/?_gl=1%2A1k3e75y%2A_up%2AMQ..%2A_ga%2AMTcyMjc5ODQzMy4xNjcxNzA1OTYw%2A_ga_02T9YV6MT2%2AMTY3MTcwNTk1OS4xLjEuMTY3MTcwNTk2MS4wLjAuMA..

3. Xie F, Gaebel K, Perampaladas K, Doble BM, Pullenayegum E. Comparing EQ-5D Valuation Studies: A systematic review and Methodological Reporting Checklist. Value in Health. 2013;16(3).

4. Szende A, Oppe M, Devlin NJ. EQ-5D value sets: Inventory, Comparative Review and User Guide. 1st ed. Dordrecht: Springer; 2007.

5. Sheather SJ. Density Estimation. Statistical Science. 2004;19(4):588–97.

6. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Chichester: John Wiley & Sons; 2011.

7. Baio G. Bayesian methods in health economics. Boca Raton: Chapman & Hall/CRC; 2013.

8. Grieve R, Nixon R, Thompson SG. Bayesian hierarchical models for cost-effectiveness analyses that use data from cluster randomized trials. Medical Decision Making. 2010;30(2):163–75.

9. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. Health Economics. 2001;10(4):303–15.

10. Pullenayegum EM, Chan KKW, Xie F. Quantifying Parameter Uncertainty in EQ-5D-3L Value Sets and Its Impact on Studies That Use the EQ-5D-3L to Measure Health Utility: A Bayesian Approach. Medical Decision Making. 2016;36(2):223–33.

11. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics. 2002;21(2):271–92.

12. Kharroubi SA, Brazier JE, O’Hagan A. Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. Social Science & Medicine. 2007;64(6):1242–52.

13. Kharroubi SA, Brazier JE, Roberts J, O’Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. Journal of Health Economics. 2007;26(3):597–612.

14. Kharroubi S, O'Hagan A, Brazier J. Estimating utilities from individual health preference data: a nonparametric Bayesian method. J Royal Statistical Soc C. 2005;54(5):879-895.

15. Shams S, Pullenayegum E. Reducing uncertainty in EQ-5D value sets: The role of Spatial Correlation. Medical Decision Making. 2019;39(2):91–9.

16. Che M, Xie F, Thomas S, Pullenayegum E. Bayesian models with spatial correlation improve the precision of EQ-5D-5L value sets. Medical Decision Making. 2023;43(5):587–94.

17. Dolan P, Gudex C, Kind P, Williams A. The Measurement and Valuation of health. First Report of The Main Survey. University of York: Centre for Health Economics; 1994.

18. Dolan P, Gudex C, Kind P, Williams A. The Measurement and Valuation of health. Final Report on the Modelling of Valuation Tariffs. University of York: Centre for Health Economics; 1995.

19. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring Preferences for Health States Worse than Death. Medical Decision Making. 1994;14(1):9–18.

20. Williams AH, Gudex C, Kind P, Dolan P. Health State Valuations from the British General Public, 1993 [Internet]. [data collection]. UK Data Service. SN: 3444. 1995. Available from: http://doi.org/10.5255/UKDA-SN-3444-1

21. Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. 1st ed. London: Chapman & Hall; 1996.

22. Gamerman D. Markov chain Monte Carlo. 1st ed. London: Chapman & Hall; 1997.

23. Robert CP, Casella G. Introducing Monte Carlo methods with R. 1st ed. New York: Springer; 2010.

24. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. 3rd ed. Boca Raton: Chapman & Hall/CRC; 2013.

25. Plummer M. Just another Gibbs sampler (JAGS) [Internet]. 2017. Available from: http://mcmc-jags.sourceforge.net/

26. Rubinshtein YG. Possibility of approximating multimodal distributions by mixtures of standard probability density functions. Measurement Techniques. 1993;36(8):858–64.

27. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics. 2014;70(4):1023–32.

28. O'Hagan A, Forster J. Vol 2B Kendall's advanced theory of statistics Bayesian inference. 1st ed. New York: Arnold; 2004.

29. Broyden CG. The Convergence of a Class of Double-rank Minimization Algorithms. IMA Journal of Applied Mathematics. 1970;6(3):222–31.

30. Fletcher R. A new approach to variable metric algorithms. The Computer Journal. 1970;13(3):317–22.

31. Goldfarb D. A Family of Variable-Metric Methods Derived by Variational Means. Mathematics of Computation. 1970;24(109):23.

32. Shanno DF. Conditioning of Quasi-Newton Methods for Function Minimization. Mathematics of Computation. 1970;24(111):647.

33. Dai Y-H. A perfect example for the BFGS method. Mathematical Programming. 2012;138(1-2):501–30.

34. Raftery AE, Lewis SM. [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. Statistical Science. 1992;7(4):493–7.

35. Geweke J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In Bayesian Statistics. 1992;4:168–93.

36. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Statistical Science. 1992;7(4):457–72.

37. Su Y-S, Yajima M. R2jags: Using R to Run 'JAGS'. R package version 0.7-1 [Internet]. 2021. Available from: http://cran.r-project.org/web/packages/R2jags/index.html

38. Ripley B, Venables B, Bates D, Hornik K, Gebhardt A, Firth D. MASS: Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-60 [Internet]. 2023. Available from: https://cran.r-project.org/web/packages/MASS/index.html

39. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. London; 2013. (Process and methods guides). Available from: https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781

40. Gray A, Rivero Arias O, Leal J, Dakin H, Ramos Goni JM, Ramos Goni JM, et al. How important is parameter uncertainty around the UK EQ-5D-3L value set when estimating treatment effects? 2012. Unpublished report. Available from: http://www.ces-asso.org/sites/default/files/Gray%20et%20al%20HESG%20submitted%201.pdf

41. Ara R, Wailoo A. Using Health State Utility Values in Models Exploring the Cost-Effectiveness of Health Technologies. Value in Health. 2012;15(6):971–4.

42. Anonymous; 2020. Details omitted for double-anonymized reviewing.