How Health State Utilities Used in Cost-Effectiveness Models Are Currently Identified, Reviewed, and Reported

Roberta Ara, PhD; **John Brazier, PhD**, University of Sheffield, Sheffield, UK; **Andrew Lloyd, DPhil**, Acaster Lloyd Consulting Ltd, Oxford, England, UK; **Hélène Chevrou-Séverac, PhD**, Celgene International, Boudry, Switzerland

KEY POINTS

The reporting standards describing the sources, actual values, justification for final choice, and application of health state utilities in cost-effectiveness models are currently poor.

The selection of health state utilities used are rarely informed by literature reviews, and fundamental details such as the preference-based measure used, the sample size, and details of patients' health condition are rarely reported.

Poor practice has been the norm rather than the exception, but authors of an ISPOR Task Force Report are hopeful that their new checklist and associated recommendations will help lead a rapid change in practice.

Health state utilities (HSU) are used to generate quality-adjusted life years (QALY) in cost-effectiveness models to inform budgetary policy decisions in healthcare. It is recommended that all parameters in these models are informed by a systematic review of the literature. For HSUs in particular, while these may not need to be totally exhaustive, the aim of searches informing the review should be "to identify the breadth of information needs relevant to a model and sufficient information such that efforts to identify further evidence would add nothing to the analysis."[1,2] One rationale behind this recommendation is to enable reviewers to determine that evidence has not been identified "serendipitously, opportunistically or preferentially."[3]

> It has been shown that differences in the preference-based weights, the baseline values, and the methods used to combine evidence, can influence the results generated to such an extent that they could potentially influence a policy decision.

However, a quick Google search quickly identified articles demonstrating this recommendation is not followed by either authors or peer reviewers of articles submitted to academic journals.[4] Indeed, the descriptions of the HSUs used are so lax as to question the validity of some of the evidence, and a skeptical reimbursement agency might be concerned about the motivation of the authors. For example, the entire description of the HSUs used in one model published in 2012 is summarized by: "The utility estimates for each health state were based on a prior estimate." There is no other text in this peer-reviewed article relating to HSUs other than a value and an additional reference in a table. On checking, we discovered there are no HSUs in the source cited within the text. The second source (published in 2007) provided in the table cited a publication from 1997 for the HSUs. After locating this third article, we discovered the evidence used was elicited directly from a small group of patients using standard gamble techniques. So an article published in 2012 incorrectly cited one source, inferring the evidence was relatively recent through citing a 2007 publication, but actually used evidence elicited before 1997, derived from methods that do not meet reference case for many agencies.

In an article published in 2011, the full description of the evidence used for HSUs stated, "Data on utilities for specific health states were identified using the Cost-Effectiveness Analysis Registry as well as Medline searches." It took 3 iterations to identify the original source for some of the HSUs, the earliest of these dated 1986 (the majority were published in early 1990s). Many of the reported HSUs do not match those in either the cited sources or the original source studies. The method or measure used to obtain the HSUs differs for each of the 5 health states in the model. Three different studies provide HSUs elicited directly from patients using time trade-off. The measures used to get HSUs for the remaining 2 health states are less clear: the cited/original sources have a) evidence collected using 6 different generic preference-based measures but it is unclear which actual HSUs were used because the values in the article do not match those in the source; and b) evidence collected using the Quality of Well-Being scale and/or the SF-36. Again, it is unclear which evidence was used because the original source provides data for health dimensions, not the required HSUs. So again, an article published in 2011 cited inappropriate sources and used evidence collected over 2 decades earlier. It also did not provide all the values used and is presenting evidence collected using a variety of methods and measures within the same model.

A third article (published in 2010) provides cynical readers additional food for thought. The authors reported, "All patients in the >

model had an assigned initial baseline utility, which was updated as the patient ages based on values reported in the literature, Utility weights for each model event were based on consensus estimates reported in the literature, as noted in Table 2." The authors did not provide the actual HSUs used for the baseline, nor did they state which method or measure was used to obtain these. As in the previous example, the cited source provides evidence collected from 6 different generic preference-based measures, thus it is not possible to determine what data were used. For the 5 different "model events," either the HSUs reported were not in the referenced source or they did not match the values in the original sources. For example, the authors reported decrements of 0.037 and 0.0175 for coronary artery bypass graft and percutaneous transluminal coronary angioplasty, respectively, while the original sources reported these were 0.042 and 0.059, respectively. Surprisingly, the authors reported an HSU of 0.5 for the "death" health state. Death is an absorbing health state and one wonders if the analyst allows dead people to accumulate 0.5 QALYs just in the first year, or every year over the full lifetime horizon modelled. Finally, they state "multiplicative utility calculations were performed," whereby the 'joint' utility value was the product of the individual utility values." It isn't obvious how this is physically modelled, given that their reported values are disutilites. So this article cited inappropriate sources, did not report all the data used, "tweaked" the evidence with no explanation, and were extremely vague regarding how the HSUs are used to calculate the QALYs.

The above study was funded by industry but the following example demonstrates that poor practice is not limited to that community. Authors of a recent UK National Clinical Guideline made no attempt to identify the most relevant or appropriate evidence simply stating, "We adopted the same utility multipliers for health states as HTA X, 2007 (these were determined following a systematic review), supplemented with values used by Y for states not included in HTA X's model."[5] Whilst acknowledging "they are sourced from multiple different studies, conducted in different settings, and which elicited quality of life preferences using different methods. As a result they may not be entirely consistent." What the authors don't clarify is that several of the values used were estimated or adjusted in some way by authors of the cited article and thus do not match the source values.

The following example illustrates the huge variations in HSUs chosen when evidence is selected in an ad hoc manner, and the difference in methods used when applying these in models. Looking at evidence in a review comparing the HSUs from 6 articles (published in either 2010 or 2011) reporting cost-effectiveness analyses of prophylactic interventions for cardiovascular disease; 4 use constant HSUs (range 0.76 to 1) for the baseline (ie, patients at risk of a cardiovascular event with no history of cardiovascular disease) while 2 use age-adjusted values. The absolute decrements for a coronary event range from -0.05 to -0.24, and for stroke range from -0.11 to -0.5. One of the studies reported that HSUs are multiplied together whilst all the others use absolute decrements. One applied the event decrement for just one year and then allowed the utility to revert to the baseline. Another calculated HSUs by weighting values with the numbers of events observed in a specific trial, then applying zero QALYs after the event. Given the proximity of the publication dates and the similarities in interventions and target populations under appraisal, one might justifiably expect that the HSUs and methods used to apply these would be more consistent across these articles.
We know that the measures and methods used to obtain HSUs make a difference to the values obtained. First, there are substantial differences in the possible HSUs when comparing across the ranges in HSUs obtained from generic preference-based measures. For example, the HUI ranges from -0.36 to 1, while the SF-6D ranges from 0.30 to 1.[6] It has been shown that this can result in different HSUs for the same health condition even when collected from the same people at the same point in time. [7] Second, even the same measure can produce different HSUs from the same sample depending on the individual country preferences weights that are applied to the initial responses to the questionnaire.[8] We also know from multiple sources providing population norms that on average HSUs decline by age irrespective of setting or measure used[9-10], and the mean is never equal to full health irrespective of age or gender.[11] It has also been shown that differences in the preference-based weights, the baseline values, and the methods used to combine evidence, can influence the results generated to such an extent that they could potentially influence a policy decision.[12-13]

A recent review of the literature illustrates that the examples above are not outliers. Looking at recent articles exploring the cost-effectiveness of pharmaceutical interventions in cardiovascular disease, the authors identified 24 studies published since 2015. Of these, just one reported they undertook a literature review to inform the HSUs and just 6 correctly referenced the original sources for all the HSUs. None of the studies provided basic details of the studies or samples used to obtain the HSUs such as the sample size, details of the health condition, timing of data collection, etc. Half did not report which measure was used to collect any of the HSUs, 6 of the studies used HSUs from at least 3 different measures, and just 2 used HSUs from the same measure for all health states in the model. There was substantial variation in the HSUs used for the baselines and large discrepancies in the values used for the individual health states.

This reinforces the importance and need for robust and transparent methods to justify the evidence selection and choice. An ISPOR Task Force has been reviewing the issues encountered when identifying, selecting, and using HSUs in cost-effectiveness models and is expected to publish a report later this year. A checklist is provided for critiquing the appropriateness of the HSU evidence — including search strategies, the review process, and the selection of HSUs used — and the methods that are employed when applying the evidence in the cost-effectiveness model.

The Task Force report is not simply designed to help those wishing to undertake a systematic review of utilities for a cost-effectiveness model. Instead, we think that this report and the SPRUCE checklist also should be used as a tool by reviewers of manuscripts and reports of modelling work to determine their suitability or validity. The Task Force recommends that HTA bodies, academic review groups, model developers, and journal reviewers use the recommendations from the work and the checklist to improve the quality of models. It is clear how important utilities can be to inform cost-effectiveness models, and the examples above indicate the manner in which they are being sourced in published studies. We hope that the Task Force report will help to lead a change in practice among modellers regarding the manner in which this information is gathered and used.

Peer reviewers should not allow cost-effectiveness modellers to assign HSUs to the "caveat" box any longer; they have got away with this for far too long. The poor standards that are currently accepted as the norm do have implications. They undermine the rational for using the cost per QALY which is to facilitate comparison of interventions across diseases and treatments. One wonders if this practice would be so readily accepted if the evidence was describing the clinical effectiveness of an intervention.

REFERENCES

1. Paisley S. Identification of evidence for key parameters in decision-analytic models of cost effectiveness: A description of sources and a recommended minimum search requirement. *Pharmacoeconomics*. 2016;34(6), 597-608. doi:10.1007/s40273-015-0372-x.

2. Glanville J, Paisley S. Searching for evidence for cost-effectiveness decisions. In Shemilt I, Mugford M, Vale L, Donaldson C, Marsh K, eds, *Evidence-Based Decisions and Economics: Health Care, Social Welfare, Education and Criminal Justice*. 2nd ed. Hoboken, NJ: Wiley-Blackwell; 2010: 79-92.

3. Kaltenthaler E, Tappenden P, Paisley S, Squires H. (2011). Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models (NICE DSU Technical Support Document: 13). Sheffield: NICE DSU. http://nicedsu.org.uk/. Accessed June 1, 2018.

4. Brazier J, Ara R, Helene Chevrou-Severac. Utilities in economic evaluation: using best practices where international guidelines provide insufficient detail. ISPOR 19th Annual European Congress; October 29-November 2, 2016; Austria. https://www.ispor.org/Event/ReleasedPresentations/2016Vienna#issuepanelpresentations Accessed June 12, 2018.

5. National Institute for Health and Care Excellence. Cardiovascular disease: risk assessment and reduction, including lipid modification Clinical guideline (NICE Guideline CG181); 2014. Available at: https://www.nice.org.uk/guidance/cg181/ evidence/lipid-modification-update-full-guideline-243786637. Accessed June 3, 2018.

6. Brazier J, Ara R, Rowen DL, Chevrou-Severac H. A review of generic preference-based measures for use in cost-effectiveness models. *Pharmacoeconomics*. 2017;35(suppl 1):21-31.

7. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics*. 2004;13(9):873-884.

8. Janssen B, Szende A. (2014) Population Norms for the EQ-5D. In: Szende A, Janssen B, Cabases J, eds. *Self-Reported Population Health: An International Perspective Based on EQ-5D*. Dordrecht, The Netherlands: Springer, Dordrecht; 2014.

9. Hanmer J, Lawrence WF, Anderson JP, et al. Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores. *Medical Decision Making*. 2006;26(4):391-400.

10. Fryback DG, Dunham NC, Palta M, et al. US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. *Medical Care*. 2007;45(12):1162.

11. Fryback DG, Lawrence WG. Dollars may not buy as many QALYs as we think: A problem with defining quality of life adjustments. *MDM*. 1997;17:276-284.

12. Hernandez Alava M, Wailoo A, Grimm S, et al. EQ-5D-5L versus EQ-5D-3L: The impact on cost effectiveness in the United Kingdom. *Value Health*. 2018;21(1):49-56.

13. Ara R, Brazier JE. Populating an economic model with health state utility values: moving toward better practice. *Value Health*. 2010;13(5):509-518.

Additional Information:

The preceeding article is based on a workshop given at ISPOR 2018.

To view this presentation, go to https://www.ispor.org/docs/default-source/presentations/1386.pdf?sfvrsn=bb54e551_1.

For more information on the Health State Utility Estimates in Cost-Effectiveness Models Task Force, go to: https://www.ispor.org/member-groups/task-forces/health-state-utility-estimates-in-cost-effectiveness-models.