Expert evaluation of EQ-5D-5L

Prepared by Denzil G Fiebig

August 20, 2019[1]

Introduction

As a general comment, it seems likely that the resolution of the exchange between the authors of the EEPRU review and those of the value set, will require finding some sensible middle ground. Hernández-Alava et al. (2018), hereafter the EEPRU review, were very thorough in their evaluation resulting in a long list of criticisms levelled at the value set production documented in Devlin, Shah et al. (2018a) and Feng et al. (2018). Some of these criticisms are generic and not specific to the EQ-5D-5L analysis for England and some are potentially more problematic than others. The value set authors, Devlin, van Hout et al. (2018b), have vigorously defended the criticisms levelled at the existing value set and I see merit in some but not all their rebuttal arguments. Ultimately, finding this sensible middle ground is made difficult by the lack of a gold standard to provide a baseline comparator.

What follows has been organized in order to directly answer the questions outlined in Annex 1 of the Agreement for Expert Advice with NICE although many of my responses are interconnected by my overall position on the production of the value set. I have had general discussions with colleagues about some of the issues in this report but none of these discussions have directly impacted the answers that I have provided.

[1] This work was funded by an unrestricted grant from the EuroQol Research Foundation. The data for the published valuation set are owned by OHE.

Data quality

1. Do the valuation set data reflect the preferences of the public in England adequately?

The EEPRU review has been especially critical of the data quality used to generate the EQ-5D-5L value set although some of the issues that were raised were generic to stated preference methods and not specific to the EQ-5D-5L and some could be viewed as relatively minor or at least not especially impactful on the final value set outcomes.

Experimental design

In terms of experimental design, the EEPRU review was critical of both the sample size and the number and type of health states that were directly evaluated. Obviously, these are not independent. In the general stated preference literature formal decision rules are rarely used to motivate choice of sample size and so the likely determinant is the available research budget and/or literature norms; see de Bekker-Grob et al. (2015). As such, comparing sample sizes across different studies is not a terribly enlightening exercise.

The criticism of the coverage of health states used, had two elements; (i) that the chosen health states didn't necessarily reflect how commonly they appear in cost-effectiveness studies, and, (ii) that a very small number were chosen relative to the total number of possible health states that needed to be evaluated. The first of these was successfully rebutted in Devlin, van Hout et al. (2018b), where they argue that formal statistical properties of the design are more important than accounting for prevalence, an argument supported in studies including Yang et al. (2018). It is the second element of coverage that Devlin, van Hout et al. (2018b) do not directly address but which I feel is much more important.

The EEPRU review makes a substantive point about coverage by casting this discussion in the general context of model misspecification and the need to be able to conduct a more "robust" analysis. While not formally defined, robust here means the capacity to explore for possible model misspecifications and as such this data issue is closely related to modelling questions that follow.

The preferred hybrid model used to generate the value set, has two features that I will highlight in my responses. The first is that it is relatively complex especially in its treatment of heterogeneity and the second is that at its core is a main effects specification in terms of domain levels. Discussion of this first feature will be deferred to Questions 2 and 4. The second feature is relevant here in the discussion of experimental design although it too will be discussed in response to Question 4. The EEPRU review is not explicit in the types of misspecification that they view as threats to "robust" analysis but to me the key threat involves ignoring potential interaction effects between the domains, especially at worse levels of health.

Consider the situation where "extra" resources are available to increase the sample size. The position of the EEPRU review is clear; the number of distinct design points shown to respondents should be increased. The alternative is to allocate more observations to each of the initially chosen design points. If the design allows estimation of the required main effects in a relatively efficient manner, then this second strategy is entirely defensible. Limited coverage of the design space is a natural outcome of exercises such as the one being considered here, and given all required effects are identified, changing the sample size is reflected in estimation precision but does not compromise the resulting estimates. This is essentially the position of the value set authors.

In the current context the strategy of increasing the number of distinct design points provides benefits only to the extent that you worry about interaction effects but then you're addressing a different design problem. By highlighting the limited coverage of design points, the EEPRU review has flagged a potential problem. Whether it is really a problem depends crucially on the assumption that the main effects specification is adequate. This issue will be reconsidered in the context of Question 4 where I will argue why limited coverage of health states is indeed a problem.

Problematic data

All data misbehave and so there will always be data quality issues associated with any empirical work. In the case of stated preference data, high on the list of challenges is that respondents may not totally engage with the survey, possibly because of their hypothetical nature. Some seemingly problematic responses can come from respondents who are either not engaged with the survey and/or those who are relatively insensitive to differences in attribute levels. Conceptually, these are different and making the distinction important but empirically difficult. The EEPRU review concedes this issue but makes no attempt to directly disentangle the two sources of "problems" but instead provides a detailed account of the types and extent of anomalies.

While the EEPRU review was very comprehensive there was no rating of problems in terms of severity and I thought some of the issues included were relatively minor. For example, yes, there may be "bunching" in TTO responses, but this is not an unusual phenomenon in other data collections and does not prove fatal for the subsequent analyses. And yes, a discrete choice experiment (DCE) in isolation can't provide value sets but here it is not being used in isolation. Concerns were raised about the assumption of independence within the DCE tasks. Why should this be problematic? A priori this seems to be a reasonable assumption and even if violated wouldn't present problems for estimating the levels of the value set.

However, this still leaves some substantive issues such as the extent of respondent engagement with the data collection, interviewer effects, methods of dealing with problematic data and non-response bias that are not satisfactorily resolved. A common response by the value set authors was along the lines that what they are doing is consistent with the approach of others. This is not overly satisfactory if everyone is making the same mistake. There should at least be some recognition where there is a need for better methods of data collection. For example, take the response to the sample size issue. I agree this is an issue of secondary importance, but it remains the case that a formal justification of sample size choice is a reasonable element of best practice in stated preference methods; see de Bekker-Grob (2015). I can't see why this was not recognized by the value set authors.

As a second example, the EEPRU review flagged a potential problem with the DCE data when respondents were indifferent to states A and B. One would hope for a random choice in such cases rather than say a systematic default of choosing A. In their response, Devlin, van Hout et al. (2018b) contend they "…routinely monitored the number of respondents always picking one alternative (A or B) … and found this type of response behaviour very rarely occurred". But this is not the point, and the presence of such behaviour is difficult to disentangle from the analyses in the EEPRU review. A simplified random effects probit specification of equation (3) from the EEPRU review, requested as part of clarifying questions, does in fact indicate a modest and significant bias towards health state A after controlling for level differences in the health states. Given randomization of health states to A and B it is unlikely this is a major concern, but it is another indicator that problems exist, and they need to be recognized.

The foregoing discussion implies that there is no definitive answer to the original question of whether the valuation set adequately reflects the English preferences. On balance there do seem to be several concerns relating to data quality that could be improved quite substantially in subsequent data collections. This is reinforced by the audit of protocol compliance and interviewer effects in EuroQol (2019) and EuroQol's decision to update the original valuation protocol, EQ-VT 1.0, that was used in the data collection under review. These together would seem to vindicate the concerns raised by the EEPRU review and dictate a new data collection.

Modelling

2. **Considering the model that informs the published 5L valuation set:**
    a. **Is there evidence of convergence failure? If so, please comment on the strength of this evidence and the implications for the validity of the model.**
    b. **Is it possible to achieve convergence (e.g. by changing the model parameters or specifications, or by estimating a model based on only TTO or discrete-choice experiment data instead of a hybrid model)?**

The preferred hybrid model in Feng et al. (2018) requires several normalizations to ensure the model is identified and even with these imposed, convergence problems are possible as indicated in the EEPRU review. The response in Devlin, Shah et al. (2018b) stresses that the results are plausible and parameter estimates are relatively insensitive across different models, data and estimation methods. Implicitly they are arguing that even if there was convergence failure it apparently doesn't matter in terms of the results. But the space of plausible results is quite large and even small differences in parameter estimates may translate into large value set differences which is what we care about. Moreover, the supporting evidence of stability, presented in the Figure on p. 13 of the appendix to Devlin, van Hout et al. (2018b), is confusing. The DCE data in isolation cannot identify the required parameters, so what is being compared here? Also, the results do not seem to be consistent with their counterparts in Tables 3 and 4 of Feng et al. (2018). These issues were posed in questions to the value set authors but were not resolved by their answers.

The need to resolve these issues is predicated on the assumption that the preferred hybrid model in Feng et al. (2018) is the most appropriate one for generating the value set. I have serious doubts that this is in fact the case and as such consider the issues raised here are moot. As implied by 2(b), much simpler specifications of the types considered in Feng et al. (2018) are not subject to these potential problems. I would still argue that collection of both TTO and DCE data for use in a hybrid model is to be recommended, but as I argue in Question 4 there are likely to be preferable specifications of this hybrid model for the primary task of producing a value set that are simpler and avoid these potential convergence problems.

As a postscript here, Question 2 suggests a "solution" that involves the use of only one part of the data and by implication that this might be a preferred approach in future modelling. I do not agree and consider a hybrid TTO/DCE approach to be sensible. One aspect of the data collection not in dispute, is that answering these types of questions can be cognitively challenging for some people. As such, having two sources of information from two somewhat different types of questions provides a convenient internal sensitivity check. The alternative of selecting one method would need to be based on evidence of superior performance. Such evidence does not exist, and it is difficult to see how one could gather such evidence without a "gold standard” to act as a comparator. However, the EEPRU review makes the reasonable claim that the TTO data is the “dominant source of information”. The argument being the DCE does not include duration in the health state. But the inclusion of duration as an attribute in a DCE is feasible. Respondents are asked to compare health states with potentially different durations and as with other attributes duration is varied as part of the design; see for example Viney et al. (2014). Following such an approach would eliminate this criticism and strengthen the case for this hybrid approach in future data collections. See Stolk et al. (2019) for further discussion on employing this so-called DCE duration approach.

3. **The valuation set authors state that “modelling does not assume that all TTO responses are ‘accurate’. The modelling approaches were selected to reflect the characteristics of the data, following careful assessment of individual respondent level data”. They state that the modelling methods also account for interviewer effects (see page 4 of Devlin et al. response to the EEPRU report [Devlin et al. 2018b]). Does the modelling approach chosen by Devlin et al. (2018a) and Feng et al. (2018) adequately account for the characteristics of the data?**

There is recognition on the part of the valuation set authors that there were data issues but their attempts to address these issues do not present a strong case supporting their contention that “modelling approaches were selected to reflect the characteristics of the data”. They do exclude some respondents based on implausible responses, and they do constrain estimated level differences to avoid “logically inconsistent” results. Both would be standard in such exercises and not specific to the data issues here. Other than these two, I found no evidence of any specific modelling strategies that directly address concerns about interviewer effects and differential engagement across respondents.

Instead, in their response in Devlin, van Hout et al. (2018b), they seem to be arguing that there is no reason to suspect the problematic data introduced any biases in their modelling approach which is far removed from being proactive in accounting for problematic data issues. Nevertheless, as with the answer to Question 1, I have some sympathy with the response of the value set authors. There may be a problem because the core results, represented by the value set, are sensitive to features of the data that have not been explicitly modelled. But this is a suspicion and it is not obvious that there are in fact any biases in the current results given the approach taken. No evidence has been presented in the EEPRU report that the problems they have identified, such as the extent of respondent engagement with the data collection, interviewer effects and non-response bias; do in fact lead to substantial biases.

4. **Are there particular choices in the model that cause you concern? Please provide your rationale, specific recommendations for alternative approaches and, where possible, supporting evidence (for example, outcome of sensitivity analyses performed by the valuation set authors or EEPRU). Please also explain the magnitude of your concern – are any issues grave enough to mean that the model should not be used to inform resource allocation decisions in England? In particular, please consider the 4 concerns raised in the EEPRU quality assurance report, listed in the table [see table 1 of the document showing the questions set by NICE].**

There are two aspects of model choice that especially concern me and were previously mentioned in my Question 1 responses. The first is that the final hybrid model is relatively complex especially in its treatment of heterogeneity and hence is directly related to concern (B) that forms part of this question. The need to model heterogeneity was not well motivated by Feng et al. (2018) but it seems from Devlin, van Hout et al. (2018b) that "…the research proposal for the EQ-5D-5L value set for England study (as approved by the NIHR Policy Research Programme) specifically stated that the research would address the observed heterogeneity of the population …". Heterogeneity in responses may be of interest in its own right and if so, may very well require an even more detailed representation of heterogeneity that the one being proposed. But this is a separate issue and arguably one not overly relevant to the primary task of producing a value set. If what is needed is a single population value set for all possible health states, what is the argument for specifying a model that allows for individual heterogeneity, that once estimated, needs to be averaged out in order to get the required estimates?

The basic modelling problem bears a close resemblance to that faced in the forecasting literature. Within-sample data are used to fit a model that then produces predictions for out-of-sample outcomes. This forecasting literature emphasizes that while complex models are better able to produce good within-sample fits, their superiority over simpler models is not maintained when one considers performance out of sample. In part, this is due to an overfitting problem but even a theoretically correct model may provide inferior out-of-sample predictions to a simpler approximate model because of the relatively large impact of estimation error. This type of result is highlighted in Clark and McCracken (2012), who formalize the general issue of trade-offs in forecast accuracy associated with noise in parameter estimation. Also, see the literature comparing homogeneous and heterogeneous panels, i.e. whether coefficients on regressors are specified to vary over individuals or not. Here a common theme is that the simpler, homogenous specifications are preferred for forecasting; see for example Baltagi (2008). While there is a temporal element to much of this literature it is entirely relevant here as the construction of the value set is an exercise in out-of-sample prediction. The design chooses a subset of health states for which in-sample observations become available for estimation and then predictions are generated for the entire set of health states.

Without a compelling argument for specifying a model with heterogenous effects, the Feng et al. (2018) preference for more complicated models is not warranted. The fact that they produce better within-sample fit statistics is not sufficiently strong evidence to support their use. Naturally this discussion is predicated on requiring a "single" value set. Requiring multiple value sets that are applied according to observable patient characteristics and/or specific health interventions would substantially change the primary modelling objective. Moving in such a direction would have serious policy implications and represent a major departure from current practice and goes far beyond the scope of the current exercise.

My second fundamental concern with the model choice relates to allowing for interactions between the EQ-5D levels in the model specification. There seems to be enough evidence elsewhere to suggest that the inclusion of interactions is warranted, and it matters for the value set that is produced; see for example the review of Mulhern et al. (2019) and references therein.

I know Feng et al. (2018) say they addressed the possible inclusion of interactions, but I thought this is an issue that deserved more justification. Concentration on fit as a discriminator is misguided here. Recall Question 1, where concern was expressed about whether the design had the power to in fact accurately assess the need for these interaction effects. This is where the EEPRU review concerns about coverage are relevant. If the possibility of interaction effects is a key element of the specification, then limited design coverage is a problem. Consider the simplest possible case of two domains each with two levels giving four possible health states. If one specifies a model with just main effects, then you only need a design that includes observations in three of the four possible states to estimate the main effects parameters. This model could then be used to predict the value for the health state omitted from the design. If, however, one specifies a different model with an interaction between the domains then observations in all four states are required to ensure the identification of all parameters. But different designs while achieving identification may differ in their ability to precisely estimate the specified interaction effect.

There is nothing in the documentation provided by the value set authors that provides any confidence that there is sufficient coverage of health states to enable one to accurately determine the presence of interactions. I am not even sure that such effects are identified given the design and so evidence that they could not be well estimated with the available data is not a convincing argument that interactions do not play a role in value set determination.

Viney et al. (2011) and Viney et al. (2014) both conclude that a main-effects-only specification is inferior to one including interactions of dimensions at their worst levels. These models indicate that the decrement in utility associated with moving to the worst level in one dimension depends very much on whether any of the other dimensions are already at this worst level. The decrement due to the first instance of moving to the worst level is large, but subsequent dimensions moving to a worst level had a relatively smaller disutility. This is a result that seems a priori reasonable. Given this position, it is necessary to use a design that allows for such interactions, at least at the worst levels of health.

Overall, these arguments provide further support for Recommendation 2 in Hernández-Alava et al. (2018) that calls for a “statistical modelling process that is robust and fit for purpose”.

Turning to the specific issues that were flagged in the Question 4 table:

A: Approach to handling valuations of +1

I agree with the EEPRU review authors that the treatment of this issue in Feng et al. (2018) is incorrect. This is not a censoring problem. The value set, by design, has a limiting value of unity.

B: Approach to heteroskedasticity and heterogeneity

I agree with the EEPRU review authors that the treatment of heteroskedasticity seems inconsistent across models for the TTO and DCE responses and the motivation for such adjustments confuses variability in respondent outcomes and the need to match population characteristics.

As has been stressed above, in my initial response to Question 4, the need to model heterogeneity was not well motivated. If all that is needed is a single value set for all possible health states what is the argument for specifying a model that allows for heterogeneity that ultimately is not exploited? The “Restricted hybrid” model reported in Table 3 of Feng et al. (2018) seemed a priori to be a good choice for a baseline model, although it does not include interactions. It is not possible to compare the results for this model with others provided in Tables 3-5 of Feng et al. (2018) in terms of parameter estimates and instead the value sets estimates need to be compared. Using the reported estimates, Table A compares the value set results for the “Restricted hybrid” model (which does not account for heterogeneity) and the authors’ preferred “Multinomial slope” model used to generate the value set.

Table A: Comparison of index values across two models for selected health states*

Health state	**Multinomial slope**	**Restricted hybrid**
11211	0.951	0.949
22222	0.703	0.706
33333	0.595	0.597
44444	-0.094	-0.100
55555	-0.284	-0.235

* These calculations are based on the parameters estimates provided in Feng *et al*. (2018) and hence are subject to rounding errors.

The comparison in Table A indicates these two models produce very similar value set results. While the valuation-set authors' preferred "Multinomial slope" model provides a superior within sample fit, this does not imply a superiority in terms of value set estimates.

As a general point, interpreting comparisons across models is difficult because there is no gold standard, but to the extent that they are useful in highlighting where there are differences, if any, they should be done using the value set estimates. Ultimately these are what is required and are in a metric that observers are more likely to readily understand; see Viney et al. (2011) and Viney et al. (2014) for examples.

C: Possible conflict between distributional assumptions for TTO and DCE parts of the model

I agree that it seems unnecessary to use different distributional assumptions in the TTO and DCE parts. One could easily have maintained normality for both parts, as is done in follow up analyses undertaken in response to clarification questions. Having made the choice though, it was not clear to me how exactly the "inconsistency in the distributional assumptions" could lead to biased parameter estimates or more specifically to the estimated values for the health states which is what we care about. This is another example of where the EEPRU review has flagged an issue worth investigating but where there is no evidence that it is actually a problem.

D: Prior distributions in the model: whether they are well-justified, how informative they are, and how sensitive the model results are to them

All empirical work involves the imposition of "priors"; which data to use, what variables to include. The strength of a Bayesian approach is the capacity to include prior information in a formal way. Such priors can be very subjective and subject to dispute. As such most applied Bayesian analysis requires uninformative priors. Here the ultimate "test" is whether the results are sensitive to the choice of priors.

Recalling my response to Question 2, I am not convinced by the response of Devlin, van Hout et al. (2018b) that the results are insensitive. Moreover, it is preferable to have this comparison in terms of the value set and not parameter estimates. Given my responses to other questions, I do not consider resolution of this issue to be a priority.

Conclusions and recommendations

5. **In your opinion, should resource allocation decisions in England (including NICE evaluations) use utility values derived using the 5L valuation set for England?**

No. There is a strong case for using a 5L value set in preference to a 3L version. The issue is then whether the current 5L value set that has been proposed should be used. This I would not support.

6. **If the answer to question 5 is NO:**
    a. **What action do you recommend to create a 5L valuation set that would be suitable for informing resource allocation decisions in England (including NICE evaluations)? Please be explicit about whether you believe new data collection is required or if you recommend different modelling approaches of the current data set.**
    b. **In the interim, whilst the actions specified above are being done, should resource allocation decisions in England be based on the existing 5L valuation set for England?**

There are good reasons for value sets to be routinely updated; preferences for different health states change, as does the composition of the population and methods and protocols to elicit these preferences are continually improving. It seems that now is an opportune time for an update to take place and possibly to formalize how often such updates should occur. Given the importance of the issues at stake, for which there is no disagreement, a regular cycle of review and updating is desirable.

I would not recommend revisiting the existing data with refined methods. Valid criticisms exist for both the data and the methods used to produce the existing 5L value set for England. The value set authors maintain that the problems with the data are exaggerated and what's more any new data collection is likely to generate a value set similar to this existing set. Even if this is the case, a new data collection has the advantage of removing any doubt that the resulting value set is as good as can be provided with existing best practice methods. It is necessary to restore trust in the English value set. The disadvantage is the cost involved in funding the new data collection and the delay in replacing the current system based on the old 3L value set. Obviously, this is a trade-off for others to evaluate. In the interim there seems little choice but to continue with the old 3L value set.

Declarations of interest

In line with NICE's Policy on Conflicts of Interest, I declare the following. I serve on the Board of the Centre for Health Economics Research and Evaluation where Brendan Mulhern is employed. I serve on the Board of the Centre for Health Economics where Stephen Pudney is Adjunct Professor.

References

Baltagi, B.H. (2008), “Forecasting with panel data”, Journal of Forecasting 27, 153-173.

Clark, T.E. and McCracken, M.W. (2012), “In-sample test of predictive ability: A new approach”, Journal of Econometrics, 170, 1-14.

de Bekker-Grob, E.W., Donkers, B., Jonker, M.F. and Stolk, E.A. (2015), “Sample size requirements for discrete-choice experiments in healthcare: a practical guide”, Patient 8, 373-384.

Devlin, N.J., Shah, K.K., Feng, Y., Mulhern, B. and van Hout, B. (2018a), “Valuing health-related quality of life: An EQ-5D-5L value sets for England”, Health Economics 27, 7-22.

Devlin, N.J., van Hout, B., Shah, K.K., Mulhern, and Feng, Y. (2018b), “Response to: Quality review of a proposed EQ-5D-5L value sets for England”, unpublished report provided by NICE.

EuroQol (2019), “QC report for England”, unpublished report provided by NICE.

Feng, Y., Devlin, N.J., Shah, K.K., Mulhern, B. and van Hout, B. (2018), “New methods for modelling EQ-5D_5L value sets: An application to English data”, Health Economics 27, 23-38.

Hernández-Alava, M., Pudney, S. and Wailoo, A. (2018), “Quality review of a proposed EQ-5D-5L value sets for England”, EEPRU report [http://www.eepru.org.uk/wp-content/uploads/2017/11/eepru-report-eq-5d-5l-27-11-18-final.pdf]

Mulhern, B., Norman, R., Street, D.J. and Viney, R. (2019), “One method, many methodological choices: A structured review of discrete choice experiments for health state valuation”, PharmacoEconomics 37, 29-43.

Stolk, E., Ludwig, K., Rand, K., van Hout, B. and Ramos-Goñi, J.M. (2019), “Overview, update and lessons learned from the international EQ-5D-5L valuation work: Version 2 of the EQ-5D-5L valuation protocol”, Value in Health 22, 23-30.

Viney, R., Norman, R., King, M.T., Cronin, P., Street, D.J., Knox, S. and Ratcliffe, J. (2011), “Time Trade-Off Derived EQ-5D Weights for Australia”, Value in Health 14, 928-936.

Viney, R., Norman, R., Brazier, J., Cronin, P., King, M.T., Ratcliffe, J. and Street, D.J. (2014), “An Australian discrete choice experiment to value EQ-5D health states”, Health Economics 23, 729-742.

Yang, Z., Luo, N., Bonsel, G., Busschbach, J. and Stolk, E. (2018), “Selecting health states for EQ-5D-3L valuation studies: Statistical considerations matter”, Value in Health 21, 456-461.