Response to: Quality review of a proposed EQ-5D-5L value set for England

From the study team:

Nancy Devlin[a,b], Ben van Hout[b] (joint principal investigators)
Koonal Shah[a,b], Brendan Mulhern[b,c] and Yan Feng[a] (co-investigators)

[a] Office of Health Economics
[b] School of Health and Related Research, University of Sheffield
[c] Centre for Health Economics Research and Evaluation, University of Technology Sydney

29 November 2018

Hernández-Alava et al. (2018) (hereafter, H-A) report a review of the EQ-5D-5L value set for England reported by us in Devlin et al. (2018)[1] and Feng et al. (2018)[2]. The review was commissioned from the Policy Research Unit in Economic Evaluation in Health and Care Interventions (EEPRU) by the Department of Health and Social Care (DHSC) for England because of the policy relevance of our work; in particular, the impact of using these values to inform National Institute for Health and Care Excellence (NICE) decisions. We support and agree with the principle of using independent experts to review and validate economic modelling that can have a bearing on policy.[3]

H-A are extremely critical of every aspect of the EQ-5D-5L value set for England study. Indeed, it is surprising that they find *almost nothing* in the study design, methods, data or modelling that they approve of.

To provide some context to H-A's review, it is worth noting that:

(a) The project was overseen by a Steering Group chaired by the head of R&D at the Department of Health (DH) for England; its members comprising senior economists from the DH, senior members of the NICE technology appraisal team, a NICE technical appraisal committee chairperson and UK academics with experience in conducting value set studies and their use in economic evaluation. All aspects of the study design, the characteristics of the data generated, and a wide variety of alternative modelling approaches were presented in detail and discussed at Steering Group meetings. The work as reported in our papers in *Health Economics* reflects the guidance we received.

(b) The study protocol was informed by a 10-year programme of methodological research by the EuroQol Group[4] and studies funded by the Medical Research Council.[5] The methods were 'state of the art' when we commenced work in 2012. Value sets

[1] Devlin, N., Shah, K., Feng, Y., Mulhern, B. and van Hout, B., 2018. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Economics*, 27, pp.23-28.
[2] Feng, Y., Devlin, N., Shah, 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, pp.7-22.
[3] Macpherson, N., 2013. *Review of quality assurance of Government analytical models: final report*. London: HM Treasury.
[4] Oppe, M., Devlin, N., van Hout, B., Krabbe, P.F.M. and de Charro, F., 2014. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, 17(4), pp.445-453.
[5] https://www.sheffield.ac.uk/scharr/sections/heds/mvh/pret

generated from the same protocol have been implemented by other HTA systems e.g. in the Netherlands[6] and Canada.[7]

(c) Both valuation study papers went through rigorous peer review for publication in *Health Economics*. This was preceded by extensive efforts to disseminate early findings, in academic and other forums, for discussion and debate.

(d) The differences in quality-adjusted life year (QALY) estimates that arise from implementing our EQ-5D-5L value set, compared to the EQ-5D-3L value set still being used by NICE – concerns about which were the catalyst to the H-A review – were fairly predictable given the well-known methodological problems and unusual characteristics of the UK EQ-5D-3L value set. We are aware of subsequent studies, both in the UK and elsewhere, using the same or a similar EQ-5D-3L valuation protocol, that were unable to replicate the characteristics of that UK EQ-5D-3L value set (e.g. Tsuchiya et al., 2006[8]). For this reason, we consider it likely that any future EQ-5D-5L valuation study using similar methods will have characteristics more like those of the EQ-5D-5L value set we report in Devlin et al. (2018) than those of the EQ-5D-3L value set from 1997. H-A's recommendation to set aside all previous work and 'return to the drawing board' on methods will delay this transition to higher values – but it is unlikely to be avoidable.

Below, we summarise the key points made in the H-A review, and respond briefly to them. Detailed responses are provided in the accompanying technical appendix. We divide H-A's comments into four categories: (1) general concerns about time trade-off (TTO) as a method; (2) general concerns about the research design and study protocol developed for the valuation of EQ-5D-5L internationally; (3) specific concerns about the data collected in England; and (4) specific concerns about the modelling methods we used.

1. Concerns about TTO

H-A: Many participants find it difficult to engage with TTO tasks and carry them out accurately. It is possible that TTO is simply not an adequate basis for valuation.

Our response: There are valid concerns about TTO generally (although we would have emphasised different ones from those that H-A mention). But, equally, there are concerns about *other* available methods for obtaining stated preferences for health states. There has considerable research on mapping out the challenges associated with TTO, and best practice has been carefully defined to address those challenges to a far greater degree than for other candidate methods. Each of the alternative methods has (different) advantages and limitations, and at this moment there no evidence that any other method outperforms TTO. Moreover, there is a clear theoretical relation between TTO and the QALY; the same cannot be said of the discrete choice experiment (DCE) technique. There are ongoing efforts to develop DCE methods (e.g. to include duration), but this is still in an experimental phase. The EuroQol Group's decision to include both TTO and DCE in its

[6] Zorginstituut Nederland., 2016. *Guideline for economic evaluations in healthcare*. Available at: https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-for-economic-evaluations-in-healthcare

[7] CADTH, 2017. *Guidelines for the Economic Evaluation of Health Technologies: Canada — 4th Edition*. Available at: https://www.cadth.ca/dv/guidelines-economic-evaluation-health-technologies-canada-4th-edition

[8] Tsuchiya, A., Brazier, J., Roberts, J. 2006. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *Journal of Health Economics* 25(2):334-346.

protocol for valuing EQ-5D-5L was made on the basis of a careful assessment of the available evidence in this field and to explore new grounds.

2. Concerns about the study design

H-A: The study design/protocol used in the study, the EuroQol Group's 'EQ-VT', was flawed. For example: (a) the health states included should reflect the states encountered in cost effectiveness analyses; (b) the sample size is too small; and (c) members of the general public may not have enough experience of ill health to inform their valuations.

Our response:

a) H-A's arguments about the selection of states suggest a lack of understanding about the role and meaning of 'coverage' in value set studies, and the power of experimental design methods. There is ample evidence that the selection of health states for inclusion in valuation studies should focus on the statistical properties of the health states rather than on how commonly they occur. Please see our technical appendix for a full explanation.

b) The comparison of the sample size in this study with that of the Health Survey for England is specious. The latter is a population health survey, data from which are intended to be used to understand how health differs between population sub-groups and regions. A value set study aims to produce average utilities that are representative of the general public's preferences. It does not need to be powered to produce value sets for sub-groups.

c) As H-A acknowledge, there are good (normative) reasons for selecting a general public sample. This is NICE's requirement for a value set as indicated in its methods guide[9] and our Steering Group was clear that the general public was the relevant sample. We agree that it would be interesting to explore ways of eliciting values informed by experience, but this was outside the scope of our study.

3. Concerns about the data

H-A: (a) The data are 'experimental' as they were generated using an early version of the protocol which has subsequently been improved; (b) data quality fails to meet the standards for policy applications: much of the data is logically inconsistent or otherwise potentially misleading; (c) the response rate is unacceptably low.

Our response:

(a) Our study was undertaken in the first 'wave' of national value sets, together with Spain, Netherlands, China and Canada. These studies represented the state of the art at the time. Experience from them, including data quality monitoring processes developed by our research team, were subsequently further developed and incorporated as standard procedures in later studies and an updated version of the protocol.[10] Together, these have improved some aspects of the data and represent the current state of the art. However, we can predict certain data characteristics that would be observed in any new study using the latest version of EQ-VT, and indeed in any valuation study using

[9] NICE, 2013. *Guide to the methods of technology appraisal 2013*. London: NICE.
[10] Stolk, E., Ludwig, K., Rand, K., van Hout, B. and Ramos-Goñi, J.M., forthcoming. 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*.

alternative protocols. More research, as H-A recommend, can further improve methods, but no method or protocol will ever be 'perfect', and for policy makers, *'waiting for toto* (i.e. the perfect protocol) *may not be a cunning strategy in a practical exercise'* (quote adapted from Sen, 1992).

(b) H-A claim to have found a high inconsistency rate in our data (92.2%). However, the definition of inconsistency H-A use is flawed. H-A define inconsistencies to include ties: for example, where a respondent gives state 55555 the same value as 45555. Such ties can represent entirely plausible preferences, so to judge them as being inconsistent relies on H-A making a strong value judgement. See the technical appendix for further explanation of this and other problems with the 'anomalies' H-A assert. Our research team was aware of the data characteristics (including characteristics that H-A failed to identify, most notably issues relating to interviewer effects) and developed strategies to deal with these, both in subsequent data collection and in the choice of modelling methods. More generally, a number of statements in H-A's review point to misunderstandings about TTO tasks.

(c) The recruitment method involved systematic sampling of dwellings across England. The response rate was in line with what was expected for studies of this kind. H-A compare our response rate with that of the Health Survey for England, but this is inappropriate given the very different nature of the questions included in the survey. A better comparison would be with other TTO studies conducted in England. For example, Rowen et al. (2011)[11] report a response rate of 40%.

4. Concerns about the modelling

H-A: There are many potential problems with the modelling approaches. For instance: (a) the model might be sensitive to the priors that were chosen; (b) the models assume that all TTO responses are 'accurate'; (c) there is dependence between values on a within-respondent basis that we did not take into account.

Our response:

(a) H-A cast aspersions about a number of problems that *could* arise in principle – but do not actually demonstrate these with respect to the data. These problems could have been recited without consulting the data. For example, with respect to priors, we tested the sensitivity of our models to alternative priors and found them to be robust.

(b) The 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.

(c) The analysis of relationships within-respondent 'anomalous' values is flawed: the 'anomalies' are themselves defined by relationships between values, so the analysis is tautological. We identify, in our technical appendix, many other aspects of H-A's review which are problematic.

The study team conducted extensive modelling of the data, only a fraction of which was reported in the papers in *Health Economics*.

The study team has confidence in the value set we have published, for two reasons. First, there is striking similarity in the findings from the TTO results and the DCE results.

[11] Rowen, D., Brazier, J., Young, T., Gaugris, S., Craig, B.M., King, M.T. and Velikova, G., 2011. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value in Health*, 14(5), pp.721-731.

Both methods point towards similar weights for the dimensions and similar values for the levels within the dimensions (as shown in the Figure in the technical appendix to this note). Second, the data show distributions which are broadly similar to those from other countries and we are confident that the statistical approach captures the error distributions in such a way that the mean estimates are a reliable representation of the average values of the public.

Concluding remarks

In summary, H-A's criticisms about our study regarding sample size, state selection, the use of the general public and response rates are not valid. There are comments in H-A's review which suggest a lack of familiarity with methods for eliciting stated preference data and TTO in particular. With respect to data quality issues: there *are* issues with certain characteristics of the data – and these are reflected in our choice of modelling approaches. Moreover, such issues arise in all such studies – so this is a question of degree. Unfortunately, H-A's review offers no insight on that, as it identifies as problematic some data characteristics which could be entirely consistent with people's preferences. On modelling, H-A identify a number of *potential* problems, but do not appear to have ascertained whether these problems actually exist in practice. We are puzzled by this, given that this is where the H-A's expertise lies, and our expectation that this review would focus on analytic modelling.[3]

Many of the issues raised by H-A regarding modelling have already been tested by us, but simply could not be reported in the *Health Economics* papers due to space limitations. Indeed, the study team spent considerable time investigating a wide range of alternative models and their properties. It had been our intention to report a number of these alternative models in our manuscript from the project – but this suggestion was very firmly rejected by our Steering Group, who recommended we publish one 'final' model only, in order to avoid uncertainty and gaming by potential users.

Notwithstanding the limitations of H-A's review, we remain open to the central challenge as to whether these data, and the value set we have produced from them, are fit for use in decision making. We would suggest that the principal question for policy makers is this:

If a new study of EQ-5D-5L values for England were commissioned, would it lead to markedly different values compared to those reported in the current EQ-5D-5L value set?

While issues regarding the accuracy of responses are pertinent, our belief is that the way they have been addressed in the modelling ensures that resulting values are a legitimate reflection of the preferences of the general public in England. We do not anticipate that markedly different values would arise from a newly commissioned study.

The transition away from the EQ-5D-3L and its value set is both necessary and inevitable – the EQ-5D-3L is demonstrably inferior to the EQ-5D-5L.[12] We therefore welcome a constructive dialogue with DHSC and NICE about next steps, and whether additional data collection, using the latest version of the EQ-VT, is warranted. We have a number of practical suggestions for how to proceed, to minimise the cost and protracted delays that H-A's recommendations entail.

[12] Janssen, M.F., Bonsel, G.J. and Luo, N., 2018. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. *Pharmacoeconomics*, 36(6), pp.675-697.

Please refer to the technical appendix where we respond in detail to each of the points raised by H-A.

We are aware that H-A's response to this response has been published on the EEPRU website, alongside the quality review itself. We stand by our concluding remarks in view of H-A's response.

Postscript: Finally, and completely separate from any point of substance: we are sure that others will, like us, be surprised by the tone of H-A's review. It contains content and statements that are of questionable relevance, and appear to have been included for impact or to imply things which are not substantiated (for just one of many examples, see Figure 3.1). Readers can draw their own conclusions about this.