

Report on EQ5D-5L UK Valuation Study

**Keith R Abrams, PhD CStat
Professor of Medical Statistics & NIHR Senior Investigator
Emeritus Biostatistics Research Group Department of
Health Sciences University of Leicester, UK**

24th September 2019

Note – this report is briefer than it would have otherwise been as I was on extended compassionate leave during July and August 2019.

Introduction

This report considers the work undertaken by the valuation set authors and quality assurance by EEPRU. In doing so it focuses mainly on the model specification used and the manner in which it was implemented.

Data

No extensive exploration of the data used was undertaken for this report as it was assumed that other experts would comment on that. However, it is noted that others (both EuroQoL and EEPRU) have identified a number of inconsistencies with some of the data that has been collected. Whilst it is not inconceivable that these inconsistencies may have an impact on the model results (possibly exacerbated by the way in which the model is formulated – see below) this would have to be verified in further work (especially if the model were to be modified or re-formulated).

Model Specification

There are a number of issues to note regarding the specification and formulation of the model as described in Feng *et al* (2018) and the WinBUGs code supplied by the valuation set authors.

An important aspect of the data collection and model is that each individual contributes a number of data points to the overall dataset. This is true for both the TTO and DCE data, and also the hybrid dataset. Consequently it is important to

allow for the possibility that this will induce correlation and also that there will be individual-to-individual variation, and the EuroQoL authors do this by adding a random effect to their model. The way in which they do this is via a multiplicative term (γ) – see equation 3 in Feng et al. This is unusual given that the models are linear either on the natural scale (for TTO) or log odds scale (for DCE). For example a recently published paper by Pickard *et al* (2019) on the US EQ-5D-5L Valuation Study uses a more usual additive random effect – see equation 1 of Pickard *et al* (2019).

There are a number of consequences of the using a multiplicative random effect (and the way in which it has been included). Essentially it means that the model is non-linear, and given that Markov Chain Monte Carlo (MCMC) methods are used to estimate the model parameters, it is of concern that this could induce instability into the model. Aside from re-formulating the model to include an additive random effect, the model could also be re-formulated so that a multiplicative random effect is parameterised more naturally, for example using $\exp(\gamma)$. However, another consequence of a multiplicative formulation is that the two variance components (the random effect and the error term) are on different scales – this also might have an impact on model performance. The impact of the random effect is also assumed constant across dimensions. In addition for the hybrid model any heterogeneity that is captured is also assumed to operate equally across TTO and DCE. This might or might not be a realistic assumption.

As regard the hybrid model (including both TTO and DCE data) – this includes common regression parameters in the two sub-models (with appropriate rescaling due to one being on the natural scale and the other on the log odds scale). However, it would appear that any correlation, due to individuals contributing data to both sub-models, is captured by both models having a common multiplicative random effect. Both aspects could have been dealt with by adopting a hierarchical model formulation with additive random effects on the regression terms – this is a more natural way for modelling what are essentially multivariate data, and in fact was adopted by Pickard *et al* (2019) – see equation 4. Aside from any theoretical reasons why TTO or DCE generated data may be preferable to derive a valuation set, a thorough exploration of how different types of “hybrid model” can capture both

correlation and differences in regression parameter estimates based on TTO and/or DCE generated data is necessary. However, in any future re-formulated model the combined effect of having additive random effects, a hierarchical structure and using MCMC methods to estimate the model parameters may well improve model stability and computational performance.

A more minor comment is that the way heteroscedasticity is modelled uses the mean (linear predictor) – this may or may not be appropriate, but an alternative would be to include individual dimensions/components, or functions of them, as necessary. This was the approach adopted by Pickard *et al* (2019) – see equation 3.

In additional work, the valuation set authors have investigated potential interviewer effects – certainly for a hybrid model with gamma distributed multiplicative random effects. However, the WinBUGs files which produce the results in Feng *et al* (2018) do not appear to include this extension.

Model Implementation

As mentioned previously the model parameters were estimated using MCMC methods (reported in Feng *et al* (2018)), though they have also been estimated using Maximum Likelihood (ML) methods. Whilst I share the valuation set authors' frustration that these results were not included in the final paper, nor were a set of comprehensive sensitivity analyses – both in terms of sensitivity to prior distributions (for the Bayesian approach) and an assessment of convergence, it is absolutely vital that these are in the public domain (as well as the code and the data). It is not sufficient for just the code and data to be and expect others to undertake them – though clearly they may well wish to explore specific aspects. However, reviewing the Quality Assurance undertaken by EEPRU and further sensitivity analyses, it is clear that the model performance is less than adequate in terms of convergence. It should be noted that it is impossible to show that a model has converged using MCMC methods, only that one has failed to show non-convergence. A consequence of this is that using MCMC diagnostic tools/tests (for example in CODA/BOA) may lead to a false sense of confidence, i.e. a test may fail to find evidence of non-convergence, but this cannot be interpreted as that a model has definitely converged. Hence, why extensive sensitivity analyses are required, for example

assessing the effect of length of burn-in, sample, starting values, auto-correlation etc on the model results. It is noted that the valuation set authors have stated that the ML results are broadly similar to the Bayesian results published, and whilst this may appear reassuring they nevertheless struggle to estimate the model parameters, and this may be because of the way in which the model is specified (see above).

A key aspect of any Bayesian analysis is the specification of prior distributions, the justification of these and an assessment of the sensitivity of the results to them.

Whilst these are clearly specified by Feng *et al* (2018) there appears to be no justification for the precise hyper-parameters used or an interpretation of what the prior distributions represent for each of the model parameters. As stated earlier this needs to be part of extensive and comprehensive sensitivity analyses.

Both of the above two issues – convergence (and assessment thereof) and sensitivity to prior distributions of course can depend upon model specification, and if the model were to be re-formulated (as described above) then these would still be required.

Whilst a variety of models are explored in Feng *et al* (2018), they are compared using the Deviance Information Criterion (DIC). The DIC is a measure of relative fit of the models, and so it can be used to compare alternative models as Feng *et al* have done. However, what would also be of important to see is a measure of absolute fit, for example the posterior residual deviance. One model may provide a better fit than another one, but still by a poor model (in terms of predictive ability etc).

Discussion & Next Steps

In summary, there appear to be issues with both the model specification and its implementation/performance (or at the very least extensive and comprehensive sensitivity analyses being in the public domain in addition to the code and data). These issues are aside from any concerns regarding the data. As such, use of the results contained in Devlin *et al* (2018) and Feng *et al* (2018) for UK health policy making at this moment in time would appear to be unwise, and could be subject to challenge if in fact they were.

The next steps (irrespective of any concerns regarding the data and even if new data were to be collected) should involve further work exploring a re-formulation of the model (with the model adopted by Pickard *et al* (2019) used as a starting point, i.e. specifically the use of an additive random effect and their approach to modelling heteroscedasticity, regardless of whether a TTO only, DCE only or a hybrid model is adopted) together with a clearly presented set of extensive and comprehensive sensitivity analyses if MCMC methods are used. It should be noted that whilst Pickard *et al* (2019) is only one of a number of country-specific valuation sets to be published (for example others include; Xie *et al*, 2016; Purba *et al*, 2017; Shafie *et al*, 2019), and was not identified in any systematic manner, future work should explore both data generation and modelling approaches adopted in *all* published valuation sets so that genuine country-to-country differences can be assessed and understood. For example, Xie *et al* (2016) and Purba *et al* (2017) adopt the hybrid and modelling approach of Oppe & van Hout (2010), evaluated by Rowen *et al* (2014) and used by the Feng *et al* (2018), whilst Shafie *et al* (2019) adopt a model formulation similar to Pickard *et al* (2019). It is possible (and in fact quite likely) that any re-formulation of the model used by Feng *et al* (2018) as suggested above will alleviate the implementation/performance issues that are apparent with its use.

Declarations of interest

In line with NICE's Policy on Conflicts of Interest, I declare the following indirect interests:

- Strategic & Methods HTA Advice to IPSEN (May 2018 – May 2018)
- Strategic & Methods HTA Advice to Sanofi (May 2018 – Dec 2018)
- Strategic & Methods HTA Advice to Pfizer (Jan 2018 – present)
- Strategic & Methods HTA Advice to Roche (Jan 2018 – present)
- Strategic & Methods HTA Advice to NovoNordisk (May 2018 – May 2019)
- Strategic & Methods HTA Advice to SIRTEX (Sept 2018 – present)
- Strategic & Methods HTA Advice to Abbvie (Oct 2018 – April 2019)
- Strategic & Methods HTA Advice to Takeda (Nov 2018 – Dec 2018)
- Strategic & Methods HTA Advice to Leo (Feb 2019 – March 2019)
- Strategic & Methods HTA Advice to Janssen (April 2019 – May 2019)
- Strategic & Methods HTA Advice to Amaris (April 2019 – present)
- Partner & Director, Visible Analytics Limited – HTA consultancy company (April 2019 – present)

References

Devlin NJ, Shah KK, Feng Y, Mulhern B van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics*. 2018;27:7–22. DOI: 10.1002/hec.3564

Feng Y, Devlin NJ, Shah KK, Mulhern B, van Hout B. New methods for modelling EQ-5D-5L value sets: An application to English data. *Health Economics*. 2018;27:23–38. DOI: 10.1002/hec.3560

Oppe, M & van Hout B. (2010). The optimal hybrid: experimental design and modeling of a combination of TTO and DCE. In Yfantopoulos J (Ed.), 27th Scientific Plenary Meeting of the EuroQol Group—Proceedings. Rotterdam: EuroQol Group Executive Office.

Pickard AS, Law EH, Jian R, Pullenayegum E, Shaw JW, Xie F, Oppe M, Boye KS, Chapman RH, Gong CL, Balch A, Busschbach JJV. United States Valuation of EQ-5D-5L Health States Using an International Protocol. *Value in Health*. 2019; 22(8):931–941. <https://doi.org/10.1016/j.jval.2019.02.009>

Purba FD, Hunfeld JAM, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Ramos-Goñi JM, Passchier J, Busschbach JJV. The Indonesian EQ-5D-5L Value Set. *Pharmacoeconomics*. 2017 Nov;35(11):1153-1165. doi: 10.1007/s40273-017-0538-9.

Rowen D, Brazier J, & van Hout B. A comparison of methods for converting DCE values onto the full health-dead QALY scale. *Medical Decision Making*, 2014; 35(3): 328–340.

Shafie AA, Vasan Thakumar A, Lim CJ, Luo N, Rand-Hendriksen K, Md Yusof FA. EQ-5D-5L Valuation for the Malaysian Population. *Pharmacoeconomics*. 2019 May;37(5):715-725. doi: 10.1007/s40273-018-0758-7.

Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, Poissant L, Johnson JA; Canadian EQ-5D-5L Valuation Study Group. A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. *Med Care*. 2016 Jan;54(1):98-105. doi: 10.1097/MLR.0000000000000447.

Acknowledgement

This work was commissioned by NICE and funded via an unrestricted grant from the EuroQol Research Foundation to NICE. The data for the published valuation set are owned by the Office of Health Economics.