PharmacoEconomics (2018) 36:637–640
https://doi.org/10.1007/s40273-018-0622-9

COMMENTARY

3L, 5L, What the L? A NICE Conundrum

Nancy Devlin[1,2] · John Brazier[2] · A. Simon Pickard[3] · Elly Stolk[4]

Published online: 26 February 2018

1 Introduction

For many years, the National Institute for Health and Care Excellence (NICE) has recommended use of the EQ-5D-3L (3L) [1] and its value set for the UK [2]. Since 2011, an expanded-level instrument, the EQ-5D-5L (5L), has been available [3] and value sets now exist to support its use, including a value set for England [4, 5]. This poses a challenge for NICE. Should it recommend the 5L rather than the 3L?

This is neither a trivial nor merely academic matter: the choice of whether to use the 5L (and English value set) or the 3L (and UK value set) is likely to impact estimates of quality-adjusted life-years (QALYs) and incremental cost-effectiveness ratios (ICERs). The size and direction of that impact will depend on the disease and the nature of the health problems. In general, where technologies improve self-reported health, estimates of QALY gains will often be smaller with the 5L [6]. In contrast, where technologies extend the length of life, estimates of QALY gains will be higher (to varying degrees): each year of additional life is assigned a higher utility. The ultimate impact on health technology assessment (HTA) will depend on whether the differences between the 3L and 5L push ICERs from one side of the cost-effectiveness threshold to the other.

Given the implications for NICE's technology appraisal process, and other decisions informed by EQ-5D data, the Department of Health for England has called for an independent validation of the 5L value set, given its relevance to policy [7].

In 2017, NICE released a 'position statement' [8] stating that:

1. The 3L value set continues to be used for reference-case analyses.
2. Where 5L data have been collected, reference-case analyses should calculate utilities by mapping the 5L descriptive system data onto the 3L value set, using the van Hout et al. [9] mapping function.
3. NICE supports sponsors of prospective clinical studies continuing to use the 5L to collect data on quality of life.

A further position statement is planned for August 2018, to be informed by evidence from various studies underway. These include studies commissioned by the English Department of Health to investigate the implications for past NICE technology appraisals had the 5L been used, and to collect 3L and 5L data in parallel to further improve functions for mapping from one to the other. Other studies, funded by the EuroQol Group, are also underway, investigating various aspects of the relationship between the 3L and 5L across disease areas.

The 3L and its UK value set has occupied a special place in NICE's technology appraisal process since its inception, therefore any transition will inevitably pose challenges; for example, reconciling potential inconsistencies between past and future decisions. Given that evidence will continue to be submitted using both the 3L and 5L for years to come,

The original version of this article was revised due to a retrospective Open Access Order.

✉ Nancy Devlin
ndevlin@ohe.org

[1] Office of Health Economics, 105 Victoria Street, London SW1B 6QT, UK

[2] School of Health and Related Research, University of Sheffield, Sheffield, UK

[3] Department of Pharmacy Systems, Outcomes and Policy, University of Illinois, Chicago, IL, USA

[4] EuroQol Research Foundation, Rotterdam, The Netherlands

if *both* value sets are able to be used, there is a risk of inconsistency between decisions being made in the future. HTA in other countries may also face similar issues.

Given the difficulties with any transition away from the 3L, is there a case for NICE to adopt the 5L as its preferred instrument? Papers in this issue of Pharmacoeconomics, which are cited in this commentary, address that question by investigating comparative performance of the 3L and 5L.

2 3L vs. 5L Descriptive Systems

There are two sources of differences between the 3L and the 5L: [1] the way they *describe* patient health via the health state classifier; and [2] the way they *value* health using preferences obtained from the general public. It is the combination of these two key elements that determines estimates of QALYS. Therefore, an assessment of the merits of the two instruments needs to consider both.

While the 3L and 5L contain the same five dimensions, there are other, important differences between them. Most obviously, the 5L has increased the number of levels from 3 to 5 and the total number of health states described from 243 to 3125. There are also differences in the descriptors, most notably for the worst level of mobility: 'confined to bed' in the 3L has been replaced with 'unable to walk about' in the 5L.

Because of its expanded-level structure, the 5L has the potential to capture the health of subjects more accurately than the 3L, but there is an increase in cognitive burden from offering more choice that may result in lower response rates and perhaps greater measurement error from not knowing which level to choose. Ultimately any measurement benefits from the increased descriptive system must be empirically demonstrated. Papers in this issue, as well as others recently published, suggest these advantages are being realised. Advantages of the 5L over the 3L include:

(a) A reduction in the ceiling effect: The 3L suffers from a ceiling effect, i.e. respondents reporting no problems on any dimension despite (e.g. slight) problems being present. The effect is reinforced by the large gap, in most 3L value sets, between full health and the next best state (in the 3L UK value set, valued at 0.88). In many 3L studies, more than 40% of subjects self-report full health, which dropped by 10% using the 5L [10–12]. Larger and smaller reductions in ceiling effects have been reported elsewhere, reflecting differences in the study samples, e.g. [13–15].

(b) Reduced clustering on just a few states: The lack of granularity in the 3L descriptive system imposes constraints on the self-report of health. Observations tend to cluster on a few health states [15, 16]. The 5L consistently produces considerably more unique health states than the 3L, as shown by Buchholz et al. [17]. For example, Feng et al. [18] reported that just three health states accounted for almost 75% of respondents on the 3L, while a similar proportion of respondents on the 5L were accounted for by 12 health states.

The clustering of descriptive data on the 3L is also reflected in the characteristics of utility-weighted 3L data. 3L health states are relatively far apart on the value scale; for example, the presence or absence of extreme problems in practice predicts almost perfectly whether utility is above or below 0.5. The distribution of utility-weighted 5L data is less prone to this sort of artefactual clustering [16].

(c) Improved ability to discriminate between patient groups/subgroups: The 5L has better discriminative ability, as demonstrated by improved ability to detect differences between subgroups defined by severity at a given sample size [13, 19, 20]. 5L users thus benefit from lower sample size requirements within samples of patients [21]. Although the 3L *seemingly* has better ability to detect differences between patients and a general population group, this is an artefact [13, 17]. The 5L has improved ability to measure health accurately at the top of the scale and therefore provides finer differences between mild ill-health states and full health at the top of the scale, whereas the 3L has much larger steps between levels 2 and 1. As a result, the 3L can overestimate health gains and produce biased ICERs.

(d) Improvements in the 5L with respect to problems with mobility: Abandoning the 3L level 3 descriptor 'confined to bed' constitutes an important improvement in the 5L. Level 3 problems on mobility are rarely observed in 3L data. For example, among patients about to receive hip replacement surgery in the National Health Service, *none* reported a level 3 problem [22]. In effect, in most settings, the 3L only has two dimensions on mobility: no and some problems. Consequently, the 3L will underestimate benefits of treatments that improve severe problems with mobility [13].

Overall, this evidence suggests that the 5L retains the benefits of 3L—its brevity and validity in a wide range of conditions—and produces a more accurate measurement of patient health than the 3L. At the same time, there is no evidence for lower completion rates, and the increase in the number of levels has reduced the amount of variability.

3 5L Versus 3L Utilities

The impact on HTA of the differences between the 3L and 5L descriptive systems becomes apparent only after attaching health state values, the properties of which vary between value sets.

Mulhern et al. [23] point to important differences between the UK 3L and England 5L value sets. Compared with the 3L value set, the entire distribution of the 5L values has shifted to the right and has a shorter tail. The minimum value is higher and there are substantially fewer values <0. While the distribution of 3L values has larger gaps, 5L values show a more even distribution.

Are these differences improvements? Until the external validation of the England 5L value set concludes, the jury is still out. But it is instructive to reflect on the causes of these differences.

First, there are differences in the preferences data they are based on. Both used time trade-off (TTO), but values <0 were elicited very differently. Furthermore, the 5L value set uses both TTO and discrete choice experiment (DCE) data. The value sets were generated at different points in time (1997 vs. 2017) and preferences for health may have changed in the interval—a potential reason to revisit value sets for all preference-based measures [24]. Furthermore, the 5L valuation protocol [25] benefited from two decades of methodological advances. Paired with the additional change in descriptors in the mobility dimension, there is no reason to expect that 3L and 5L would produce the same values.

Second, there are differences in the way the value sets are modelled. While the 3L value set model has the merit of simplicity, the 5L value set uses innovative modelling approaches, e.g. addressing preference heterogeneity and combining TTO and DCE data via 'hybrid' models [5]. The realization that simple models can produce biased values has led to advances in modelling TTO data [26, 27]. With 5L valuation studies being conducted in the digital era, researchers have access to metadata (e.g. respondents' patterns of trading), which reveal the influence of the TTO design task on values. New methods can control for this.

In comparing the UK 3L and England 5L value sets, it should be noted that some of these differences arise because the former is somewhat unusual (e.g. compared with most other countries' 3L value sets). It has a high percentage of health states with negative values (over one-third of the 243 states have values <0, indicating that, on average, the general public considered them 'worse than being dead'). A 1996 UK replication study by Kind and Macran [28], using the same protocol, found just 12% of states were <0. In comparison, 5% of the values in the England 5L value set are <0. Similarly, the minimum value in the UK 3L value set (-0.594) is much lower than that in the replication study (-0.126). In comparison, the minimum value in the England 5L value set study is -0.285. Similar conclusions with respect to the UK 3L value set were also reported by Tsuchiya et al. [29].

In summary, there are many reasons why the UK 3L and England 5L value sets are different. Some of these reasons apply to *all* countries with 3L and 5L value sets, while others are specific to the UK/England case. The England 5L value set was one of the first 5L value set studies undertaken internationally, and learning from it benefitted subsequent studies. For example, detailed reporting of issues observed in the English data led to improvements in the protocol and data quality monitoring in subsequent studies. Nevertheless, comparison of the England 5L value set with other 5L value sets shows a broad level of agreement between them [30].

4 Concluding Remarks

The 5L was developed to improve on an instrument (the 3L), which has been widely used and has validity in a wide range of conditions. As summarised in this commentary, the 5L has a number of advantages over the 3L as a measure of self-reported health.

NICE's position statement does *not* signal a concern about the 5L descriptive system. Rather, it is a reaction to a governmental requirement to validate the England 5L value set that accompanies it.

As a decision-making entity that bears responsibility to a range of stakeholders, NICE is responding to the availability of a 5L value set with understandable care. QALY gains are often very small, therefore ICERs can be highly sensitive to the choice of value set. This underlines the importance of ensuring that any new value set is valid for use in decisions about cost effectiveness. Until that work concludes, the status of the England 5L value set is (to coin a NICE phrase) 'in research only' rather than 'recommended'.

However, what is increasingly clear is that much of the difference noted between the UK 3L and England 5L value sets is attributable to characteristics of the former. It is fairly unlikely that any new value set, whether that be for the 3L or the 5L, will have the same properties as the existing UK 3L value set, suggesting that the transitional challenge facing NICE is unavoidable. The papers in this issue help to shed light on the comparability of the 3L and 5L, and provide evidence to help inform that transition.

Compliance with Ethical Standards

Funding Nancy Devlin received funding from the EuroQol Research Foundation for her contribution to writing this commentary.

Conflicts of interest Nancy Devlin, John Brazier, A. Simon Pickard and Elly Stolk are members of the EuroQol Group. Elly Stolk is employed by the EuroQol Research Foundation as Scientific Team Leader, and Nancy Devlin was Principal Investigator of the 5L value set study for England.

References

1. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. https://www.nice.org.uk/guidance/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781. Accessed 23 Feb 2018.
2. Dolan P. Modelling valuations for Euroqol health states. Med Care. 1997;35(11):1095–108.
3. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727–36.
4. Devlin NJ, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health related quality of life: an EQ-5D-5L value set for England. Health Econ. 2017. http://onlinelibrary.wiley.com/doi/10.1002/hec.3564/full. Accessed 15 Nov 2017 **(Epub 22 Aug 2017)**.
5. 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. 2017. http://onlinelibrary.wiley.com/doi/10.1002/hec.3560/full. Accessed 15 Nov 2017 **(Epub 18 Aug 2017)**.
6. Hernandez Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, 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.
7. HM Treasury. Review of quality assurance of Government analytical models: final report. 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/206946/review_of_qa_of_govt_analytical_models_final_report_040313.pdf. Accessed 23 Feb 2018.
8. NICE. Position statement on use of the EQ-5D-5L valuation set. 2017. https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf. Accessed 14 Nov 2017.
9. Van Hout B, Janssen MF, Feng Y-S, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15(5):708–15.
10. Craig BM, Pickard AS, Lubetkin EI. Health problems are more common, but less severe when measured using newer EQ-5D versions. J Clin Epidemiology. 2014;67(1):93–9.
11. Mukuria C, Ara R, van Hout B. A comparison of the performance of the UK EQ-5D-3L and EQ-5D-5L using evidence from the GP patient survey. In: Paper presented at the 34th plenary meeting of the EuroQol Group, 20–23 Sept 2017, Barcelona.
12. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D-3L and 5L: measurement properties and association with chronic conditions and multi-morbidity in the general population. Health Qual Life Outcomes. 2014;16:12–74.
13. Janssen MF, Bonsel G, Luo N. Is EQ-5D-5L better than EQ-5D-3L? A head-to-head comparison of descriptive systems and value sets from seven countries. Pharmacoeconomics. 2018. https://doi.org/10.1007/s40273-018-0623-8.
14. Janssen B, Pickard AS, Golicki A, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res. 2013;22(7):1717–27.
15. Zamora B, Parkin D, Feng Y, Bateman A, Herdman M, Devlin N. New methods for analysing the distribution of EQ-5D data. In: Paper presented at the 34th plenary meeting of the EuroQol Group, 20–23 Sept 2017, Barcelona.
16. Parkin D, Devlin N, Feng Y. What determines the shape of an EQ-5D distribution? Med Decis Mak. 2016;36(8):941–51.
17. Buchholz I, Janssen B, Kohlman T, Feng Y-S. A systematic review on studies comparing the measurement properties of the three-level and the five-level version of the EQ-5D. Pharmacoeconomics **(submitted manuscript)**.
18. Feng Y, Devlin NJ, Herdman M. Assessing the health of the general population in England: how do the three- and five-level versions of EQ-5D compare? Health Qual Life Res. 2015;13:171.
19. Pan CW, Sun HP, Wang X, Ma Q, Xu Y, Luo N, Wang P. The EQ-5D-5L index score is more discriminative than the EQ-5D-3L index score in diabetes patients. Qual Life Res. 2015;24(7):1767–74.
20. Wang P, Luo N, Thumboo J. The EQ-5D-5L is more discriminative than the EQ-5D-3L in patients with diabetes in Singapore. Value Health Reg Issues. 2016;9:57–62.
21. Pickard AS, de Leon MC, Kohlmann T, Cella D. Psychometric comparison of the standard EQ-5D to a 5-level version in cancer patients. Med Care. 2007;45(3):259–63.
22. Oppe M, Devlin N, Black N. Comparison of the underlying constructs of EQ-5D and Oxford Hip Score: implications for mapping. Value Health. 2011;14:884–91.
23. Mulhern B, Feng Y, Shah K, Janssen B, Herdman M, van Hout B, et al. Comparing the UK E-5D-3L and English EQ-5D-5L value sets. Pharmacoeconomics. 2018. https://doi.org/10.1007/s40273-018-0628-3.
24. Pickard AS. Is it time to update societal value sets for preference-based measures of health? Pharmacoeconomics. 2015;33:191–2.
25. Oppe M, Devlin NJ, van Hout B, Krabbe PFM, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445–53.
26. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJ, Stolk E. Quality control process for EQ-5D-5L valuation studies. Value Health. 2017;20(3):466–73.
27. Stolk E, Ludwig K, Rand Hendriksen K, Van Hout B, Ramos Goñi JM. Overview, update and lessons learned from the international EQ-5D-5L. ViH **(submitted manuscript)**.
28. Kind P, Macran S. Valuing EQ-5D health states using a modified MVH protocol: preliminary results. In: Proceedings of the 16th plenary meeting of the EuroQol Group, 6–9 Nov 1999, Sitges.
29. Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to EQ-5D and SF-6D value sets. J Health Econ. 2006;25(2):334–46.
30. Olsen JA, Lamu AN, Cairns J. In search of a common currency: a comparison of seven EQ-5D-5L value sets. Health Econ. 2018;27(1):39–49.