Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets

Brendan Mulhern[1], Yan Feng[2], Koonal Shah[2], MF Janssen[3], Michael Herdman[2], Ben van Hout[4], Nancy Devlin[2]

1 University of Technology Sydney, Centre for Health Economics Research and Evaluation, 1-59 Quay St, Haymarket, Sydney, NSW 2000, Australia
2 Office of Health Economics, Southside, 105 Victoria Street, London, SW1E 6QT, UK
3 Department of Medical Psychology and Psychotherapy, Erasmus MC, Erasmus University, PO Box 2040, 3000 CA, Rotterdam, The Netherlands
4 Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St, S1 4DA, UK

Corresponding author:
Brendan Mulhern
e-mail: Brendan.mulhern@chere.uts.edu.au
Tel no: +61 2 9514 4725
Fax no: +61 2 9514 4730

Running header: Comparing EQ-5D value sets

Acknowledgements:
We thank the EuroQol Group for providing access to the patient data used in this study, and John Brazier for comments on an earlier draft.

Abstract:

Background:

Three EQ-5D value sets (EQ-5D-3L, crosswalk and EQ-5D-5L) are now available for cost utility analysis in the UK and/or England. The value sets' characteristics differ, and it is important to assess the implications of these differences. The aim of this paper is to compare the three value sets.

Methods:

We carried out analysis comparing the predicted values from each value set, and investigated how differences in health on the descriptive system is reflected in the utility score by assessing the value of adjacent states. We also assessed differences in values using data from patients who completed both EQ-5D-3L and EQ-5D-5L.

Results:

The distribution of the value sets systematically differed. EQ-5D-5L values were higher than EQ-5D-3L/crosswalk values. The overall range and difference between adjacent states was smaller. In the patient data, the EQ-5D-5L produced higher values across all conditions and there was some evidence that the value sets rank different health conditions in a similar severity order.

Conclusions:

There are important differences between the value sets. Due to the smaller range of EQ-5D-5L values the possible change in quality-adjusted life years (QALYs) might be reduced, but they will apply to both control and intervention groups, and will depend on whether the gain is in quality of life, survival, or both. The increased sensitivity of EQ-5D-5L may also favour QALY gains even if the changes in utility are smaller. Further work should assess the impact of the different value sets on cost effectiveness by repeating the analysis on clinical trial data.

Keywords: EQ-5D, Quality Adjusted Life Year, Utilities, psychometrics

Key points for decision makers:

- There are differences between the UK EQ-5D-3L and English EQ-5D-5L value sets.
- The choice of value set will have implications for the decision making process carried out by the National Institute of Health and Care Excellence

1. Introduction

In the economic evaluation of health interventions the quality adjusted life year (QALY) is a commonly used metric that combines length and quality of life into a single figure. The quality, or utility, weight used in the estimation of QALYs is anchored on a full health (1) to dead (0) scale, with negative values assigned to health states considered worse than dead. Utility values for health states associated with a particular condition or disease can be derived in several ways, one of which is via the use of preference based measures (PBM) of health. Of currently available PBMs, the EQ-5D [1,2] is the most widely used.

EQ-5D classifies health on five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. The original version of the EQ-5D (described as EQ-5D-3L) included three severity levels (none, some, extreme/unable to)[1], thereby describing (3^5 =) 243 health states. In the UK, utility values for EQ-5D-3L health states were derived using the Time Trade Off (TTO) preference elicitation technique [3]. The resulting 'value set' has been widely influential, and is preferred by the National Institute for Health and Care Excellence (NICE) for use in the cost utility analysis of health interventions [4]. EQ-5D-3L values are also accepted by reimbursement agencies worldwide including the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia [5] and the Canadian Agency for Drugs and Technology in Health (CADTH) in Canada [6]. The instrument itself is also used in a wide range of settings including population health surveys and routine clinical practice [7].

Notwithstanding their widespread use, research has suggested that both the EQ-5D-3L descriptive system and value set have a number of limitations. Regarding the descriptive system, it has been shown that the EQ-5D-3L is not sensitive to the important quality of life impacts of all conditions [8,9]. It may also not be sensitive to smaller changes in health as it only has three response levels in each dimension and, in general public and some patient samples, a substantial proportion of respondents report themselves as being in the best health state, i.e. no problems on any dimension (11111). This is known as a ceiling effect [10]. Regarding the value set, the procedure and modelling used to elicit values for worse than dead health states has been criticised [11]. Furthermore, the EQ-5D-3L valuation data was collected in 1993, and population preferences for different aspects of health and quality of life may have changed in this time given advances in treatment and care. Social and environmental changes may also be important.

In an effort to improve the instrument's sensitivity and reduce the ceiling effect, a five-level descriptive system, the EQ-5D-5L [12], was developed. The new instrument includes five response levels (none, slight, moderate, severe, extreme/unable to). The wording was also standardised across dimensions so that the worst level of mobility was changed from 'confined to bed' to 'unable to walk about' which is in line with the severity indicators used for the other functioning dimensions (self care and usual activities). The intermediate severity level was also standardised to be 'moderate'. The EQ-5D-5L

[1] In the EQ-5D-3L, level 3 mobility was described as 'confined to bed' not 'unable to'

increases the number of states described to (5^5=) 3,125. Research has shown improved measurement properties of the EQ-5D-5L descriptive system across a number of patient samples when compared to the EQ-5D-3L [13].

One consequence of this initiative was the need to develop value sets for the new descriptive system the reflect more up-to-date preferences of the population for health and quality of life, and this resulted in two separate developments. Firstly, an interim 'crosswalk' value set was developed so that EQ-5D-3L values could be used to predict EQ-5D-5L values [14]. Secondly, in order to elicit values for health states generated by the EQ-5D-5L descriptive system, a new valuation protocol combining TTO and Discrete Choice Experiment (DCE) methods was developed [15]. This protocol used a 'composite' TTO approach combining standard and 'lead time' TTO [15-17]. In England, health states generated by the EQ-5D-5L were valued during 2012 and 2013 using this protocol and subsequently modelled using newly developed techniques which combined TTO and DCE data in a hybrid model to produce an EQ-5D-5L value set [18,19].

Three EQ-5D value sets are therefore now available for use in cost utility analysis in the UK and/or England, those being the EQ-5D-3L value set, the crosswalk value set mapping the EQ-5D-5L descriptive system onto the EQ-5D-3L value set, and the EQ-5D-5L value set. The first two of these were developed based on valuations from respondents in the UK while the latter was based on valuations from respondents in England only. However, this is only one way in which they differ. As noted, they are also based on different descriptive systems, valuation protocols, and modelling methods. Given widespread and increasing use of the EQ-5D-5L in decision making, it is important to systematically assess the differences between the value sets, and the implications of the new values. For example in recent work it has been found that quality of life changes are valued less using the EQ-5D-5L value set [20]. At the end of 2017, NICE released a position statement regarding the use of the EQ-5D-5L stating that "the mapping function developed by van Hout et al. [14] [i.e. the crosswalk value set] should be used for reference-case analyses" until its position is reviewed in 2018 [21]. This means that the UK crosswalk is currently important in health technology assessment (HTA) carried out by NICE, and the results of studies comparing the new EQ-5D-5L value set with the crosswalk and EQ-5D-3L will inform future decisions about which measure to use. Therefore the aim of this paper is to add to the literature in this area by comparing the UK EQ-5D-3L and English EQ-5D-5L value sets, and the EQ-5D-5L and crosswalk value sets.

2. Methods

2.1. The value sets

In the sections below, EQ-5D health states are described using five digits corresponding to each dimension and each level. The dimensions are listed in the order presented on the questionnaire (Mobility-Self care-Usual activities-Pain/discomfort-Anxiety/depression). For the EQ-5D-3L, 1 represents no problems, 2 some problems, and 3 extreme problems/confined to bed. Therefore EQ-5D-3L state 22222 describes some problems on each of the five dimensions. For the EQ-5D-5L, 1 represents no problems, 2 slight problems, 3 moderate problems, 4 severe problems, and 5 extreme problems/unable to. Therefore EQ-5D-5L state 22222 describes slight problems on each dimension.

2.1.1. EQ-5D-3L

The UK EQ-5D-3L value set [3] was developed using data collected in 1993 from 2,997 general population respondents who were sampled from the postcode address file. Respondents were recruited to be representative of the non institutionalised adult population of England, Scotland and Wales, and had similar characteristics to the UK General Health Survey sample (Dolan et al 1996). Each respondent completed a face-to-face interview and valued 13 states (12 EQ-5D-3L profiles plus 'unconscious') using TTO which included one procedure for states valued better than dead, and a different process for states valued worse than dead. In total, 42 of the 243 EQ-5D-3L states were valued, with an overrepresentation of the mildest health states. The data were modelled using additive generalised least squares (GLS) regression to produce a value set ranging from 1 (for the best state, 11111) to -0.594 (for the worst state, 33333), with 34.6% of states valued as worse than dead. The model includes a constant subtracted for any move away from full health, a further decrement for each move away from 'no problems' for each dimension, and an additional term that is subtracted if any dimension is at the worst level (known as the N3 term). The value set also has a large change in utility between 11111 and the next best state (11211 which is scored at 0.883).

2.1.2. Crosswalk

Crosswalk value sets were developed by van Hout et al [14] from a multicountry study of respondents who completed both the EQ-5D-3L and EQ-5D-5L in 2010. The crosswalk used a non-parametric response mapping method to predict values that are linked to the EQ-5D-3L value set The decrements for the 'equivalent' levels of the two descriptive systems are the same. This means that the decrements for level 3 of the EQ-5D-5L (moderate problems) are the same as level 2 of the EQ-5D-3L (some problems), and those for level 5 of the EQ-5D-5L are the same as level 3 of the EQ-5D-3L. This means that the range of values is the same (55555 on the EQ-5D-5L has the same value as 33333 on the EQ-5D-3L, and, an example intermediate state, 35353 on the EQ-5D-5L has the same value as 23232 on the EQ-5D-3L). The crosswalk can link EQ-5D-5L data to a range of existing international EQ-5D-3L value sets. For the purposes of this paper, we focus on the crosswalk to the UK value set.

2.1.3. EQ-5D-5L

The English EQ-5D-5L value set [18] was developed from 996 members of the general population who were purposively sampled from the Postcode Address File. In contrast to the EQ-5D-3L, respondents representative of the population of England (as opposed to the wider UK; a UK value set reflecting the preferences of respondents in England, Scotland, Wales and Northern Ireland is due to follow) were recruited. The sample used differed slightly to the actual population as there were more older and retired people. Preferences were elicited using computer-assisted face-to-face valuation interviews that were conducted in 2012 and 2013. Respondents valued 10 EQ-5D-5L states using composite TTO [15,16], and completed seven DCE paired comparison tasks. In total 86 states were valued in the TTO exercise and 196 pairs in the DCE tasks. The data was modelled using heterogeneous hybrid approaches combining the TTO and DCE data [19]. The resulting tariff ranges from 1 to -0.285, with 5.1% of the states valued as worse than dead. The model includes a decrement for each dimension for each move away from full health, and an extra 'scalar' coefficient. The range of values is therefore smaller than for the EQ-5D-3L, despite the considerable increase in the number of possible health states. The value of the mildest health states other than 11111 (12111 and 11211) is 0.950.

2.2. *Analysis*

We carried out analysis comparing the predicted values from each of the three value sets, and also using patient reported EQ-5D-3L and EQ-5D-5L data. The patient data was taken from the crosswalk development study dataset where all respondents self-reported their health using both the EQ-5D-3L and EQ-5D-5L descriptive systems thereby enabling direct comparisons. The key comparisons carried out were between the EQ-5D-3L and EQ-5D-5L value sets, and the EQ-5D-5L value set and the crosswalk tariff.

2.3. Comparison of predicted values

2.3.1. Comparing value set models

Firstly we compared the coefficient models used to calculate the values. This was done to assess the overall magnitude of the coefficients for each dimension, and the impact of the various interaction coefficients included in each model on the values produced. We also compared the process for calculating values using an example health state.

2.3.2. Comparing value set characteristics

We assessed a range of descriptive statistics of the possible theoretical values (i.e. 243 for the EQ-5D-3L and 3,125 for the EQ-5D-5L). This included the value set range, the percentage of states valued as worse than dead, and the state with the smallest utility decrement from 11111. We looked at the modality of the overall distributions using kernel density histograms, and compared the values of selected states to demonstrate differences between the value sets.

2.3.3. Comparing value set characteristics for matched states

We carried out a comparative analysis on the states that are comparable across the EQ-5D-3L and EQ-5D-5L (i.e. the matched 243 states). The crosswalk value set is not relevant here as for these states

the values are the same as the EQ-5D-3L tariff due to the response mapping procedure used. We considered comparable states to be those from the intermediate levels of the EQ-5D-5L descriptive system (i.e. none, moderate and extreme/unable to) which, to a certain extent, 'match' the three-level states (as an example the EQ-5D-3L state 12321 is defined as comparable to 13531 on the EQ-5D-5L). We assessed similarities and differences, both for individual states and at the overall level, to highlight where the largest differences occur across the value sets.

2.3.4. Comparing differences in utility between adjacent states

Analysis was also carried out to understand how changes in severity on the descriptive system are reflected by changes in utility. This was done by assessing the values of adjacent states within the descriptive system, and comparing the differences across the three value sets. An adjacent state pair was defined as having one dimension with a one-level difference (for example calculating the change in utility between 21111 and 11111). This was done for states where only one dimension changed at a time, so we focused on the change in utility between level 3/5 and level 1 on one dimension, with the other four dimensions held at the same level. For example, for mobility we compared the increase in utility between 51111, 41111, 31111, 21111 and 11111, and we repeated this for all five dimensions. The magnitude of the change between all level changes, and the matched states, was assessed. This analysis reflects the coefficient decrements in a different way and provides an insight about how change in self-reported health would lead to change in utility in the absence of longitudinal data.

2.4. Analysis on patient data

2.4.1. Data used:

The data used to develop the crosswalk value sets were used for the analysis. The data were collected online across a range of patient groups with different health conditions who completed both the EQ-5D-3L and EQ-5D-5L descriptive systems. More information about the data collection procedure is provided in van Hout et al [14]. Respondents from seven countries took part, but the analysis reported here used only the English and Scottish data. The characteristics of the 1,501 respondents included are reported in Table 1.

2.4.2. Comparing the descriptive system and value sets

Firstly, we compared the number of respondents reporting each level of the two descriptive systems. This was done to understand how the addition of the two extra levels changes response patterns. We compared the values using density plots, and also by assessing the scores overall and across patient groups (with the exception of those with a sample size of less than 50) using one way ANOVA and mean difference statistics. We also compared the agreement between the value sets using Bland Altman plots [22]. These present the mean of two scores on the x axis and the difference on the y axis, with lines indicating the upper and lower limits of agreement (calculated as the mean difference +/- 1.96 x standard deviation) added. Agreement across the full severity range can then be assessed, with points outside the limits indicative of outliers.

3. Results

3.1. Comparison of predicted values

3.1.1. Comparing value set models

The models used to derive EQ-5D-3L and EQ-5D-5L values are displayed in Table 2. In each case, the coefficient decrements are larger for the more severe levels of each dimension and are therefore ordered as expected. Both models include a constant term, and in the EQ-5D-3L this involves a decrement of 0.081 for the move away from the best health state (11111). The EQ-5D-5L constant is 1, and the coefficients are the mean coefficients from the modelling process after the application of the latent class adjustments. The magnitude of the dimension level coefficients between the EQ-5D-3L and EQ-5D-5L varies (for example, pain/discomfort has a larger overall decrement on the EQ-5D-3L and anxiety/depression has a larger decrement on the EQ-5D-5L). The EQ-5D-3L N3 term is an extra decrement when at least one of the levels is at the most severe (i.e. level 3), and therefore this reduces the value of the more severe states. Table 2 also displays how to calculate a value for a state. The calculation of the value for EQ-5D-5L state 21223 and the equivalent EQ-5D-5L state 31335, and shows that the EQ-5D-3L value is substantially lower (0.186 vs. 0.488).

3.1.2. Comparing value set characteristics

Table 3 (adapted from Devlin et al [18]) compares the descriptive characteristics of the three value sets. The EQ-5D-5L value set has a higher value for the worst health state and substantially fewer worse than dead values. Also, the decrement from the best (11111) to next best health state (11211) is smaller for the EQ-5D-5L value set. This is expected given differences in labelling (e.g. 11211 describes 'slight' problems performing usual activities in the five-level instrument and 'some' problems in the three-level version). In all three value sets, pain/discomfort has the largest overall decrement (but not at the less severe levels), while self-care and usual activities have the smallest.

Figure 1 compares all unique theoretical values for the three value sets. The results demonstrate that the range for the EQ-5D-3L and crosswalk is different from the EQ-5D-5L. The large coefficients for level 3 on the EQ-5D-3L (and the impact of the N3 term) means that there is a higher density of lower values. The EQ-5D-5L is unimodal, whereas the EQ-5D-3L has multiple clusters as has previously been observed [23].

3.1.3. Comparing value set characteristics for matched states

Figure 2 displays the values of the comparable states from the EQ-5D-3L and the EQ-5D-5L value sets ordered by descending EQ-5D-5L value. EQ-5D-3L values are consistently lower across the full severity range. Figure 3 shows a histogram of the differences for each comparable state across the value sets, and a box plot of the mean difference by utility score category as a proxy for severity (1 to 0.500; 0.499 to 0.200; 0.199 to 0; <0). The mean difference is large overall at 0.312 (sd 0.102; range 0 to 0.484), and significantly increases as severity increases ($F_{3,239} = 196.0$, $p < 0.001$). Only 16 (6.6%) of 243 states have a mean difference smaller than 0.1, and 40 (16.4%) states have a difference of at least 0.4. The state with the largest difference is 32131 (53151 on EQ-5D-5L) (0.484) and the state with the smallest difference (excluding the best state is 11212 (11313 on EQ-5D-5L) (0.023).

3.1.4. Comparing differences in utility between adjacent states

Table 4 displays the change in utility between adjacent and matched states. Comparisons of the matched states demonstrates that the change in adjacent states is substantially larger for the three-level tariff across all five dimensions. This may suggest that the use of the EQ-5D-3L value set would tend to result in larger QALY gains for purely quality of life-improving interventions. Regarding EQ-5D-5L, the largest change in value occurs in the move from severe (level 4) to moderate (level 3) reported health problems. In contrast, the largest change in the crosswalk value set is between extreme/unable to (5) and severe (4) which is comparatively small in the EQ-5D-5L value set. The change in the crosswalk values from slight (2) to no problems (1) is larger than for EQ-5D-5L. This means that interventions resulting in an improvement in both mild and more severe health may result in larger QALY gains if the crosswalk values were used.

3.2. Comparisons using patient data

3.2.1. Comparing the descriptive system and value sets

Table 5 displays the dimension level responses to the EQ-5D-3L and EQ-5D-5L and shows that the largest impact of the addition of the two intermediate levels (slight and severe) is to spread the 'some' responses on the EQ-5D-3L between levels 2 to 4 on the EQ-5D-5L. The introduction of 'slight' modestly reduces the ceiling effect as respondents move away from reporting no problems given the increased sensitivity. There is clear dispersion of scores from 'some' on the EQ-5D-3L across 'slight', 'moderate' and 'severe' on the EQ-5D-5L.

Figure 4 compares the EQ-5D-3L, EQ-5D-5L and EQ-5D-5L crosswalk values. For the EQ-5D-3L there is not only a large decrease in values in the very mild area (due to the upper gap reflected by the large constant), but also in the moderate area around the values 0.25 to 0.45. In contrast the EQ-5D-5L has a smoother distribution. This reflects a benefit of EQ-5D-5L: the increased sensitivity results in a much smoother transition between adjacent values that are closer together than on the EQ-5D-3L. The crosswalk value set distribution is more similar to the EQ-5D-5L, and the lack of EQ-5D-3L values in the range between approximately 0.25 and 0.45 is not apparent.

Figure 5 compares the EQ-5D-3L and crosswalk with the EQ-5D-5L and shows that there are differences in values across the entire severity scale, but greater variation for more severe health states (where the mean utility value is lower for the EQ-5D-3L and crosswalk). Figure 6 displays Bland Altman plots comparing EQ-5D-3L and EQ-5D-5L, and EQ-5D-5L and crosswalk scores. There is evidence of disagreement between values across the severity scale, where the difference is outside the +/- 2 sd range. Disagreement means more diverse utility scores for states of a similar severity.

The mean difference between the EQ-5D-3L and EQ-5D-5L values as reported by the patient sample is 0.073 (range -0.944 to 0.880 calculated as EQ-5D-5L minus EQ-5D-3L). Some respondents gave apparent inconsistent responses and this results in the wide range overall. For example, the difference

of -0.944 results from a patient reporting 21111 on EQ-5D-3L and 44444 on EQ-5D-5L. Comparing the EQ-5D-5L and crosswalk value sets, the mean absolute difference is 0.085 and ranges from 0.002 for the states with the smallest non zero difference (44431, 42433, 43441 and 41231) to 0.429 (for state 51131).

Table 6 compares the value set scores overall and across the different health conditions, with significance statistics reported for the conditions with more than 50 patients. As would be expected, the EQ-5D-5L values are higher, and the difference is significant for the four conditions with the largest sample size (COPD, heart problems, arthritis and depression). Of the four conditions with a sample size of between 50 and 100, the difference tends towards significance for stroke and back pain but not for ADHD or rheumatoid arthritis. The percentage of states worse than dead overall and also across each condition is lower for the EQ-5D-5L. Table 6 also displays the rank order of the severity of the conditions according to the mean utility values. There is evidence of consistency for seven of the 12 conditions, including the most (Parkinson's disease) and third most (back pain) severe conditions, and the five least severe (ADHD, breathing problems, arthritis, depression and diabetes). The most variable condition is multiple sclerosis, which is second most severe according to the EQ-5D-3L, but fifth and equal sixth overall according to the crosswalk and EQ-5D-5L value sets respectively.

4. Discussion

We have compared three EQ-5D value sets that can be used to support HTA in the UK. The comparison firstly investigated differences in the 'theoretical' values possible from the value sets for health states matched across the EQ-5D-3L and EQ-5D-5L descriptive systems and secondly compared values observed in patient data.

Regarding the theoretical values, the results demonstrate that there are differences between the EQ-5D-3L and EQ-5D-5L value sets, where the EQ-5D-5L values for matched states are higher, and the overall range and therefore change between adjacent states is smaller than for the EQ-5D-3L. The distribution of values also differs. There are similar differences between the EQ-5D-5L value set and the crosswalk tariff given that the latter is linked to the EQ-5D-3L value set. However it is also worth noting that some underlying features of the preferences, and therefore utility scales, are similar. For example, the overall importance of each dimension to the overall value is similar, with only one difference (where the rank order of the dimensions is the same, apart from two dimensions, mobility and anxiety/depression, changing position in the ordering in the EQ-5D-5L value set), and the relative distance between the levels for different dimensions is similar.

Regarding the observed values from the patient data, the EQ-5D-5L value set produces higher values overall and across all of the conditions included, and the differences are generally significant. This is expected given the overall increase in the values of matched states and reduction in the overall utility scale. There is some evidence that the value sets rank different health conditions in a similar order,

particularly the most and least severe conditions as measured by the descriptive systems. However this requires further exploration across a larger range of conditions.

There are a number of possible reasons why the EQ-5D-3L and EQ-5D-5L value sets differ. These include differences in the samples used in terms of demographics and country. The EQ-5D-3L value set was based on a representative sample of England, Scotland and Wales, whereas the EQ_5D-5L was based on just an English sample. This may have implications for decision making in the jurisdictions that are not represented. However, the project team has since collected EQ-5D-5L valuation data for the other countries in the UK so will be able to compare using a more representative sample (albeit one that is smaller than that used for the EQ-5D-3L). Potential changes in population demographics and preferences over time (from 1993 to 2013), is another possible reason why the value sets demonstrate differences. For example the population is getting older [24], and this might impact on preferences for different health dimensions. One indication of change in preferences over time might be the increased magnitude of the anxiety/depression dimension given increased focus on the detrimental aspects of mental health conditions in policy [25], and reduction in stigma surrounding conditions such as depression [26]. Even without the development of the EQ-5D-5L, the currently used EQ-5D-3L value set is outdated and therefore would require updating anyway. Overall the dimension preference structure between the EQ-5D-3L and EQ-5D-5L is similar, with only one inversion (anxiety/depression and mobility) which is encouraging given the differences between the studies. This may demonstrate that the order of preferences for the five areas of health described by the EQ-5D may be generally consistent over time.

Other reasons why the value sets may differ relate to the descriptive system and the valuation method used. Firstly regarding the descriptive system, the EQ-5D-5L uses more consistent wording, particularly for the more severe levels, and it is possible that the change in labelling of the mobility dimension (from 'confined to bed' to 'unable to walk about') has impacted the values, where mobility has a smaller weighting in the EQ-5D-5L than in the EQ-5D-3L. The increase in levels and associated sensitivity also may impact the magnitude of the difference and transition between the intermediate levels and therefore the overall value set.

Secondly, the valuation method differs, particularly regarding the process used to value states worse than dead which was problematic for the EQ-5D-3L [11]. The methodological change to a new approach to eliciting values < 0, the lead time TTO, meant that the lowest possible value for an EQ-5D-5L health state in the protocol used was -1 [15,27]. In contrast the minimum value was -39 in the Dolan study [3], which was rescaled to -1. This therefore led to a reduction in the overall scale. The inclusion of DCE tasks in the EQ-5D-5L valuation also provides a different type of valuation data focusing on the choices between states rather than measuring direct values for states as is the case with TTO. The development of innovative modelling methods combining TTO and DCE data in one model [28,29] provide further reasons for differences in the value sets. The modelling process for the EQ-5D-5L data also developed heterogeneous models for the TTO data only [19], and further work is underway to

model the EQ-5D-3L valuation data applying the methods developed for the EQ-5D-5L [30]. It is also worth noting that a partial replication of the original EQ-5D-3L valuation study was carried out by Macran and Kind [31]. In this study the authors used a smaller health state design but a similar TTO process to Dolan [3] and estimated an EQ-5D-3L value set with quite different characteristics. For example, the value for the worst state was substantially higher (-0.126 vs. -0.594), and the amount of negative states was substantially lower (12.3% vs 34.6%). This value set is more in line with other EQ-5D-3L value sets developed internationally [32], and provides a useful counterpoint for comparisons between the value sets included in this study.

There are also large differences in the proportion of states valued as worse than dead (i.e. with a negative value) and the associated values assigned to these states, which has resulted in a smaller range for the EQ-5D-5L. One of the key criticisms of the EQ-5D-3L value set was the process used to value and subsequently model states worse than dead which led to the large range observed [11] which may not realistically reflect population preferences. The protocol for the development of the EQ-5D-5L value set introduced a new method for the valuation of states worse than dead which bounded all observed values on a -1 to 1 scale [15,17]. This has reduced the overall proportion of negative values, and moved the anchor value of 0 (i.e. the state equivalent to dead). Further work could compare the characteristics of the health states that have a values close to zero across different value sets.
However the impact of the change in negative values on HTA is unclear, as it is not well established how often states that are worse than dead actually appear in cost effectiveness models. There are differences in the proportions of negative states in different conditions, where the proportion is similar across the EQ-5D-3L and EQ-5D-5L in Parkinson's disease, but quite different for MS and COPD, for example. This might be due to changes in the magnitude of the decrement associated with the key dimensions for each condition. As the overall range of EQ-5D-5L values is smaller, the change in QALYs (for estimates generated from quality of life changes) might be reduced across the whole scale for states both better and worse than dead. This depends on the descriptive data, where respondents could show no change on the 3L (i.e. 'some' problems both before and after) whilst showing a change on the 5L (move from "moderate" to "slight") leading to higher QALY gains.

It is also useful to compare the scale of the English EQ-5D-5L value set with those from other countries that were developed using the same valuation protocol [15]. For example, the Dutch value set has a minimum value of -0.446 with around 15% of states valued negatively [33]. The Spanish EQ-5D-5L value set has a minimum value of -0.224 [34]. Differences between countries could be due to cultural differences in preferences as well as the use of different modelling approaches. Further work should compare EQ-5D-5L value sets from different countries in more detail.
It is unclear how the differences between the value sets indicated in both analysis of the estimates and patient data will impact the HTA process. This is because the utility values will be applied to both treatments and their comparators, and therefore to some extent the differences may be even, and the estimates of improvements in quality of life between arms of a clinical trial could be similar using the EQ-5D-3L or EQ-5D-5L value sets. The increased sensitivity of the EQ-5D-5L in terms of the addition

of two extra response levels, and the change possible across the levels may also favour QALY gains even if the changes in utility are smaller. An added complexity is whether the gain is linked to improving quality of life or extending length of life, and the interaction between the two. This requires further investigation on clinical trial data, which is a key part of this programme of research, and has also been investigated by other researchers who found different cost effectiveness estimates based on the value set used [20].

There are also implications for the NICE reference case, and further decision making based on their recently released position statement regarding the use of the EQ-5D-5L. The improvement in the methods used to both collect and model the valuation data, and the increased use of the improved descriptive system, make a strong case for the use of the new EQ-5D-5L value set. The EQ-5D-3L value set has benefits if the instrument is still being used in trials and other settings, but is based on societal preferences from decades ago. The crosswalk draws on the EQ-5D-3L values so is prone to the same issues as that value set. There is also the potential for 'gaming' where the crosswalk may be used instead of the EQ-5D-5L value set to potentially inflate QALY gains (as the utility range, and therefore change between states, is larger). One important point is how to compare results of cost utility analyses using the EQ-5D-5L against those using the EQ-5D-3L and establishing the cost per QALY thresholds that should be used. Further work is required to explore this.

The main limitation of this study is that we have not tested the impact of the value sets on any clinical trial data which would have enabled us to directly compare QALY estimations. This would allow us to test some of the issues raised in data previously used for cost utility analysis, and is the next planned stage of this programme of research. It will also be important to compare the psychometric performance, and impact on cost utility analysis, of the EQ-5D-5L descriptive system and value set with those of other widely used generic measures. In particular comparisons with version two of the SF-6D (SF-6Dv2) [35] which has been valued using DCE with duration methods would be useful.

5. Conclusions

In conclusion we have demonstrated key differences in the theoretical and observed values from three EQ-5D value sets that can be used in HTA. The value sets will lead to differences, and the use of the EQ-5D-5L value set will have implications for the decision making process carried out by NICE and may require revision to the guidelines used for the economic evaluation of health technologies.

6. Data availability statement

The crosswalk data used in this study were funded and are owned by the EuroQol Research Foundation. The data are not publicly available as analysis is ongoing, but requests for data sharing can be made to the EuroQol Research Foundation

7. Compliance with ethical standards

7.1. Conflict of interest:

All authors (BM, KS, YF, BvH, MFJ, MH and ND) are members of the EuroQol Research Foundation (the copyright holders of the EQ-5D-3L and EQ-5D-5L). No funding was received for this study, but the crosswalk study data collection was funded by the EuroQol Research Foundation.

7.2. Author contributions:

BM lead the data analysis and interpretation and was primarily responsible for drafting the manuscript. YF, KS, MH and ND supported the data analysis and interpretation, and commented on and amended the draft manuscript. MFJ and BvH were involved in the collection of the data underpinning the study, the interpretation of the data analysis, and revising the manuscript. BM acts as overall guarantor of the work.

8. References

1 Brooks R. EuroQol: The current state of play. Health Policy. 1996;37:53-72.

2 Devlin N, Brooks R. EQ-5D past, present and future. Applied Health Econ Health Policy. 2017;15(2):127-37.

3 Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35(11):1095-108.

4 National Institute of Health and Care Excellence. Guide to the methods of technology appraisal. London: NICE; 2013.

5 Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. Canberra: Australian Department of Health; 2015.

6 Canadian Agency for Drugs and Technology in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada; 2017.

7 Appleby J, Devlin N, Parkin D. Using Patient Reported Outcomes to Improve Health Care. London: Wiley-Blackwell; 2015.

8 Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, Lloyd Jones M, Paisley S, O'Cathain A, Barkham M, Knapp M, Byford S, Gilbody S, Parry G. A systematic review, psychometric analysis and qualitative assessment of Generic Preference-Based Measures of Health in Mental Health Populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess. 2014;18:34.

9 Longworth L, Yang Y, Young T, Mulhern B, Hernandez-Alava M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P. 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.

10 Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ. 2005;13(9):873-84.

11 Lamers LM. The transformation of utilities for health states worse than death: consequences for the estimation of EQ-5D value sets. Med Care. 2007;45(3):238-44.
12 Herdman, M., Gudex, C., Lloyd, A., Janssen, M.F., 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.
13 Janssen MF, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P, Busschbach J. 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.
14 van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, Lloyd A, Scalone L, Kind P, Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value in Health. 2012;15(5):708-15.
15 Oppe, M., Devlin, N.J., van Hout, B., Krabbe P.F.M., de Charro, F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17:445-53.
16 Janssen MF, Oppe M, Versteegh M, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013;14(1):5-13.
17 Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goñi, J.M., Luo, N., 2016. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics, 34, pp.993-1004.
18 Devlin N, Shah K, Feng Y, Mulhern BJ, van Hout B. Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. Health Econ. 2017;doi:10.1002/hec.3564.
19 Feng Y, Devlin N, Shah K, Mulhern BJ, van Hout B. New Methods for Modelling EQ-5D-5L Value Sets: An Application to English Data. Health Econ. 2017;doi:10.1002/hec.3560
20 Hernandez Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, Meads D, O'Dwyer J, Barton G, Irvine L. EQ-5D-5L versus 3L: The impact on cost-effectiveness in the United Kingdom. Value Health. 2017;doi:10.1016/j.jval.2017.09.004.
21 National Institute of Health and Care Excellence. EQ-5D-5L: NICE Position Statement. NICE. 2017. https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf).
22 Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–310.
23 Parkin D, Devlin N, Feng Y. What determines the shape of an EQ-Index distribution? Medical Decis Mak. 2016; doi:10.1177/0272989X16645581.
24 Office of National Statistics. Census 2011. London: Office of National Statistics. 2011.
25 Layard R. A New Priority for Mental Health. London: London School of Economics; 2015.
26 Rüsch N, Angermeyer M, Corrigan P. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. European Psychiatry. 2005;20(8):529–39.
27 Devlin N, Tsuchiya A, Buckingham K, Tilling C. A uniform Time Trade Off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Econ. 2011;20(3):348-61.
28 Ramos-Goñi J.M, Pinto-Prades J.L, Cabasés J.M, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care. 2014;55(7):e51-e58.

29 Rowen D, Brazier J, van Hout B. A Comparison of Methods for Converting DCE Values onto the Full Health-Dead QALY Scale. Med Decis Mak. 2014;35:328-40.

30 Feng Y, Devlin N, van Hout B. Revisiting the MVH value set: applying 5L modelling methods to 3L valuation data., Barcelona: EuroQol Research Foundation Plenary; 2017.

31 Macran S, Kind P. Valuing EQ-5D health states using a modified MVH protocol: Preliminary results. Sitges: EuroQol Research Foundation Plenary; 1999.

32 Szende A, Oppe M, Devlin N. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Rotterdam: Springer; 2007.

33 Versteegh M, Vermeulen K, Evers S, de Wit GA, Prenger R, Stolk E. Dutch Tariff for the Five-Level Version of EQ-5D. Value Health. 2016;19(4):343-52.

34 Ramos-Goñi JM1, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and Modeling of EQ-5D-5L Health States Using a Hybrid Approach. Med Care. 2017;55(7):e51-e58.

35 Mulhern B, Brazier J. Developing SF-6D-V2: The classification system. Qual Life Res. 2014;23:49.

Table 1: Demographic characteristics of the crosswalk data used for the comparative analysis

Demographic	N(%)
N	1,501
Country	
Scotland	500
England	1,001
Age	
Mean (sd)	57 (16)
Range	19 - 94
Gender male	734 (49)
Education	
Left school with no qualifications	485 (32)
Left school with some qualifications	339 (23)
College degree/further education	377 (25)
Degree/postgraduate/professional	300 (20)
EQ-5D visual analogue scale	
Mean (SD)	60.3 (21.4)
Range	0 to 100
Condition	
COPD	320 (21)
Heart problems	251 (17)
Arthritis	250 (17)
Depression	250 (17)
Rheumatoid arthritis	87 (6)
Stroke	85 (6)
Back pain	70 (5)
ADHD	69 (5)
Diabetes	45 (3)
Parkinson's	37 (3)
Breathing problems	22 (2)
Multiple sclerosis	15 (1)

Table 2: Comparing the EQ-5D-3L and EQ-5D-5L models

Parameters		EQ-5D-3L	EQ-5D-5L[2]	Value calculation (21223/31335)	
				EQ-5D-3L	EQ-5D-5L
Constant		0.081			
EQ-5D dimensions					
Mobility	None	0	0		
	Slight		0.058		
	Some/Moderate	0.069	0.076	0.069	0.076
	Severe		0.207		
	CTB[1]/Unable to	0.314	0.274		
Self-care	None	0	0	0	0
	Slight		0.050		
	Some/Moderate	0.104	0.080		
	Severe		0.164		
	Unable to	0.214	0.203		
Usual Activities	None	0	0		
	Slight		0.050		
	Some/Moderate	0.036	0.063	0.036	0.063
	Severe		0.162		
	Unable to	0.094	0.184		
Pain/discomfort	None	0	0		
	Slight		0.063		
	Some/Moderate	0.123	0.084	0.123	0.084
	Severe		0.276		
	Extreme	0.386	0.335		
Anxiety/depression	None	0	0		
	Slight		0.078		
	Some/Moderate	0.071	0.104		
	Severe		0.285		
	Extreme	0.236	0.289	0.236	0.289
Interactions					
	N3 term	0.269		0.269	
Value of state				1-0.081-0.069-0-0.036-0.123-0.236-0.269=0.186	1- (0.076+0+0.063+0.084+0.289) = 0.488

[1]CTB: Confined to bed; [2] Mean coefficient from the Bayesian regression with the latent class adjustment applied

Table 3: Overall descriptive characteristics of the three value sets (modelled values)

	EQ-5D-3L value set	EQ-5D-5L crosswalk	EQ-5D-5L value set
Range	1 to -0.594	1 to -0.594	1 to -0.285
% health states worse than dead	34.6%	26.7%	5.1%
Dimension importance order §	Pain/Discomfort Mobility Anxiety/depression Self-care Usual Activities	Pain/Discomfort Mobility Anxiety/Depression Self-care Usual Activities	Pain/Discomfort Anxiety/depression Mobility Self-care Usual Activities
Health state values			
'Mildest' state (11211)*	0.883	0.906	0.950 (11211/12111)
'Moderate' state (22222 (3L) or 33333 (5L))	0.516	0.516	0.593
'Worst' state (33333 (3L) or 55555 (5L))	-0.594	-0.594	-0.285

§Importance is judged by the size of the coefficient for level 5 in each dimension.

Table 4: Comparing the change in utility between adjacent health states

	EQ-5D-5L value set			Crosswalk value set			EQ-5D-3L value set		
EQ-5D-5L state	Value	Difference	Difference matched[a]	Value	Difference	Difference matched	EQ-5D-3L state	Value	Difference
11111	1.000			1.000			**11111**	1.000	
21111	0.942	0.058		0.877	0.123				
31111	0.924	0.018	0.076	0.850	0.027	0.150	**21111**	0.850	0.150
41111	0.793	0.131		0.813	0.037				
51111	0.726	0.067	0.198	0.336	0.477	0.514	**31111**	0.336	0.514
11111	1.000			1.000			**11111**	1.000	
12111	0.950	0.050		0.846	0.154				
13111	0.920	0.030	0.080	0.815	0.031	0.185	**12111**	0.815	0.185
14111	0.836	0.084		0.723	0.092				
15111	0.797	0.039	0.123	0.436	0.287	0.379	**13111**	0.436	0.379
11111	1.000			1.000			**11111**	1.000	
11211	0.950	0.050		0.906	0.094				
11311	0.937	0.013	0.063	0.883	0.023	0.117	**11211**	0.883	0.117
11411	0.838	0.099		0.776	0.107				
11511	0.816	0.022	0.121	0.556	0.220	0.327	**11311**	0.556	0.327
11111	1.000			1.000			**11111**	1.000	
11121	0.937	0.063		0.837	0.163				
11131	0.916	0.021	0.084	0.796	0.041	0.204	**11121**	0.796	0.204
11141	0.724	0.192		0.584	0.212				
11151	0.665	0.059	0.251	0.264	0.320	0.532	**11131**	0.264	0.532
11111	1.000			1.000			**11111**	1.000	
11112	0.922	0.078		0.879	0.121				
11113	0.896	0.026	0.104	0.848	0.031	0.152	**11112**	0.848	0.152
11114	0.715	0.181		0.635	0.213				
11115	0.711	0.004	0.185	0.414	0.221	0.434	**11113**	0.414	0.434

a The 'difference matched' calculation refers to the difference between states that are matched across the EQ-5D-3L and EQ-5D-5L (e.g. 31111 on the EQ-5D-5L is equivalent to 21111 on the EQ-5D-3L)

Table 5: Dimension level responses across the EQ-5D-3L/EQ-5D-5L (English and Scottish data)

Dimension responses	EQ-5D-3L (n,%)	EQ-5D-5L (n,%)
Mobility		
None	506 (33.7)	435 (29.0)
Slight		392 (26.1)
Some/Moderate	983 (65.5)	377 (25.1)
Severe		277 (18.5)
CTB[a]/Unable to	12 (0.8)	20 (1.3)
Self-care		
None	951 (63.4)	907 (60.4)
Slight		301 (20.1)
Some/Moderate	517 (34.4)	201 (13.4)
Severe		74 (4.9)
Unable to	33 (2.2)	18 (1.2)
Usual Activities		
None	464 (30.9)	390 (26.0)
Slight		447 (29.8)
Some/Moderate	881 (58.7)	358 (23.9)
Severe		228 (15.2)
Unable to	156 (10.4)	78 (5.2)
Pain/discomfort		
None	380 (25.3)	303 (20.2)
Slight		447 (29.8)
Some/Moderate	947 (63.1)	449 (29.9)
Severe		243 (16.2)
Extreme	174 (11.6)	59 (3.9)
Anxiety/depression		
None	672 (44.8)	571 (38.0)
Slight		444 (29.6)
Some/Moderate	721 (48.0)	324 (21.6)
Severe		111 (7.4)
Extreme	108 (7.2)	51 (3.4)

a CTB: Confined to Bed (Level 3 of the EQ-5D-3L Mobility dimension)

Figure 1: All unique theoretical values

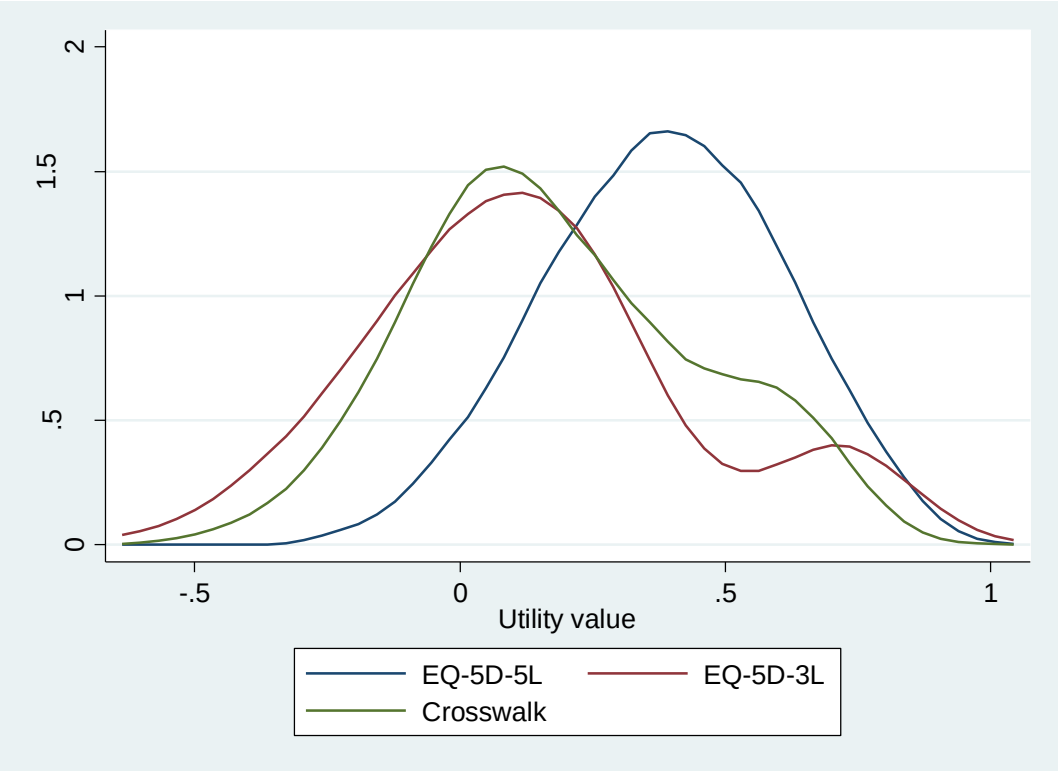


Figure 2: Values of comparable states ordered by EQ-5D-5L value

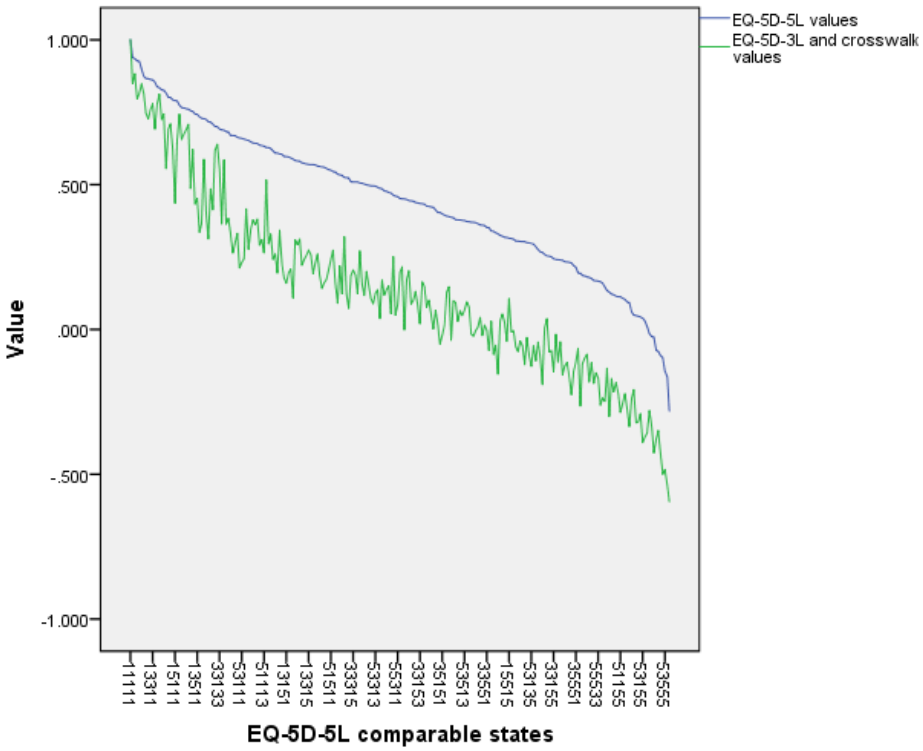
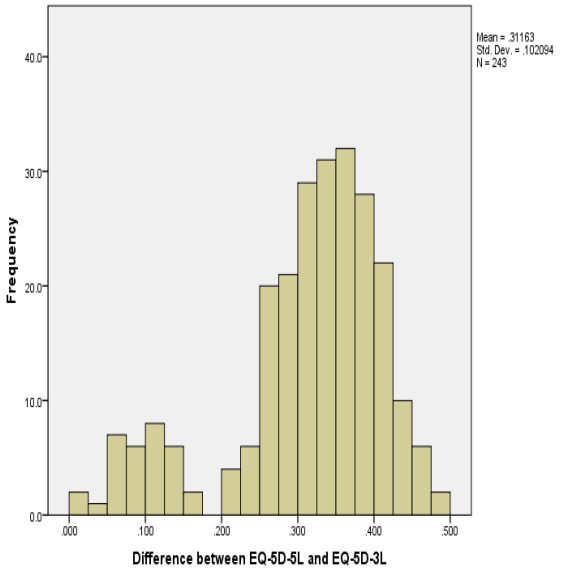


Figure 3: Histogram and boxplot of differences between the EQ-5D-3L and EQ-5D-5L value sets



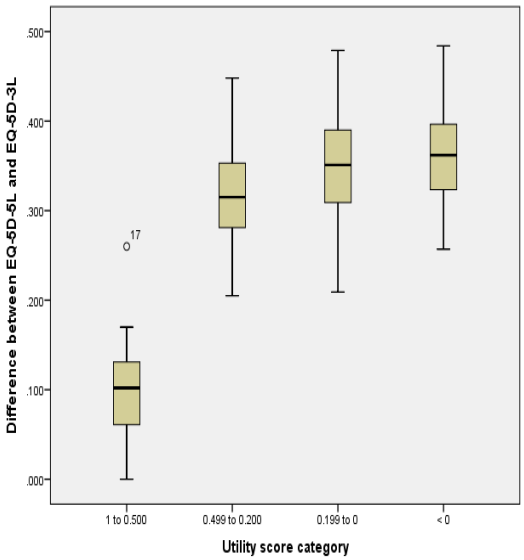


Figure 4: Comparison of all EQ-5D-3L ,EQ-5D-5L and EQ-5D-5L crosswalk values

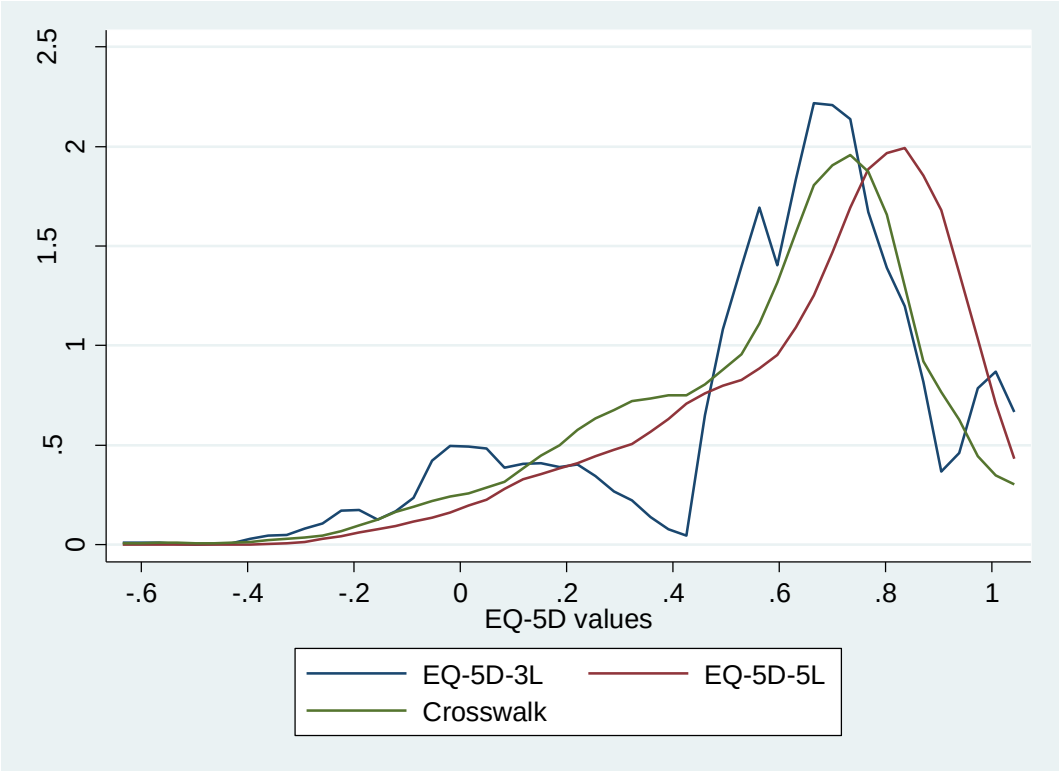


Figure 5: EQ-5D-3L and crosswalk patient values ordered by EQ-5D-5L

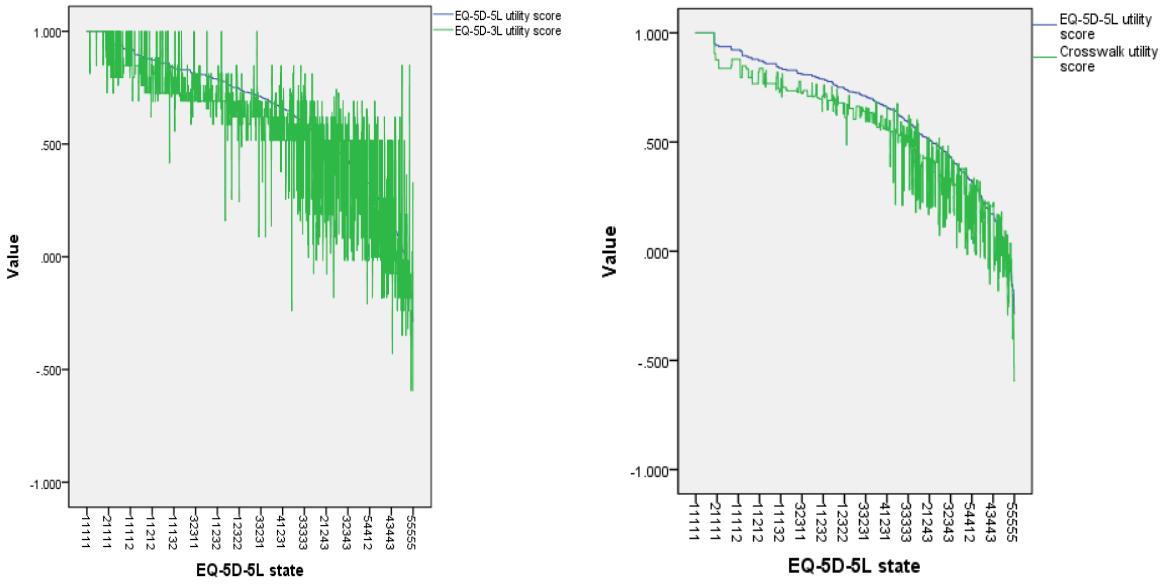


Figure 6: Bland Altman plots comparing EQ-5D-3L and EQ-5D-5L, and EQ-5D-5L and crosswalk scores