CrossMark

CURRENT OPINION

Is Meta-Analysis for Utility Values Appropriate Given the Potential Impact Different Elicitation Methods Have on Values?

Tessa Peasgood[1] · John Brazier[1]

Abstract A growing number of published articles report estimates from meta-analysis or meta-regression on health state utility values (HSUVs), with a view to providing input into decision-analytic models. Pooling HSUVs is problematic because of the fact that different valuation methods and different preference-based measures (PBMs) can generate different values on exactly the same clinical health state. Existing meta-analyses of HSUVs are characterised by high levels of heterogeneity, and meta-regressions have identified significant (and substantial) impacts arising from the elicitation method used. The use of meta-regression with few utility values and inclusion criteria that extend beyond the required utility value has not helped. There is the potential to explore greater use of mapping between different PBMs and valuation methods prior to data synthesis, which could support greater use of pooling values. Researchers wishing to populate decision-analytic models have a responsibility to incorporate all high-quality evidence available. In relation to HSUVs, greater understanding of the differences between different methods and greater consistency of methodology is required before this can be achieved.

Key Points for Decision Makers

Searching and synthesis of health state utility values (HSUVs) to populate decision models should incorporate all good-quality evidence, but the variability of utility scores by elicitation methods generates a problem for pooling values through meta-analysis.

Stricter inclusion criteria for meta-regression or meta-analysis of HSUVs may help.

There is potential for greater use of mapping algorithms between HSUVs prior to meta-analysis, although careful consideration should be given to the appropriateness of the mapping function and the additional level of uncertainty associated with mapped values.

1 Introduction

The evaluation of healthcare technologies is increasingly reliant upon decision-analytic models. Where quality-adjusted life-years (QALYs) are used as the overall outcome measure for a decision model, each health state included in the model requires a health-related quality-of-life score or health state utility value (HSUV). Good practice in parameter estimation relies on the principles of evidence-based medicine, hence, aims to include all (unbiased) evidence and employ formal evidence synthesis techniques, with systematic review and meta-analysis [1] being the highest level of evidence. That said, the diversity of

✉ Tessa Peasgood
T.Peasgood@sheffield.ac.uk

[1] School of Health and Related Research, University of Sheffield, Regent Court, Regent Street, Sheffield S1 4DA, UK

methods for generating QALYs [2] and the variability across the values generated by these different methods leads to a quandary over whether meta-analysis of utility values will be appropriate.

We are interpreting utility here to mean a measure of the social judgement of the value of a particular health state. Health economists use a number of different methods to extract that value, resulting in the same health state being attributed different (sometimes really quite different) utility scores. This variability arises from four factors: (1) who is asked (and when) to value health states (patients, ex-patients, or members of the public); (2) the technique used to extract preferences and estimate values [the most common being time trade-off, standard gamble (SG), visual analogue scale (VAS) and discrete choice experiment]; (3) different variants of each of the general method (such as the exact question wording, the mode of administration or the use of props); and (4) different preference-based measures (PBMs) or instruments with different descriptive systems, including different items and response options, valued using different methods.

Meta-analysis provides a means to pool data collected across a number of studies and produce a weighted average of the measure of interest, thereby, generating a more precise measure. Most HSUV studies report more than one mean utility value (e.g. patients may complete more than one PBM); consequently any meta-analysis of HSUVs needs to adjust for the fact that these values will be correlated. Given the potential sources of variability of HSUVs, it is unsurprising that conventional tests find that pooled HSUVs reveal considerable heterogeneity (e.g. [3, 4]).

2 Existing Use of Meta-Analysis and Meta-Regression for Utility Values

Meta-regressions [5] allow researchers to explore heterogeneity and the impact of different elicitation methods. Existing meta-regressions (see Table 1) on HSUVs have found substantial differences in values between elicitation methods.

These differences are worryingly large. Indeed, Sturza [6], reporting on her meta-regression for lung cancer, argued that since methodological factors affect utility values, lung cancer researchers "should avoid direct comparisons on lung cancer utility values elicited with dissimilar methods" (p. 691).

Some HSUV synthesis has avoided some of these problems by only using meta-analysis on the EQ-5D (Peasgood et al. [14] for osteoporosis states; Doth et al. [15] for pain states) as this is the measure explicitly preferred by the National Institute for Health and Care Excellence (NICE) [16]. Others have conducted a separate meta-analysis for each overall method or instrument (Liem et al. [17] for renal replacement therapy states; Post et al. [18] for stroke; Mohiuddin and Payne [19] for depression). Whilst a weighted average of EQ-5D values may be adequate for NICE Health Technology Appraisal submissions, for non-NICE submissions, we are left with a decision as to which value to use to populate a decision model. This choice is likely to impact substantially upon the mean values used (e.g. Mohiuddin and Payne [19] reported a pooled SG value for mild depression of 0.69 compared with only 0.56 for the pooled EQ-5D estimate) and on the final incremental cost-effectiveness ratios [20]. Furthermore, a meta-analysis on one particular instrument or

Table 1 Some example coefficients on utility instruments and elicitation methods in meta-regressions

References	Health states	Coefficient on utility instrument/elicitation method (all with $p < 0.05$)	Reference case
Sturza [6]	Lung cancer	Assessment of quality of life (AQoL) [7]: −0.263	SG
McLernon et al. [3]	Chronic liver disease states	TTO: 0.116; transformed VAS: 0.152	EQ-5D
Si et al. [4]	Hip fracture	SG: 0.36	EQ-5D
	Vertebral fracture	Health Utilities Index (HUI) [8]: 0.22	EQ-5D
Lung et al. [9]	Diabetes	TTO or SG: 0.068	EQ-5D
Wyld et al. [10]	Chronic kidney disease	Mapped EQ-5D: −0.14	TTO
Bremner et al. [11]	Prostate cancer	Quality of Well-being (QWB) [12]: −0.09	TTO
Djalalov et al. [13]	Colorectal cancer	SG: −0.13	TTO

SG standard gamble, *TTO* time trade-off, *VAS* visual analogue scale

method results in considerable loss of evidence and information, which goes against the researcher's responsibility to incorporate all high-quality evidence available.

3 Recommendations

How do we use the very best evidence under the circumstances of considerable parameter variation across methodologies? The problem may not be as bad as it at first seems. It may be that these elicitation method differences identified in meta-regressions are inflated. Firstly, some meta-regressions for HSUVs have been conducted on fairly small numbers of utility values. Secondly, meta-regressions have included values that do not appear to be measuring the same thing, i.e. the utility score on a scale of 0 (dead) to 1 (full health) representing how the relevant society views the value of a particular clinical health state.

Meta-regressions with only a few studies and considerable study heterogeneity run the risk of showing false positives [21]; hence, a dummy variable for the elicitation method may appear to be statistically significant when it is not. Whilst there are no hard and fast rules for the appropriate sample size in meta-regression, a ratio of at least ten studies to each covariate is often recommended [5]. For meta-regressions of effectiveness, a minimum of four studies in a categorical subgroup variable has been recommended [22], while more are required to conduct significance testing. Meta-regressions of HSUVs have been conducted with small numbers of utility values (e.g. McLernon et al. [3] conducted a meta-regression with nine covariates and 40 utility values), and some have very few utility values in each category (e.g. Wyld et al. [10] included a covariate for Short Form 6 dimension with only one utility value identified that used this instrument).

The pooling of utility values should only be attempted where the data are valuing the same clinical health state for the appropriate population. The breadth of the health state for which utility values are sought should be dictated by the economic model, and utility values should confidently reflect that exact health state required. Vignettes, which verbally describe a particular (hypothetical) clinical health state to allow individuals who are not in that particular health state to estimate a utility score, may have a useful role in populating economic models in the absence of any other utility values. However, they introduce another layer of uncertainty and may offer no additional benefit when values on the actual desired health state are available. In the meta-regression by Sturza [6], values derived from asking members of the public to link lung cancer vignettes to an EQ-5D state are included alongside direct patient EQ-5D responses without recognition of the superiority of the latter evidence. Making a judgement on whether a study is identifying a utility for the appropriate health state requires detailed information on the exact study population (including study selection, drop out, missing values and clinical diagnosis), and this is unfortunately not always available [19]. When in doubt, preference should be for including only studies where it is reasonable to assume that the utility refers to the desired population.

The pooling of utility values should also only include utilities anchored on the dead to full-health scale. This would exclude values where the top anchor is symptom free (which would exclude some values used in Bremner et al. [11]) or 'normal' rather than full health (which would exclude some values used in Peasgood et al. [23], Tengs and Lin [24, 25] and Sturza [6]). Where there is uncertainty on whether the values really are utility scores, such as when the assessment method is not stated, these should not be included (which would exclude some values used in Tengs and Lin [25]).

It is possible that some PBMs may not adequately identify important aspects of a particular clinical health state. Where there is strong psychometric evidence that a particular instrument lacks validity for the health condition of interest (e.g. see Longworth et al. [26] for a review), a synthesis that excludes those values will be useful for sensitivity analysis.

Where an economic model is to be used to support decision making in a particular country, the desired utility values are those that give the social value of the health state as judged by the relevant population from that country. Utility scores using tariffs from other countries reflect different sets of preferences, and unless it is believed that preferences should be universal, or the value sets are very similar, the rational for pooling utilities that use different country-specific tariffs is not clear. Considerable inter-country differences in the social tariff of the EQ-5D have been identified, with differences varying across the EQ-5D distribution [27]. Including a country-specific tariff dummy, hence, shifting the intercept, will not capture this variability across the distribution or differences in the weight given to different items in the instrument. To include utility data from other countries would require patient level data to enable the appropriate social tariff to be applied or a mapping from one country tariff to another using more sophisticated methods (e.g. [28]).

Even where we have included only utility values on the same clinical health state, the identified utility values are still likely to show variability across instruments and elicitation methods. For PBMs, it is likely that the different descriptive systems drive the variation as much as differences in valuation method [29]. Including the instrument as an intercept term on meta-regression is a limited approach as it does not pick up the relative weights attributed to the different domains within an instrument (including zero if

the item is not included at all). An alternative approach would be to use mapping between instruments, at the aggregate or, if possible, the individual patient level. Whilst mapped values may still differ in terms of both mean and variance compared with direct values (e.g. Wyld et al. [10] found EQ-5D values mapped from Short Form 12 and Short Form 36 to have different values to direct EQ-5D values) and may not be feasible where descriptive content does not substantially overlap, where mapping is possible, the pooling of mapped-utility values could offer a means of generating an estimate that incorporates more of the relevant evidence and has a smaller variance. That said, consideration should be given to the quality of the mapping function, particularly at the ends of the distribution [30], and the appropriateness of the population on which the mapping function was based.

In addition to generating a pooled mean value, consideration also needs to be given to an assessment of uncertainty of the parameter. Ara and Wailoo [31] note that this should incorporate the uncertainty from any mapping functions used, the uncertainty from tariff scores and uncertainty from the output of the descriptive system.

More generally, pooling HSUVs would be aided if there was a greater consistency of valuation methods between instruments. Where instruments adopt different descriptive systems, effort could still be made to generate a social tariff that adopts a standardised methodology. This would facilitate greater understanding of the source of differences between instruments.

The advantages of adopting a systematic review of utility values to populate economic models are clear—the adoption of a clear methodology to follow in terms of searching (see [32]) and transparent reporting of findings. This includes details of study characteristics that would allow modellers to select the most appropriate value [33] for both the main model and any sensitivity analysis. The advantage of including a meta-analysis or meta-regression is the use of all available good-quality evidence in generating the value to be used. Yet even with stricter inclusion criteria (excluding values that are not the appropriate utilities), we are still likely to be left with a considerable degree of heterogeneity across utility values. Higgins [34] has presented the case that in relation to study effect sizes ''any amount of heterogeneity is acceptable, providing both that the predefined eligibility criteria for the meta-analysis are sound and that the data are correct.'' (p. 1158). Where we are aiming to measure the same thing—the social value of a particular health state—we ought to be able to combine values. More work is required on understanding sources of variation in utility values, particularly, variation driven by differences in the descriptive system.

For England and Wales, the current NICE methods guide states that when it is necessary to take HSUVs from the literature "the methods of identification of the data should be systematic and transparent. The justification for choosing a particular data set should be clearly explained. When more than one plausible set of EQ-5D data is available, sensitivity analyses should be carried out to show the impact of the alternative utility values" [16]. This does not then imply a requirement for meta-analysis on EQ-5D values at present. However, given the growing number of publications that incorporate meta-analysis or meta-regression of HSUVs, this guidance may change in the future.

References

1. Sutton AJ, Higgins J. Recent developments in meta-analysis. Stat Med. 2008;27(5):625–50.
2. Brazier J, Ratcliffe J, Salomon J, Tsuchiya A. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.
3. McLernon DJ, Dillon J. Donnan PT. Health-state utilities in liver disease: a systematic review. Med Decis Making. 2008;28(4):582–92.
4. Si L, Winzenberg TM, de Graaff B, Palmer AJ. A systematic review and meta-analysis of utility-based quality of life for osteoporosis-related conditions. Osteoporos Int. 2014;25(8):1987–97.
5. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to meta-analysis. New York: Wiley; 2011.
6. Sturza J. A review and meta-analysis of utility values for lung cancer. Med Decis Making. 2010;30(6):685–93.
7. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the Assessment of Quality of Life (AQoL-8D) multi attribute utility instrument. Patient. 2014;7(1):85–96.
8. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): concepts, measurement properties and applications. Health Qual Life Outcomes. 2003;1(1):54.
9. Lung TW, Hayes AJ, Hayen A, Farmer A, Clarke PM. A meta-analysis of health state valuations for people with diabetes: explaining the variation across methods and implications for economic evaluation. Qual Life Res. 2011;20(10):1669–78.

10. Wyld M, Morton RL, Hayen A, Howard K, Webster AC. A systematic review and meta-analysis of utility-based quality of life in chronic kidney disease treatments. PLoS Med. 2012;9(9):e1001307.

11. Bremner KE, Chong CA, Tomlinson G, Alibhai SM, Krahn MD. A review and meta-analysis of prostate cancer utilities. Med Decis Making. 2007;27(3):288–98.

12. Kaplan R, Bush J, Berry C. Health status: types of validity and the index of wellbeing. Health Serv Res. 1976;11(4):478–507.

13. Djalalov S, Rabeneck L, Tomlinson G, Bremner KE, Hilsden R, Hoch JS. A review and meta-analysis of colorectal cancer utilities. Med Decis Mak. 2014;34(6):809–18.

14. Peasgood T, Herrmann K, Kanis JA, Brazier JE. An updated systematic review of health state utility values for osteoporosis related conditions. Osteoporos Int. 2009;20(6):853–68.

15. Doth AH, Hansson PT, Jensen MP, Taylor RS. The burden of neuropathic pain: a systematic review and meta-analysis of health utilities. Pain. 2010;149(2):338–44.

16. National Institute for Health and Clinical Excellence, Guide to the Methods of Technology Appraisal, 2013.

17. Liem YS, Bosch JL. Myriam Hunink MG. Preference-based quality of life of patients on renal replacement therapy: a systematic review and meta-analysis. Value Health. 2008;11(4):733–41.

18. Post PN, Stiggelbout AM, Wakker PP. The utility of health states after stroke: A systematic review of the literature. Stroke. 2001;32(6):1425–9.

19. Mohiuddin S, Payne K. Utility values for adults with unipolar depression: Systematic review and meta-analysis. Med Decis Making. 2014;34:666–85.

20. Adams R, Craig B, Veale D, et al. The impact of a revised EQ-5D population scoring on preference-based utility scores in an inflammatory arthritis cohort. Value Health. 2011;14(6):921–7.

21. Higgins J, Thompson S. Controlling the risk of spurious findings from meta-regression. Stat Med. 2004;23:1663–82.

22. Fu R, Gartlehner G, Grant M, Shamliyan T, Sedrakyan A, Wilt TJ, Trikalinos TA et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol. 2011;64(11):1187–97.

23. Peasgood T, Ward SE, Brazier J. Health state utility values in breast cancer. Expert Rev Pharmacoecon Outcomes Res. 2010;10(5):553–66.

24. Tengs TO, Lin TH. A meta-analysis of utility estimates for HIV/AIDS. Med Decis Making. 2002;22(6):475–81.

25. Tengs TO, Lin TH. A meta-analysis of quality-of-life estimates for stroke. Pharmacoeconomics. 2003;21(3):191–200.

26. Longworth L, Yang Y, Young T, Hernandez Alva M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P, Keetharuth A, Brazier J. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: systematic review, statistical modelling and survey. Health Technol Assess. 2014;18(9):1–224.

27. Karlsson JA, Nilsson JÅ, Neovius M, Kristensen LE, Gülfe A, Saxne T, Geborek P. National EQ-5D tariffs and quality-adjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. Ann Rheum Dis. 2011;70(12):2163–6.

28. Kharroubi SA, O'Hagan A, Brazier JE. A comparison of United States and United Kingdom EQ-5D health states valuations using a nonparametric Bayesian method. Stat Med. 2010;29(15):1622–34.

29. Richardson J, Iezzi A, Khan MA. Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. Qual Life Res. 2015. doi:10.1007/s11136-015-0926-6.

30. Hernández Alava M, Wailoo A. A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes. Med Decis Making. 2014;34(7):919–30.

31. Ara R, Wailoo A. Using health state utility values in models exploring the cost-effectiveness of health technologies. Value Health. 2012;15:6.

32. Papaioannou D, Brazier J, Paisley S. NICE DSU Technical Support Document 9: the identification, review and synthesis of health state utility values from the literature. 2013.

33. Sampson CJ, Tosh JC, Cheyne CP, Broadbent D, James M. Health state utility values for diabetic retinopathy: protocol for a systematic review and meta-analysis. Syst Rev. 2015;4(1):15.

34. Higgins JP. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol. 2008;37(5):1158–60.