**Comparison of the EQ-5D-5L and the EQ-5D-3L using individual patient data from REFORM trial**

Belen Corbacho1§, Ada Keding1, Ling-Hsiang Chuang2, Juan Manuel Ramos-Goñi3, Kalpita Joshi1, Sarah Cockayne1, David J Torgerson1

**Department(s) and institution(s) to which the work should be attributed**

1 York Trials Unit, Department of Health Sciences, University of York, York, United Kingdom

2 Department of Epidemiology and Global Health, Umeå University, Sweden

3 Maths in Health B.V., Rotterdam, The Netherlands

**ORCID:**

Belen Corbacho: orcid.org/0000-0002-2359-0379

Ada Keding: orcid.org/0000-0002-1182-887X

Ling-Hsiang Chuang:

Juan Manuel Ramos-Goñi:

Kalpita Joshi: orcid.org/0000-0001-8472-5698

Sarah Caockayne: orcid.org/0000-0002-1288-5497

David J Torgerson: orcid.org/0000-0002-1667-4275

**Details of corresponding author**

Belen Corbacho

ARRC Building, Department of Health Sciences, University of York, York, YO10 5DD, UK

Email: [belen.corbacho@york.ac.uk](mailto:belen.corbacho@york.ac.uk)

Tel: +44 (0) 1904 321852 / Fax: +44 (0) 1904 321387

**Comparison of the EQ-5D-5L and the EQ-5D-3L using individual patient data from the REFORM trial**

**Abstract**

**Background:** This study compares the 5-level version of the EQ-5D (5L) with the 3-level version EQ-5D (3L) in older adults using individual patient data from the REFORM (REducing Falls with Orthoses and a Multifaceted podiatry intervention) trial.

**Methods:** EQ-5D-5L and EQ-5D-3L were administered to men and women (n=151) over the age of 65 years alongside the REFORM trial.The two versions of the EQ-5D were assessed in terms of feasibility, level of consistency, ceiling effect and discriminatory power. We also undertook a comparison of the performance of different EQ-5D-3L and EQ-5D-5L value sets.

**Results:** The proportion of participants that returned a complete questionnaire was higher for the 5L (96.7%) than for the 3L (92.7%). Missing values among dimensions were on average 1.59% (5L) and 1.45% (3L). The ceiling effect was reduced from 18.2% (3L) to 6% (5L). On average the proportion of inconsistent responses between both descriptive systems was 3.25%. Redistribution from 3L to 5L showed valid results for the majority of consistent level combinations, with slight inconsistency in the case of Anxiety/Depression. For the 5L, 67 unique health states were observed for the 5L compared to 27 for the 3L. The absolute informatively improved with the new classification system (5.48 for 5L versus 3.91 for 3L) and relative discriminatory power improved slightly on average (0.90 for 5L versus 0.84 for 3L). The mean difference between the EQ-5D-5L and EQ-5D-3L values was 0.091 (range -0.345 to 0.505); whilst the mean difference between the EQ-5D-5L and the crosswalk values was 0.082 (range -0.035 to 0.293).

**Conclusion:** In the REFORM clinical trial involving an elderly population, our study supported the feasibility and convergent validity of both EQ-5D-3L and EQ-5D-5L. Results suggest that the 5L improves the ceiling effect and discriminatory power. The EQ-5D-5L scores were significantly higher than both EQ-5D-3L and crosswalk.

**Keywords:** Health related Quality of life, EQ-5D-3L, EQ-5D-5L, elderly population

**Introduction**

The National Institute for Health and Care Excellence (NICE) develop evidence-based guidelines on the most effective ways to diagnose, treat and prevent disease and ill health. Part of the evaluation includes a health economic component. Typically, these evaluations use a cost-utility analysis, where health gains are normally expressed in terms of quality-adjusted life years (QALYs), and decisions about whether a treatment is efficient or not are determined by whether the cost per QALY is below a certain threshold. The QALY considers both the quantity and quality of life generated by healthcare interventions. QALYs are estimated following a three-step approach. The first step requires the need for assessing heath related quality of life experienced by patients from generic instruments such as the EQ-5D. The value tariff of patients’ EQ-5D scores are then converted into a utility index score on a scale from 0 (dead) to 1 (perfect health), with negative values for health states considered worse than death. Finally, the EQ-5D index score is used as the quality adjusted component in the calculation of QALYs, which are estimated by multiplying the time spent in each health state over time with its corresponding utility value.

The EQ-5D™ is widely used as a measure of health in economic evaluations and is designed for self-completion by the respondent. NICE [1] and other reimbursement agencies [2, 3] recommend the use of the EQ-5D, as it is a generic quality of life instrument that can be applied to a wide range of health conditions. Therefore, the EQ-5D is commonly included in trials such as those funded by the National Institute for Health Research (NIHR), such as the REFORM (REducing Falls with Orthoses and a Multifaceted podiatry intervention) trial ([ISRCTN68240461](https://doi.org/10.1186/ISRCTN68240461)).

There are currently two versions of the EQ-5D that researchers can use for adults; the original EQ-5D-3L (five dimensions of health with three levels of problems); and a new EQ-5D-5L more complex version (same five dimensions of health with increased five levels of problems). The EQ-5D-3L has an associated utility value set based on estimates of the preferences of the UK general population [4]. There is also a value set for the EQ-5D-5L available for England [[5](#_ENREF_1)]; in addition, utilities for the 5L can be derived using the crosswalk by van Hout *et al* [[6](#_ENREF_1)].

Following on the publication of the English value set, NICE, in collaboration with the Department of Health and Social Care, commissioned a review to evaluate the quality of the data and the modelling techniques used to derive the EQ-5D-5L valuation set for England. While the EQ-5D-5L valuation set is under review, NICE supports funders to use the 5L version of the EQ-5D to collect data on health related quality of life in randomised controlled trials (RCTs), and recommends that utility values should be calculated using the crosswalk developed by van Hout *et al* (2012) [[6](#_ENREF_1)].

Although the EQ-5D-5L implies an improvement of the descriptive system [7], it remains important to explore the use of the new EQ-5D-5L in clinical trials and its potential to improve the sensitivity of the original 3L and reduce ceiling effects. The aim of our study is to compare the use of the EQ-5D-5L to the EQ-5D-3L in the context of the REFORM trial. We compared both versions of the EQ-5D instrument in terms of their feasibility, level of consistency, ceiling effect and discriminatory power. We also investigated the differences in the utility values generated by both valuation systems for the participants in the trial.

**Methods**

***REFORM study design and participants***

The REFORM trial was a pragmatic multicentre cohort RCT in England and the Republic of Ireland. The design involved the recruitment of an observational cohort from which eligible, consenting participants (≥65 years old) were randomised into the trial to receive either a podiatry intervention (n = 493), including foot and ankle strengthening exercises, foot orthoses, new footwear if required, and a falls prevention leaflet, or usual podiatry treatment plus a falls prevention leaflet (n = 517). Recruitment took place through 37 NHS podiatry clinics in primary or secondary care. Participants were ineligible if they were <65 years of age; reported neuropathy, dementia, or another neurological condition; were unable to walk household distances; had a lower limb amputation; or were unwilling to attend their podiatry clinic for a REFORM appointment. Participants in the cohort were eligible for inclusion in the trial if they had had a fall in the past 12 months, or a fall in the past 24 months requiring hospital attention or reported worrying about falling. The primary outcome was the incidence rate of self-reported falls per participant in the 12 months following randomisation. Secondary outcomes included: proportion of fallers and those reporting multiple falls, time to first fall, fear of falling, Frenchay Activities Index, Geriatric Depression Scale, foot pain, health related quality of life, and cost-effectiveness. The study (and this sub study) was approved by the East of England National Research Ethics Committee. The REFORM protocol [8], and trial results in terms of effectiveness and cost-effectiveness results have been published elsewhere [9, 10].

The participants were asked to complete a consent form to indicate they wished to take part in the study; they were also informed about the opportunity to take part in this EQ-5D sub-study. The EQ-5D data used for this analysis were collected from a sample of participants in the REFORM trial (n = 151); hence the dataset was a pre intervention dataset. Each participant was asked to complete a baseline questionnaire which included both the 3L and 5L versions of the UK English language EQ-5D. Participants completed in order: (1) EQ-5D-5L questions and associated visual analogue scale (EQ-VAS); and then (2) EQ-5D-3L and associated EQ-VAS. The 5L version was administered first to avoid participants not using the ‘in-between’ levels 2 and 4 of the EQ-5D-5L [11].

***Instruments***

*EuroQol*

The EQ-5D comprises of two parts: classification of five dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and a visual analogue scale (VAS), which records participant’s overall evaluation of their health on a scale from 100 (best imaginable health) to 0 (worst imaginable health). The 3L classification system considers three severity levels (1=no problems, 2=some/moderate problems, 3=severe problems) which defines 243 health states ranging from 11111 (full health) to 33333 (worst health). In the new EQ-5D-5L version the level of each of the five domains has been increased from 3 levels to 5 levels (1=no problems, 2=slight problems, 3= moderate problems, 4=severe problems and 5=unable/extreme problems) and also standardised wording across dimensions. In addition, in the mobility dimension the response option “confined to bed” has been replaced with “unable to walk” as the most severe level. Therefore, the number of possible health states has increased for the 5L to 3,125 ranging from 11111 (full health) to 55555 (worst health).

*Other self-reported health outcomes*

The self-reported presence of ten long-term health conditions (Diabetes, Arthritis, Osteoporosis, ALS/Lou Gehrig’s disease, Multiple Sclerosis, Parkinson’s disease, Huntingdon’s disease, Alzheimer’s disease, Depression and Dementia) was collected at baseline for REFORM participants. Besides, participants also completed other related health instruments, including the Frenchay Activities Index (FAI) [12] and the Geriatric Depression Scale (GDS) [13]. The FAI is a 15-item instrument to assess a broad range of activities of daily living by recording the frequency with which each item or activity was undertaken over the previous three or six months. The scale provides a summed total score from 15 to 60 (higher scores represent higher activity levels). The GDS is a 15-item scale to assess geriatric depression, with each item representing a negative state of mind, with which respondents either agree or disagree. The GDS score ranges from 0 to 15 (higher scores represent greater levels of depression), with a scores greater than five considered the threshold for mild depression and scores greater than 10 considered the threshold for severe depression.

***Data analysis***

*Feasibility*

The feasibility was explored in terms of missing responses. For each EQ-5D version, we assessed the proportion of participants with (i) completely missing responses (all five dimensions missing); (ii) partially missing responses (at least one domain completed); and (iii) missing responses by domain. We also computed the proportion of participants not completing the VAS.

*Redistribution analysis*

The distribution of the 3L and 5L responses was shown in terms of percentage of each level reported. The redistribution patterns of the responses from the 3L to 5L for each dimension were also reported in terms of

percentage. An inconsistent 3L response by the same REFORM participant was defined as being at least two levels away from the 5L response given. The mapping between 3L and 5L responses was recorded, and the mean size of inconsistencies calculated following the redistribution diagram and inconsistency weights proposed by Janssen *et al* [14] (Table 1). We analysed the redistribution properties of the consistent response pairs as the proportion of the 3L-5L response pairs within each 3L response level (e.g. 1, 2 and 3) and the corresponding mean and median VAS values.

**Table 1. Redistribution diagram and inconsistency weights for the relationship between the 3-5L, described by Jassen *et al* [14]**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **EQ-5D Response Level (5L)** | | | | |
|  |  | **1** | **2** | **3** | **4** | **5** |
| **EQ-5D Response Level (3L)** | **1** | *Consistent* | *Consistent* | Inconsistent (weight 1) | Inconsistent (weight 2) | Inconsistent (weight 3) |
| **2** | Inconsistent (weight 1) | *Consistent* | *Consistent* | *Consistent* | Inconsistent (weight 1) |
| **3** | Inconsistent (weight 3) | Inconsistent (weight 2) | Inconsistent (weight 1) | *Consistent* | *Consistent* |

*Ceiling effect*

Assuming the elderly REFORM participants had at least slight or some problems in one or more dimensions of health, we compared ceiling effects between both systems by estimating the proportion of participants responding with no problems in each dimension and also the proportion of participants reporting problems in every domain (i.e. 11111). We would expect the 5L to have a lower ceiling effect compared with the 3L. We estimated absolute ceiling effect (i.e. difference between the proportions) and the relative change [(ceiling3L – ceiling5L/ceiling3L) x100].

*Discriminatory power*

A key feature of any health status measure is the ability to discriminate among people at a single point in time [15]. Discriminatory power was assessed using the Shannon indices. Unlike floor and ceiling effects, the Shannon indices incorporate the frequency distribution across all categories of a health status classification system.[15].

The Shannon index is defined as:

Where represents the absolute amount of informativity captured, is the number of possible categories (levels or permutations in this study), and = , the proportion of observations in the ith category (i = 1,…, ), where is the observed number of scores (responses) in category i and is the total sample size. The higher the value of , the more information that is being captured by the system [15].

In the case of a homogeneous distribution, where responses are evenly distributed among categories, an optimal amount of information is captured and reaches its maximum (). This equals . If the number of categories () is increased, will increase, however will only increase if the additional categories are utilised [15].

Shannon’s Evenness Index () is defined as:

Shannon’s index can be considered to represent the absolute informativity of a system whereas Shannon’s Evenness index expresses the relative informativity, regardless of the number of categories [15].

Discriminatory power for the EQ-5D-3L and EQ-5D-5L was assessed by dimension and by instrument as a whole. To calculate Shannon’s indices by dimension, levels are treated as categories, so here would be the number of levels per instrument and is the proportion of responses of the th level. However, to calculate Shannon’s indices by instrument, permutations are treated as unique categories, where is the number of permutations (, is the proportion of the ith permutation and now equals [15]. For the EQ-5D-3L there are 243 possible permutations, or unique health states, and 3125 for the EQ-5D-5L.

As the number of participants in this study is N = 151, the maximum number of unique health states (243, 3125) could never have been reached a priori, and consequently maximum informativity ( ) and maximum relative informativity () would never have been reached. Using methods similar to Polinder *et al.* (2010) the observed number of unique health states in the population are used, not the theoretical possible number of health states [16].

The data shows 25 and 67 unique health states for the EQ-5D-3L and EQ-5D-5L respectively. This gives:

(EQ-5D-3L)

(EQ-5D-5L)

*Association between multimorbidity and EQ5D index score*

Utility values for the 3L were derived using the UK EQ-5D-3L value set [4]. Utility values for the 5L were derived using both the crosswalk by van Hout *et al.* (2012) [6] and the English EQ-5D-5L value set [5]. We compared the value sets scores overall and across the different health conditions.

**Results**

***Characteristics of REFORM sub-sample and EQ-5D profile***

The sub-sample comprised a total of 151 participants. Characteristics of respondents were comparable in both samples. The mean age of participants for the sub-sample was 78.3 (SD 6.25) years and 56.2% were women, compared to 78.1 years (SD 7.2) and 61.0% female for the main trial. For the sub-sample, 37% had diabetes, 7% were affected by depression, 17% had osteoporosis and 58% suffered from arthritis; compared to 33% diabetes, 9.6% depression, 15% osteoporosis and 58.6% arthritis for the main trial. Table 2 displays the dimension level responses of REFORM participants to both descriptive systems. The proportion of REFORM participants reporting slight problems was higher for each dimension of the 3L compared to the 5L. For instance, the proportion of participants reporting slight problems with mobility was 61% for 3L compared to 30% for 5L; and similarly, for pain/discomfort (74% for 3L vs 44% for 5L).

**Table 2. Dimension level responses across the EQ-5D-3L and EQ-5D-5L**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mobility** | | | | **Self-Care** | | | | **Usual Activities** | | | | **Pain/Discomfort** | | | | **Anxiety/Depression** | | | |
|  | **3L^** | | **5L~** | | **3L** | | **5L** | | **3L** | | **5L** | | **3L** | | **5L** | | **3L** | | **5L** | |
| **Level** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** | **N** | **%** |
| **1** | 58 | 39 | 48 | 33 | 120 | 81 | 113 | 76 | 78 | 52 | 62 | 42 | 28 | 18 | 19 | 13 | 106 | 71 | 96 | 64 |
| **2** | 91 | 61 | 44 | 30 | 28 | 19 | 23 | 15 | 65 | 44 | 46 | 31 | 110 | 74 | 66 | 44 | 42 | 28 | 40 | 27 |
| **3** | 0 | 0 | 37 | 25 | 0 | 0 | 12 | 8 | 6 | 4 | 29 | 19 | 11 | 7 | 49 | 33 | 1 | 1 | 11 | 7 |
| **4** | - | - | 18 | 12 | - | - | 1 | 1 | - | - | 12 | 8 | - | - | 12 | 8 | - | - | 0 | 0 |
| **5** | - | - | 0 | 0 | - | - | 0 | 0 | - | - | 0 | 0 | - | - | 3 | 2 | - | - | 2 | 2 |

^3L: Level 1 (no problems), Level 2 (some problems), Level 3 (extreme problems). ~5L: Level 1 (no problems), Level 2 (slight problems), Level 3 (moderate problems), Level 4 (severe problems), Level 5 (unable)

***Feasibility***

The proportion of completed questionnaires (i.e., health dimensions plus VAS score) was higher for the 5L (96.7%) than for the 3L (92.7%). Less than 1% of 5L questionnaires were returned blank compared to 1.3% for the 3L. Unsurprisingly the missing VAS score was higher for the 3L than 5L (4.0% vs 2.0%), as the VAS score appears two times in the questionnaire and participants might have seen this as duplication. The proportion of partially completed questionnaires (i.e., one to four missing domains) was higher for the 3L (3.31%) compared to the 5L (2.64%). Overall missing dimensions were on average 2.2 (1.45%) for the 3L and 2.4 (1.59%). The range of missing values per dimension was similar for both descriptive systems, ranging from 1.32% to 2.65% in mobility for the 5L; and from 1.32% to 1.99% in self-care for the 3L.

***Redistribution analysis***

Table 3 shows cross tabulations to the 3L and 5L indicating *inconsistent responses* in italics. On average, the proportion of inconsistent responses was 3.25%. The ‘usual activities’ domain had the highest frequency of inconsistent responses (6.76%) while self-care showed no inconsistency at all.

**Table 3. Cross tabulation for EQ-5D-3L and 5L**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **EQ-5D 5L** | | | | |
| **EQ-5D 3L** |  | *No problems* | *Slight problems* | *Moderate problems* | *Severe problems* | *Unable to/ Extreme* |
| **MOBILITY** |  |  |  |  |  |
| *No problems* | 46 | 12 | *0\** | *0\** | *0\** |
| *Some problems* | *2\** | 32 | 37 | 18 | *0\** |
| *Confined bed* | *0\** | *0\** | *0\** | 0 | 0 |
| **SELF-CARE** |  |  |  |  |  |
| *No problems* | 112 | 7 | *0\** | *0\** | *0\** |
| *Some problems* | *0\** | 15 | 12 | 1 | *0\** |
| *Unable to* | *0\** | *0\** | *0\** | 0 | 0 |
| **USUAL ACTIVITIES** |  |  |  |  |  |
| *No problems* | 57 | 16 | *4\** | *0\** | *0\** |
| *Some problems* | *5\** | 30 | 23 | 7 | *0\** |
| *Unable to* | *0\** | *0\** | *1\** | 5 | 0 |
| **PAIN/DISCOMFORT** |  |  |  |  |  |
| *None* | 14 | 13 | *1\** | *0\** | *0\** |
| *Moderate* | *5\** | 52 | 47 | 5 | *0\** |
| *Extreme* | *0\** | *0\** | *1\** | 7 | *3* |
| **ANXIETY/DEPRESSION** |  |  |  |  |  |
| *None* | 92 | 14 | *0\** | *0\** | *0\** |
| *Moderate* | *4\** | 25 | 11 | 0 | *\** |
| *Extreme* | *0\** | *0\** | *0\** | 0 | 1 |

\*Inconsistent responses are marked in bold

Redistribution from 3L to 5L is displayed in Table 4. The “some” responses on the 3L are reassigned between levels 2 to 4 (slight, moderate, severe) on the 5L, while the “severe” responses on the 3L are spread between level 4 (severe) and 5 (extreme) on the 5L. For the level 1 in 3L, there was always a higher proportion of 1→1 than 1→2. The most skewed relative frequency distribution was in self-care (94.1/5.9) and the least in pain/discomfort (51.9/48.1). For the level 2 in 3L, the most evenly spread proportion was in mobility (2→2: 36.8/ 2→3: 42.5/ 2→4: 20.7). Between 54.8% (pain/discomfort) and 69.4% (anxiety/depression) of the participants reporting level 2 (moderate problems) with the 3L answered 2 (slight problems) or 4 (severe problems) with the 5L. Only a few participants reported severe problems in the 3L, among these, participants chose the fourth level (severe) of the 5L in usual activities, the fifth level (extreme) in 5L for anxiety and depression; and were evenly spread between levels 4 and 5 in 5L in pain/discomfort. Generally, the median and the mean VAS scores decreased as participants move from a better to a worse level of health, indicating valid results for the majority of pair combinations of consistent responses. We observed a level of discrepancy in particular for anxiety/depression, where decreasing median VAS values were slightly higher than expected.

**Table 4. Redistribution of consistent responses from the EQ-5D-3L to the EQ-5D-5L**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dimension** | **3L** | **n** | **%** | **Pair**  **3L→5L** | **n** | **%** | **VAS 5L median\*** | **VAS 3L median** | **Difference 5L – 3L**  **Mean (SD)** |
| **Mobility** | **1** | 58 | 39.4 | **1→1** | 46 | 79.3 | 80 | 80 | 2.3 (7.4) |
|  |  |  |  | **1→2** | 12 | 20.7 | 82.5 | 82.5 | -0.1 (5.1) |
|  | **2** | 89 | 60.5 | **2→2** | 32 | 36.8 | 75 | 75 | 0.7 (3.8) |
|  |  |  |  | **2→3** | 37 | 42.5 | 65 | 65 | 2.2 (7.8) |
|  |  |  |  | **2→4** | 18 | 20.7 | 40 | 40 | -0.7 (6.8) |
|  | **3** | 0 | 0 | **3→4** | 0 | 0 | - | - | - |
|  |  |  |  | **3→5** | 0 | 0 | - | - | - |
| **Self-Care** | **1** | 119 | 81.0 | **1→1** | 112 | 94.1 | 80 | 80 | 1.9 (6.6) |
|  |  |  |  | **1→2** | 7 | 5.9 | 68 | 65 | 0.7 (7.8) |
|  | **2** | 28 | 19.0 | **2→2** | 15 | 53.6 | 50 | 59 | - 1.5 (6.7) |
|  |  |  |  | **2→3** | 12 | 42.9 | 45 | 47.5 | - 0.7 (6.3) |
|  |  |  |  | **2→4** | 1 | 3.6 | 25 | 25 | 0 |
|  | **3** | 0 | 0 | **3→4** | 0 | 0 | - | - | - |
|  |  |  |  | **3→5** | 0 | 0 | - | - | - |
| **Activities** | **1** | 77 | 52.0 | **1→1** | 57 | 78.1 | 85 | 80 | 2.0 |
|  |  |  |  | **1→2** | 16 | 21.9 | 75 | 74.5 | -0.4 (7.1) |
|  | **2** | 65 | 44.0 | **2→2** | 30 | 50.0 | 75 | 75 | 1.7 (5.4) |
|  |  |  |  | **2→3** | 23 | 38.3 | 50 | 50 | -0.9 (6.9) |
|  |  |  |  | **2→4** | 7 | 11.7 | 45 | 45 | 3.1 (5.0) |
|  | **3** | 6 | 4.0 | **3→4** | 5 | 100.0 | 40 | 37 | 1.8 (2.1) |
|  |  |  |  | **3→5** | 0 | 0 | - | - | - |
| **Pain/** | **1** | 28 | 19.0 | 1→1 | 14 | 51.9 | 80 | 79 | 0.6 (1.7) |
| **Discomfort** |  |  |  | 1→2 | 13 | 48.1 | 85 | 85 | 2.5 (4.8) |
|  | **2** | 109 | 73.6 | **2→2** | 52 | 50.0 | 80 | 80 | 2.2 (6.7) |
|  |  |  |  | **2→3** | 47 | 45.2 | 66.5 | 67 | 0.4 (6.1) |
|  |  |  |  | **2→4** | 5 | 4.8 | 50 | 50 | 0.8 (7.2) |
|  | **3** | 11 | 7.4 | **3→4** | 7 | 70.0 | 30 | 40 | -4.3 (7.9) |
|  |  |  |  | **3→5** | 3 | 30.0 | 40 | 39 | 0.3 (5.6) |
| **Anxiety/** | **1** | 106 | 71.6 | **1→1** | 92 | 86.8 | 80 | 80 | 1.3 (6.7) |
| **Depression** |  |  |  | **1→2** | 14 | 13.2 | 75 | 74.5 | -0.6 (6.4) |
|  | **2** | 41 | 27.7 | **2→2** | 25 | 69.4 | 60 | 65 | 0.8 (6.7) |
|  |  |  |  | **2→3** | 11 | 30.6 | 45 | 40 | 2.7 (3.3) |
|  |  |  |  | **2→4** | 0 | 0 | - | - | - |
|  | **3** | 1 | 0.67 | **3→4** | 0 | 0 | - | - | - |
|  |  |  |  | **3→5** | 1 | 100.0 | 20 | 20 | 0 |

\*Dimension-specific rating scale values were not available. Therefore, we used VAS scale as comparator

***Ceiling effect***

Self-care and anxiety/depression were the domains showing the highest percentage of “no problems” in both the 3L and the 5L; while Pain/Discomfort was the domain showing the lowest percentage of “no problems” (Table 5). Moving from the 3L to the 5L showed a decrease of “no problem” respondents for all domains; where Pain/Discomfort showed the highest relative reduction in ceiling effect. 17 participants (11.26%) reported full health with the 3L (11111), and when moving to the 5L these participants reported: 41.2 % perfect health; 35.3% scored ‘slight problems’ in one dimension; 11.8% scored ‘slight problems’ in two dimensions; 5.9% scored ‘moderate problems’ in one dimension; and 5.8% ‘slight’ in two dimensions and ‘moderate’ one dimension. Conversely of 9 participants (5.96%) that reported full health on the 5L only 2 participants reported some problems in one dimension of the 3L (pain/discomfort). The percentage of REFORM participants that reported full health decreased from 11.2% with the 3L to 6% with the 5L, indicating an absolute ceiling reduction of 5.2 and a relative ceiling reduction of 47.1% on the full profile.

**Table 5. Proportion of “no problem” responses for both descriptive systems**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **EQ-5D-3L** | **EQ-5D-5L** | **Ceiling effect reduction** | | |
|  | **n (%)** | **n (%)** | **Absolute** | **Relative (%)** |
| **Full Health (11111)** | 17 (11.2) | 9 (6.0) | 5.2 | 47.1 |
| **Mobility** | 58 (38.9%) | 48 (32.7%) | 6.3 | 16.1 |
| **Self-Care** | 120 (81.1%) | 113 (75.8%) | 5.2 | 6.5 |
| **Usual-Activity** | 78 (52.4%) | 62 (41.6%) | 10.7 | 20.5 |
| **Pain/Discomfort** | 28 (18.8%) | 19 (12.8%) | 6.0 | 32.1 |
| **Anxiety/Depression** | 106 (71.1%) | 96 (64.4%) | 6.7 | 9.4 |

\* REFORM participants were recruited from NHS podiatry clinics. Participants were eligible for inclusion in the trial if they had had a fall in the past 12 months, or a fall in the past 24 months requiring hospital attention or reported worrying about falling at least some of the time in the 4 weeks prior to completing their baseline questionnaire. Therefore, the ceiling effect was estimated assuming that none of REFORM participants were in full health.

***Discriminatory power***

The Shannon’s indices were calculated both by dimension and for each instrument as a whole. The results, by dimension, are shown in Table 6. It can be shown that the absolute informativity () increased for each dimension when moving from the 3L to the 5L. The relative informativity (J’) increased for all dimensions, except ‘anxiety/depression’.

**Table 6. Shannon’s Indices for the EQ-5D-3L and EQ-5D-5L by dimension.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **EQ-5D-3L** | | **EQ-5D-5L** | |
|  |  |  |  |
| **Mobility** | 0.96 | 0.61 | 1.92 | 0.83 |
| **Self-Care** | 0.70 | 0.44 | 1.06 | 0.46 |
| **Usual Activities** | 1.20 | 0.76 | 1.80 | 0.78 |
| **Pain/Discomfort** | 1.05 | 0.67 | 1.83 | 0.79 |
| **Anxiety/Depression** | 0.91 | 0.58 | 1.28 | 0.55 |

\*H’ represents the absolute amount of informativity captured. J’ represents the relative informativity of a system, regardless of the number of categories.

The Shannon Index and Evenness Index were also calculated for each instrument as a whole. When calculating the Evenness Index (J’) , varied based on the number of unique health states found in our sample of participants. The results are shown in Table 7. Again, the absolute informativity () increased when moving from the 3L to the 5L. This is to be expected due to the increase in levels. However, the relative informativity () also increased, indicating that the extra levels were being used and providing information.

**Table 7. Shannon’s Indices for the EQ-5D-3L and EQ-5D-5L overall.**

|  |  |  |
| --- | --- | --- |
|  | **EQ-5D-3L** | **EQ-5D-5L** |
|  | 3.91 | 5.48 |
|  | 4.64 | 6.07 |
|  | 0.84 | 0.90 |

\*H’ represents the absolute amount of informativity captured. J’ represents the relative informativity of a system, regardless of the number of categories.

***Association between multimorbidity and EQ5D index score***

At the overall level, the EQ-5D-5L scores were significantly higher than both EQ-5D-3L and crosswalk (Table 8). The mean difference between the EQ-5D-5L and EQ-5D-3L values was 0.091 (range -0.345 to 0.505); whilst the mean difference between the EQ-5D-5L and the crosswalk values was 0.082 (range -0.035 to 0.293). As expected, the percentage of states worse than dead was lower for the EQ-5D-5L. Overall long term health problems are associated with a reduction in utility values for all used value sets except among patients with diabetes; even though the decrements associated with specific health problems were different for both value sets.

**Table 8. Utility scores according to 3 systems.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***REFORM Sample*** | ***N*** | ***EQ-5D-3L*** | | |  | ***Crosswalk*** | | |  | ***EQ-5D-5L*** | | |  | ***Significance (p-value)*** | |
|  |  | *Mean (sd)* | *Range* | *SWD %* | *Sever order* | *Mean (sd)* | *Range* | *SWD %* | *Sever order* | *Mean (sd)* | *Range* | *SWD %* | *Sever order* | *3L-5L* | *5L-Crosswalk* |
| ***Overall*** | 151 | 0.679 (0.24) | 1 to -0.181 | 3.3 |  | 0.684 (0.20) | 1 to -0.042 | 2.0 |  | 0.766 (0.20) | 1 to 0.016 | 0 |  | 0.0000 | 0.0000 |
| ***Arthritis*** | 88 | 0.601 (0.26) | 1 to -0.181 | 5.7 | **4** | 0.608 (0.21) | 0.879 to -0.042 | 3.4 | **4** | 0.69.3 (0.22) | 0.937 to 0.016 | 0 | **3** | 0.0000 | 0.0000 |
| ***Diabetes*** | 56 | 0.681 (0.23) | 1 to -0.016 | 3.6 | 6 | 0.695 (0.19) | 1 to -0.021 | 1.8 | 6 | 0.774 (0.18) | 1 to 0.168 | 0 | 6 | 0.0000 | 0.0000 |
| ***Osteoporosis*** | 25 | 0.582 (0.28) | 1 to -0.016 | 8.0 | **3** | 0.595 (0.24) | 0.837 to -0.023 | 4.0 | **3** | 0.700 (0.22) | 0.115 to 0.937 | 0 | **4** | 0.0000 | 0.0000 |
| ***Depression*** | 10 | 0.500 (0.34) | 1 to -0.181 | 11.1 | 1 | 0.413 (0.28) | 0.75 to -0.042 | 20 | 1 | 0.479 (0.29) | 0.833 to 0.016 | 0 | 1 | 0.0000 | 0.0093 |
| ***> 1 condition*** | 78 | 0.650 (0.24) | 1 to -0.181 | 6.4 | 5 | 0.654 (0.20) | 1 to -0.042 | 3.9 | 5 | 0.736 (0.20) | 1 to 0.016 | 0 | 5 | 0.1988 | 0.0000 |
| ***> 2 condition*** | 123 | 0.544 (0.29) | 1 to -0.181 | 10.4 | 2 | 0.556 (0.24) | 0.837 to -0.042 | 6.1 | 2 | 0.646 (0.24) | 0.937 to 0.016 | 0 | 2 | 0.2635 | 0.0000 |

**Discussion**

The aim of our study was to compare the performance of the new EQ-5D-5L in the context of the REFORM trial in terms of feasibility, ceiling effect, redistribution properties and discriminatory power. Results showed that both descriptive systems showed good feasibility. Redistribution was confirmed indicating valid results for most pair combinations of consistent responses; showing that the elderly population in the REFORM trial were able to consistently respond to both the 3L and the 5L. Compared to the 3L, the 5L reduced ceiling effects. Likewise, the 5L had a higher absolute discriminatory power in all five dimensions; similarly, the relative discriminatory power was slightly better in the 5L than the 3L. The EQ-5D-5L scores were significantly higher than both EQ-5D-3L and crosswalk.

It could be argued that the completion of the 5L and the 3L was not randomised, which may have affected our study by the way participants completed the first and the second instrument. However, the same study design has been followed in other studies as this has been proved to lead to more efficient design and avoid order bias. Overall, our results are supported by other studies; ceilingeffect reductions for the 5L were expected and similar to the reductions found in other studies [11; 17-19], which confirm the 5L would be an adequate measure of health related quality of life among the elderly who suffer from two or more chronic conditions. As in previous studies [11,18-20], we found valid redistribution showing that the largest impact of the addition or two extra intermediate levels is to spread the “some” responses on 3L between levels 2 to 4 on the 5L. Regarding discriminative properties we also found a greater discriminative ability of the new 5L descriptive system [11,18, 21].

When we assessed feasibility, we observed that the amount of invalid questionnaires is lower for the 5L. This observation might have helpful implications for handling missing data in RCTs. Missing data are a frequent problem in RCTs irrespective of how well designed the data collection is; and the challenge is generally even greater for economic data. QALYs are cumulative measures hence missing dimensions on EQ-5D at one follow-up point implies that the aggregate variable (e.g. total QALYs over the trial) is also missing. Missing data can produce different cost-effectiveness results and alter decisions on the value for money of healthcare interventions. Therefore, considering this finding, the use of the EQ-5D-5L may be preferable to the EQ-5D-3L to lower the impact of missing data in cost-effectiveness analysis conducted alongside clinical trials.

Our results highlight differences in the utility values depending on the value set used in the analysis. Results showed the 5L shifts mean utilities up in the utility scale towards full health and the overall range of values is smaller compared to the 3L when administered to an elderly population. The same pattern is observed across all long-term health conditions for REFORM participants, which is more accentuated for depression. The differences in utilities produced by the UK 3L, the crosswalk and the England 5L value sets are supported by other studies [20, 22]. Values for the 5L are expected to be higher because the 3L value set has a lower minimum value and a larger proportion of states that are considered worse than dead compared to the 5L. The economic evaluation conducted alongside the REFORM trial used the EQ-5D-3L; the intervention was found to be cost-effective with a marginal gain in QALYs compared to usual care. However, we cannot compare the impact on cost-effectiveness decisions of moving from the 3L to the new descriptive system, as the 5L was only collected at baseline. Therefore, the implications of these findings for the decision-making process by NICE remain unclear. A recent simulation based study showed that 3L and 5L can produce substantially different estimates of cost-effectiveness in a number of health conditions, severities and technologies [22]. The authors concluded that interventions that improve quality of life are more likely to be considered less cost-effective if the 5L is used, while the cost-effectiveness of interventions driven by mortality rather than morbidity would be improved if the 5L is used in place of the 3L. REFORM cost-effectiveness is not driven by mortality therefore it might be possible that shifting to the 5L would have made the intervention become less cost-effective. However, this is difficult to predict. NICE’s Decision Support Unit has looked at the differences between both value sets [23]. In particular the weight given to mobility decreased in the 5L relative to the 3L; therefore, as the REFORM intervention focused on mobility we would expect lower QALYs using the EQ-5D-5L. However, at the time of writing, the impact of adopting the EQ-5D-5L value set in England is still unclear.

**Conclusion**

In this clinical trial involving an elderly population both the EQ-5D-3L and the EQ-5D-5L showed good feasibility. However, the use of the 5L reduced the ceiling effects and improved discriminatory power. Likewise, the 5L instrument is likely to lower the problem of missing data in cost-effectiveness analysis. Further research is required to explore the impact of using the new EQ-5D-5L value set on estimates of QALYs gained.

**Data availability**

***Underlying data***

Full underlying (non-aggregated) data cannot be made publicly available since the ethics approval of this study does not cover openly publishing non-aggregated data.  
In order to access this data, it must be requested from the corresponding author. Data requestors will have to provide: i) written description and legally binding confirmation that their data use is within the scope of the study; ii) detailed written description and legally binding confirmation of their actions to be taken to protect the data (e.g. with regard to transfer, storage, back-up, destruction, misuse, and use by other parties), as legally required and to current national and international standards (data protection concept); and iii) legally binding and written confirmation and description that their use of this data is in line with all applicable national and international laws (e.g. the General Data Protection Regulation of the EU).

**Competing interests**

No competing interests were disclosed.

**Grant information**

The REFORM trial was funded by the NIHR Health Technology Assessment Programme (project number 09/77/01) and is published in full in the NIHR Journals Library [Health technology assessment. pp. 1-198. ISSN 2046-4924].

This manuscript presents independent research commissioned by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this manuscript are those of the authors and do not necessarily reflect those of the NHS, the NIHR HTA programme or the Department of Health.

REFORM trial registration number:ISRCTN68240461

**Acknowledgments**

We would like to thank the participants of this study for their invaluable input. We also thank Professor Gerry Richardson for his feedback on the manuscript’s results.

**References**

1. National Institute for Health and Care Excellence. Guide to the Methods of Technology Appraisal 2013. Process and methods [PMG9]. London: National Institute for Health and Care Excellence; 2013.

2. Langley PC. Dreamtime: Version 5.0 of the Australian Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC). INNOVATIONS in pharmacy. 2017 Jan 10;8(1).

3. Brady BK, Amin M, Mensinkai S. PHP26 CCOHTA GUIDELINES FOR THE ECONOMIC EVALUATION OF HEALTH TECHNOLOGIES: CANADA 2004. Value in Health. 2004 May 1;7(3):374-5.

4. Dolan P. Modeling valuations for EuroQol health states. Medical care. 1997 Nov 1:1095-108.

5. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27:7–22

6. 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 Jul 1;15(5):708-15.

7. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of life research. 2011 Dec;20(10):1727-36.

8. Cockayne, S., J. Adamson, B.C. Martin, C. Fairhurst, C. Hewitt, K. Hicks, R. Hull, A.M. Keenan, S.E. Lamb, and L. Loughrey, *The REFORM study protocol: a cohort randomised controlled trial of a multifaceted podiatry intervention for the prevention of falls in older people.* BMJ open, 2014. **4**(12): p. e006977.

9. Cockayne, S., S. Rodgers, L. Green, C. Fairhurst, J. Adamson, A. Scantlebury, B. Corbacho, C.E. Hewitt, K. Hicks, and R. Hull, *Clinical effectiveness and cost-effectiveness of a multifaceted podiatry intervention for falls prevention in older people: a multicentre cohort randomised controlled trial (the REducing Falls with ORthoses and a Multifaceted podiatry intervention trial).* Health technology assessment, 2017: p. 1-198.

10. Corbacho B, Cockayne S, Fairhurst C, Hewitt CE, Hicks K, Kenan AM, Lamb SE, MacIntosh C, Menz HB, Redmond AC, Rodgers S, Scantlebury A, Watson J, Torgerson DJ; on behalf of the REFORM study. Cost-Effectiveness of a Multifaceted Podiatry Intervention for the Prevention of Falls in Older People: The REducing Falls with Orthoses and a Multifaceted Podiatry Intervention Trial Findings. Gerontology. 2018;64(5):503-512. doi: 10.1159/000489171. Epub 2018 Jun 26. PMID: 29945150.

11. Janssen, M.F., Pickard, A.S., Golicki, D., Gudex, C., Niewada, M., Scalone, L., Swinburn, P. and Busschbach, J., 2013. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research*, *22*(7), pp.1717-1727.

12. Schuling J, De Haan R, Limburg MT, Groenier KH. The Frenchay Activities Index. Assessment of functional status in stroke patients. Stroke. 1993 Aug;24(8):1173-7.

13. Sheikh JI, Yesavage JA. Geriatric Depression Scale (GDS): recent evidence and development of a shorter version. Clinical Gerontologist: The Journal of Aging and Mental Health. 1986 Jun.

14. Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value in Health. 2008 Mar 1;11(2):275-84.

15. Janssen MF, Birnie E, Bonsel GJ. Evaluating the discriminatory power of EQ-5D, HUI2 and HUI3 in a US general population survey using Shannon’s indices. Quality of Life Research. 2007 Jun;16(5):895-904.

16. Polinder S, Haagsma JA, Belt E, Lyons RA, Erasmus V, Lund J, van Beeck EF. A systematic review of studies measuring health-related quality of life of general injury populations. BMC Public Health. 2010 Dec;10(1):1-3.

17. Ferreira LN, Ferreira PL, Ribeiro FP, Pereira LN. Comparing the performance of the EQ-5D-3L and the EQ-5D-5L in young Portuguese adults. Health and Quality of Life Outcomes. 2016 Dec;14(1):1-0.

18. Kim SH, Kim HJ, Lee SI, Jo MW. Comparing the psychometric properties of the EQ-5D-3L and EQ-5D-5L in cancer patients in Korea. Quality of Life Research. 2012 Aug;21(6):1065-73.

19. Scalone L, Ciampichini R, Fagiuoli S, Gardini I, Fusco F, Gaeta L, Del Prete A, Cesana G, Mantovani LG. Comparing the performance of the standard EQ-5D 3L with the new version EQ-5D 5L in patients with chronic hepatic diseases. Quality of Life Research. 2013 Sep;22(7):1707-16.

20. Mulhern B, Feng Y, Shah K, Janssen MF, Herdman M, van Hout B, Devlin N. Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets. Pharmacoeconomics. 2018 Jun;36(6):699-713.

21. Agborsangaya CB, Lahtinen M, Cooke T, Johnson JA. Comparing the EQ-5D 3L and 5L: measurement properties and association with chronic conditions and multimorbidity in the general population. Health and quality of life outcomes. 2014 Dec;12(1):1-7.

22. Hernandez Alava, M., A. Wailoo, S. Grimm, S. Pudney, M. Gomes, Z. Sadique, D. Meads, J. O'Dwyer, G. Barton, and L. Irvine, *EQ-5D-5L versus 3L: the impact on cost-effectiveness.* 2017.

23. Wailoo A, Hernández Alava M, Grimm S, et al. Comparing the EQ-5D-3L and 5L versions. What are the implications for cost effectiveness estimates? Decision Support Unit. http://nicedsu.org.uk/wp-content/uploads/2017/05/ DSU\_3L-to-5L-FINAL.pdf. Accessed September 22, 2020