



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect

journal homepage: [www.elsevier.com/locate/jval](http://www.elsevier.com/locate/jval)

## EQ-5D-5L versus EQ-5D-3L: The Impact on Cost Effectiveness in the United Kingdom

Monica Hernandez Alava, BSc, MSc, PhD<sup>1</sup>, Allan Wailoo, BSc, MA, MSc, PhD<sup>1,\*</sup>, Sabine Grimm, BA, MSc, PhD<sup>2</sup>, Stephen Pudney, BSc, MSc<sup>1</sup>, Manuel Gomes, BSc, MSc, PhD<sup>3</sup>, Zia Sadique, MSc, PhD<sup>3</sup>, David Meads, BA, MSc, PhD<sup>4</sup>, John O'Dwyer, MSc<sup>4</sup>, Garry Barton, BA, MSc, PhD<sup>5</sup>, Lisa Irvine, BA, MSc<sup>5</sup>

<sup>1</sup>School of Health and Related Research, University of Sheffield, Sheffield, South Yorkshire, UK; <sup>2</sup>Maastricht University Medical Centre, Maastricht, The Netherlands; <sup>3</sup>London School of Hygiene and Tropical Medicine, London, UK; <sup>4</sup>University of Leeds, Leeds, UK; <sup>5</sup>Norwich Medical School, University of East Anglia, Norwich, UK

### ABSTRACT

**Objectives:** To model the relationship between the three-level (3L) and the five-level (5L) EuroQol five-dimensional questionnaire and examine how differences have an impact on cost effectiveness in case studies. **Methods:** We used two data sets that included the 3L and 5L versions from the same respondents. The EuroQol Group data set ( $n = 3551$ ) included patients with different diseases and a healthy cohort. The National Data Bank data set included patients with rheumatoid disease ( $n = 5205$ ). We estimated a system of ordinal regressions in each data set using copula models to link responses of the 3L instrument to those of the 5L instrument and its UK tariff, and vice versa. Results were applied to nine cost-effectiveness studies. **Results:** Best-fitting models differed between the EuroQol Group and the National Data Bank data sets in terms of the explanatory variables, copulas, and coefficients. In both cases, the coefficients of the covariates and latent factors between the 3L and the 5L instruments were significantly different, indicating that moving between instruments is not simply a uniform re-alignment of the response levels for

most dimensions. In the case studies, moving from the 3L to the 5L caused a decrease of up to 87% in incremental quality-adjusted life-years gained from effective technologies in almost all cases. Incremental cost-effectiveness ratios increased, often substantially. Conversely, one technology with a significant mortality gain saw increased incremental quality-adjusted life-years. **Conclusions:** The 5L shifts mean utility scores up the utility scale toward full health and compresses them into a smaller range, compared with the 3L. Improvements in quality of life are valued less using the 5L than using the 3L. The 3L and the 5L can produce substantially different estimates of cost effectiveness. There is no simple proportional adjustment that can be made to reconcile these differences.

**Keywords:** cost effectiveness, EQ-5D, health utility.

Copyright © 2018, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

### Introduction

The EuroQol five-dimensional questionnaire (EQ-5D) comprises a descriptive system of health-related quality of life and associated tariffs or “utility” scores. The descriptive system covers five dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain/discomfort, and anxiety/depression. The original version of the EQ-5D allows respondents to indicate the degree of impairment on each dimension according to three levels (no problems, some problems, and extreme problems). This is the three-level EQ-5D (3L). The five-level EQ-5D (5L) is a new version of the instrument, which includes five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems), with the intention of improving the instrument’s sensitivity and reducing ceiling effects [1]. Tariffs are anchored around 1 for full health and 0 for states considered equivalent to death, on the basis of estimates from

samples of the general population. For the 3L version, these tariffs were based on a time trade-off valuation method. In the United Kingdom, these tariffs range from 1 for full health to  $-0.594$ . Thirty-five percent (84 of 243) of the health states are valued with a negative score. There is a gap between full health and the next level of impairment valued at 0.883. Tariffs for the 5L version are now available for England [2], Canada [3], Japan [4], Uruguay [5], The Netherlands [6], and Korea [7]. The valuation methods for the 5L used a combination of updated “lead time” time trade-off methods and discrete-choice experiments [8]. In England, this has led to a smaller range of values (from 1 to  $-0.281$ ), a smaller gap at the upper end of the distribution (0.951 is the next score after 1), and fewer values less than 0 (153 of 3125 [5%]).

The EQ-5D is one of the most widely used instruments underpinning economic evaluations conducted in terms of cost-per quality-adjusted life-year (QALY), calculated from the tariff scores. It is therefore essential to understand the implications of

\* Address correspondence to: Allan Wailoo, School of Health and Related Research, University of Sheffield, 30 Regent Street, Sheffield, South Yorkshire S1 4DA, UK.

E-mail: [a.j.wailoo@sheffield.ac.uk](mailto:a.j.wailoo@sheffield.ac.uk)

1098-3015/\$36.00 – see front matter Copyright © 2018, International Society for Pharmacoeconomics and Outcomes Research (ISPOR).

Published by Elsevier Inc.

<http://dx.doi.org/10.1016/j.jval.2017.09.004>

using the new 5L version of the instrument compared with using the 3L version. This article provides information on how the two versions of the EQ-5D relate to each other, using the UK/English tariffs. It should be noted that this 5L tariff could be subject to change as it progresses through the peer review process. We used two data sets in which respondents filled in both the 3L and the 5L instruments. We estimated the joint distribution of responses to the two instruments. This model was then used in nine cost-effectiveness studies to compare results when using directly observed 3L values with estimated 5L results.

## Methods

### Data

We used two data sets. The first was provided by the EuroQol Group (the EQG data). Between August 2009 and September 2010, the EuroQol Group coordinated and partly funded a data collection study. Its main aim was to collect data on both versions of the EQ-5D, the 3L and the 5L, to compare them in terms of their measurement properties and to generate an interim value set for the 5L using a mapping (or cross-walk) approach. The questionnaire introduced the 5L version first, followed by a few background questions (age, sex, education, etc.), and then the 3L version, the EQ-5D visual analogue scale, a set of five dimension-specific rating scales, and finally the World Health Organization (five) Well-Being Index. The study was carried out in six countries, Denmark, England, Italy, The Netherlands, Poland, and Scotland, and included eight broad patient groups (cardiovascular disease, respiratory disease, depression, diabetes, liver disease, personality disorders, arthritis, and stroke) and a student cohort (healthy population). Each country used the official EQ-5D language versions, and data were mainly collected through specialist hospitals/centers and patient recruitment agencies. All countries used paper-and-pencil questionnaires, apart from England, which used an online version. In all countries, except Italy, sampling methods ensured a wide range of severity across all the 5L and 3L dimensions.

The National Data Bank (NDB) for rheumatic diseases is a register of patients with rheumatoid disease, primarily recruited by referral from US and Canadian rheumatologists. Information supplied by participants is validated by direct reference to records held by hospitals and physicians. (A minority of cases come by self-referral, with medical details obtained by the NDB in the same way.) Full details of the recruitment process are given by Wolfe and Michaud [9]. The EQ-5D responses and other patient-supplied data are collected by various means, primarily postal and Web-based questionnaires completed directly by patients. Data collection began in 1998 and continues to the present, in waves administered in January and July of each year. In 2011, there was a switch from the 3L to the 5L version of the EQ-5D and both versions were included in the January 2011 wave. The NDB questionnaire is 27 pages long and it includes many general as well as specific questions on rheumatoid arthritis. The 5L and 3L versions are on pages 11 and 22 of the questionnaire, respectively. This wave is used to estimate the model.

### Statistical Analysis

The aim is to estimate the relationship between the two instruments. Hernandez and Pudney [10] have previously developed a flexible model that allows analysis of the joint responses to the 3L and 5L versions. Full details are provided there. Responses to the 3L and 5L versions are ordinal. The model reflects this in a system of 10 ordinal regressions, each of which is used to estimate the response level for one of the two versions of the

EQ-5D conditional on covariates (age and sex), arranged into the five health domains. The model reflects any tendency for an individual to give more or less positive responses across domains via a latent factor representing background response behavior. A copula approach is used to specify the bivariate distribution of each “3L, 5L” pair of responses. This captures the strong association between 3L and 5L responses within each health domain, without necessarily assuming that the strength of the association is the same in all parts of the health distribution. For example, someone who has experienced extreme pain may answer the pain questions in a more focused and coherent way than someone without experience of chronic pain. Five different copulas were examined (Gaussian, Clayton, Frank, Gumbel, and Joe), which reflect different types and strengths of dependence at different parts of the distribution, with the data informing the most appropriate final choice of copula.

Such statistical models are sensitive to the distributional assumptions, the usual one being normality. Mis-specification of the joint residual distribution may lead to significant bias in the estimated coefficients of the covariates, in addition to giving a distorted picture of the dependence. For this reason, mixture distributions are used to allow for non-normality in the residuals and the latent factor representing the individual's response behavior.

### Cost-Effectiveness Case Studies

We used the copula mapping models in nine cost-effectiveness case studies. All were economic evaluations based on individual patient-level data using the 3L version. We made a pragmatic decision in selecting case studies. We sought collaborators who had previously completed suitable studies using the 3L instrument and who were willing and able to replicate their study substituting predicted utility scores for the 5L instrument using a bespoke Stata command (StataCorp, College Station, TX). The included studies are as follows:

1. The Combination of Anti-Rheumatic Drugs in Early Rheumatoid Arthritis (CARDERA) trial was a double-blind, factorial designed, placebo-controlled randomized trial that compared the benefits of adding cyclosporine, high-dose step-down prednisolone, or both to methotrexate monotherapy [11]. The 3L version was administered to patients at baseline and at 6, 12, 18, and 24 months [12].
2. The Cost-effectiveness of Aphasia Computer Treatment Compared to Usual Stimulation (CACTUS) pilot randomized controlled trial tested the feasibility of comparing self-managed computer therapy combined with usual stimulation (such as participation in normal language stimulation activities and support groups) with usual stimulation alone in people with aphasia [13]. The 3L version was completed at baseline and at 3 and 8 months.
3. The Risk Adjustment in Neurocritical care (RAIN) trial compared 1) management in a dedicated neurocritical care unit versus a combined neuro/general critical care unit and 2) “early” transfer to a neuroscience center versus “no or late” transfer, for patients who initially present at a non-neuroscience center and do not require urgent neurosurgery and for patients with acute traumatic brain injury. The 3L version was completed at 3 months.
4. The Immediate Management of Patients with Rupture: Open Versus Endovascular Repair (IMPROVE) trial compared either endovascular repair or open repair of ruptured abdominal aortic aneurysm [14]. The 3L version was administered at 3 and 12 months.
5. The COUGAR-02 randomized, controlled, open-labeled trial compared docetaxel chemotherapy plus active symptom control (DXL + ASC) and ASC-only in patients in the United Kingdom with advanced adenocarcinoma of the esophagus,

**Table 1 – Descriptive statistics in the EQG and NDB estimation samples.**

	EQG sample		NDB sample	
Age (y)				
Mean (95% CI)	51.23 (50.57–51.89)		63.32 (62.99–63.65)	
Median (95% CI)	54 (54–56)		64.13 (63.78–64.46)	
SD	20.11		12.31	
Minimum	13		16.66	
Maximum	99		95.20	
Proportion female	0.53		0.81	
	EQ-5D-3L	EQ-5D-5L	EQ-5D-3L	EQ-5D-5L
Utility				
Mean (95% CI)	0.628 (0.617–0.639)	0.712 (0.703–0.722)	0.681 (0.674–0.688)	0.779 (0.773–0.784)
Median (95% CI)	0.691 (0.691–0.725)	0.802 (0.792–0.816)	0.725 (0.725–0.727)	0.823 (0.817–0.829)
SD	0.333	0.278	0.254	0.191
Minimum	–0.594	–0.281	–0.594	–0.226
Maximum	1	1	1	1
No. of health states (percentage out of possible health states)	123 (50.62)	660 (21.12)	86 (35.39)	524 (16.77)

CI, confidence interval; EQG, EuroQol Group; NDB, National Data Bank; EQ-5D-3L, three-level EuroQol five-dimensional questionnaire; EQ-5D-5L, five-level EQ-5D.

esophagogastric junction, or stomach [15]. Patients completed the EQ-5D at baseline, during clinic visits at weeks 3, 6, 9, and 12, then every 6 weeks for up to 1 year, and then every 3 months until death.

- The Attenuated dose Rituximab with ChemoTherapy in Chronic lymphocytic leukemia (ARCTIC) study was a multicenter, randomized, controlled, open, phase IIB noninferiority trial conducted in previously untreated patients with chronic lymphocytic leukemia [16,17]. It compared the combination of fludarabine, cyclophosphamide, and rituximab, which is considered conventional frontline therapy, with that of fludarabine, cyclophosphamide, mitoxantrone, and low-dose rituximab. The 3L version was completed at baseline, after three cycles of therapy, at the end of therapy, 3 months after the end of therapy, and then every 3 months after the end of therapy until 24 months postrandomization (i.e., at 6, 9, 12, 18, and 24 months postrandomization).
- The Self-Help and Relapse Prevention in Smoking for Health (SHARPISH) trial [18] sought to estimate the effectiveness and cost effectiveness of self-help booklets versus a single leaflet to prevent smoking relapse in people who had stopped smoking for 4 weeks. The 3L version was administered at baseline and at 2 months and 11 months postrandomization.
- The Weight-Reduction Activity Program (WRAP) [19] was a multicenter, nonblinded, three-arm parallel-group randomized controlled trial of two commercial weight loss programs, compared with a brief intervention in overweight adults. The 3L version was administered at baseline and at 3, 12, and 24 months.
- The Complete versus Lesion-only Revascularization for Myocardial Infarction (CvLPRIT) trial [20] randomized patients presenting with ST-segment elevation myocardial infarction with bystander stenosis to an infarct-only strategy (only treat the blocked artery that caused the heart attack) versus complete revascularization (treat the blocked artery and also treat any narrowed arteries that may cause heart attacks in future). The 3L version was administered immediately before discharge and at 12 months postdischarge.

We used the UK value sets for the 3L instrument and the English value set for the 5L instrument [21,22].

## Results

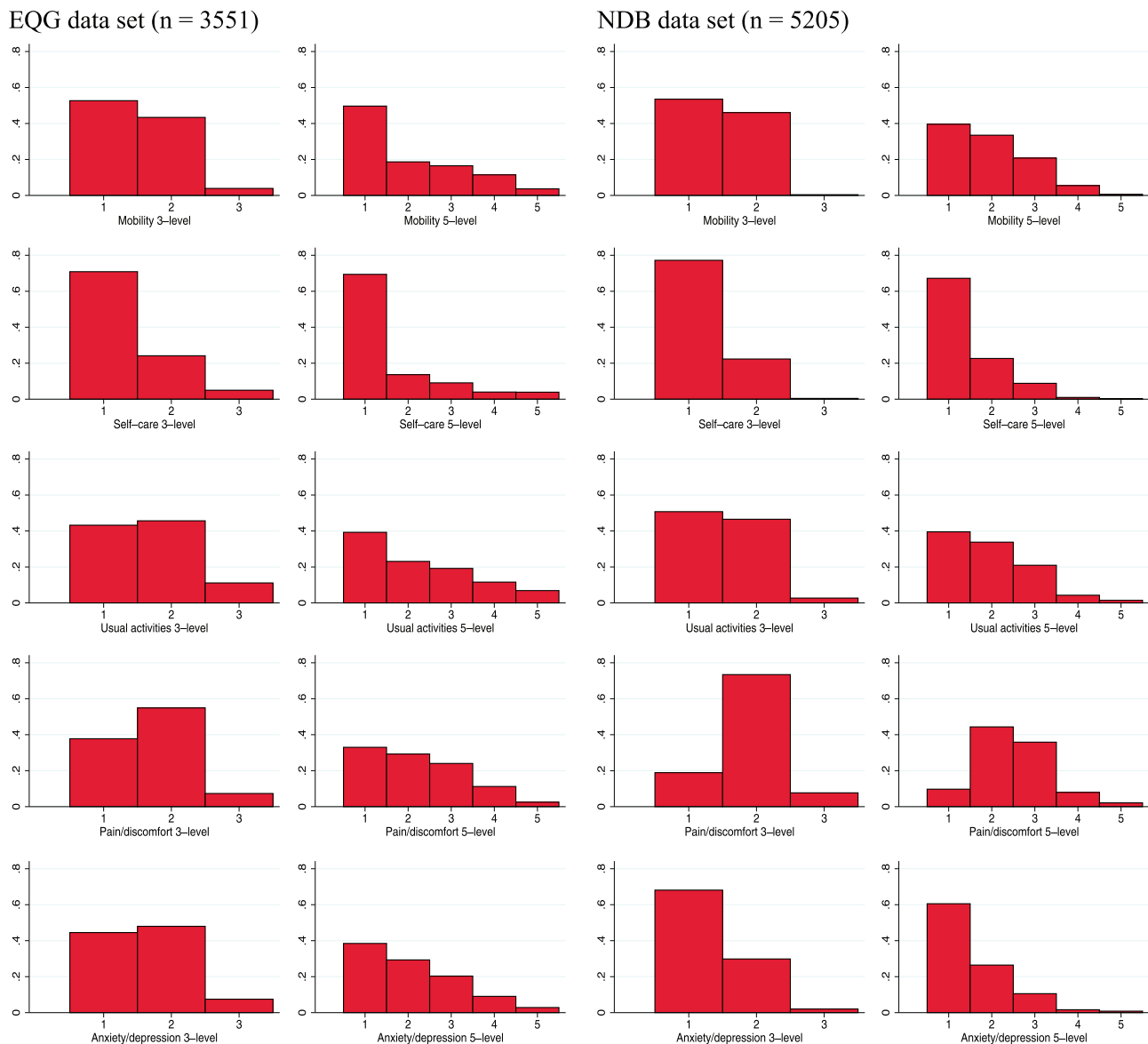
### Data Sets

After exclusion of missing values, there were final estimation samples of 3551 and 5205 respondents in the EQG and NDB data sets, respectively. The EQG sample was younger and contained more males than the NDB sample (see Table 1).

Figure 1 shows histograms of the response distributions for each dimension of the 3L and 5L versions of the EQ-5D in both data sets. There are differences both across the dimensions and between the data sets. Four distinct distributional shapes can be identified:

- Decreasing profile with a dominant mode at the first category:* This distributional shape can be seen in the self-care dimension of both the 3L and the 5L versions and in the mobility and usual activities dimensions of the 5L version in the EQG data set and on the self-care and anxiety/depression dimensions of both versions of the EQ-5D in the NDB data set.
- Decreasing profile with a heavier central section:* In the EQG data set, the pattern can be seen in the mobility dimension (3L) and in the pain/discomfort and anxiety/depression dimensions (5L). In the NDB data set, the mobility and usual activities dimensions for both the versions exhibit this shape.
- A strong mode in the center of the distribution:* This shape can be found in the pain/discomfort dimension in the 3L version in the EQG data set and in both the versions in the NDB data set.
- A mode in the center of the distribution and an almost as large first category:* This distributional shape is similar to shape 2 in that they both exhibit a decreasing profile, but shape 4 has less central concentration. This shape can be found only in the EQG data set in the usual activities and anxiety/depression dimensions of the 3L version.

In the NDB data set, both versions of the EQ-5D display the same pattern within each dimension but different shapes across dimensions: shape 1 in both the self-care and anxiety/depression dimensions, shape 2 in the mobility and usual activities dimension, and shape 3 in the pain/discomfort dimension. In contrast,



**Fig. 1 – Response histograms for EQ-5D-3L and EQ-5D-5L in the EQG data set and the NDB data set. For the 3L, level 1 = “no problems,” 2 = “some problems,” 3 = “extreme problems/unable to do.” For the 5L, level 1 = “no problems,” 2 = “slight problems,” 3 = “moderate problems,” 4 = “severe problems,” 5 = “extreme problems/unable to do.” EQ-5D-3L, three-level EuroQol five-dimensional questionnaire; EQ-5D-5L, five-level EQ-5D; EQG, EuroQol Group; NDB, National Data Bank.**

in the EQG data set, only the self-care dimension shows the same shape of distribution in both the 3L and 5L versions. In the EQG data set, the distributional shapes for all the dimensions of the 5L version are similar, displaying a decreasing profile corresponding to either shape 1 or shape 2. The 3L distributions in the EQG data set exhibit all four distributional shapes and appear more different across dimensions than in the 5L version. The variation in shape highlights the need to use flexible model specifications that do not impose the same model structure across dimensions or data sets.

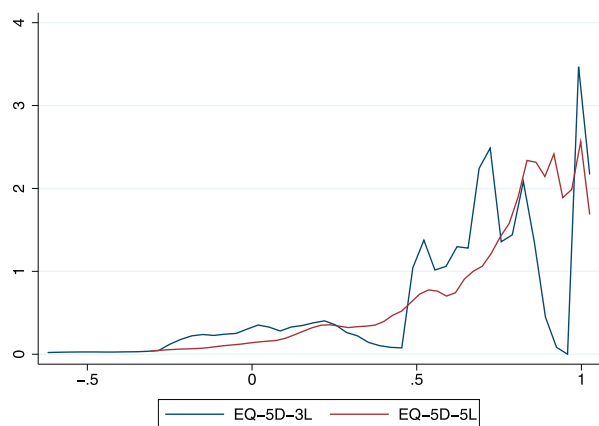
Figure 2 shows kernel estimates of the distributions of utility scores in both data sets. The 3L versions in both data sets exhibit the typical characteristics documented in the literature: a large mass of observations at 1 (full health), a gap of no observations between full health and the next feasible value (0.883), and a multimodal distribution. In both data sets, the distributions are smoother for the 5L version, especially toward the top of the

distribution. The number of individuals in full health is reduced by using the 5L version, and the mode at the bottom of the distribution around the value of 0 in the 3L distribution disappears in the 5L distribution. The mean and median of the 5L version are higher than the corresponding mean and median of the 3L version in both data sets (see Table 1). The range of the 5L version is smaller because the worst state has a utility score of  $-0.281$  compared with  $-0.594$  of the 3L version.

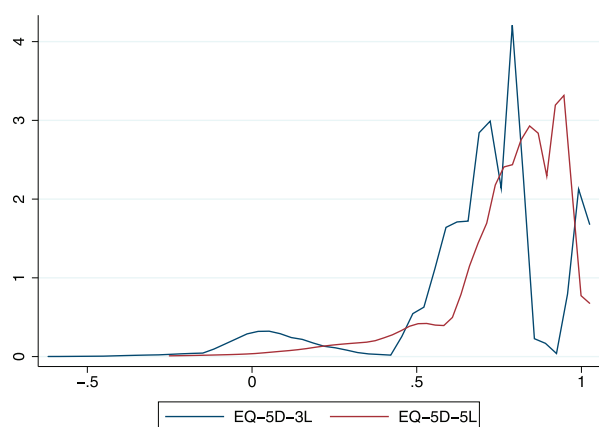
### Statistical Model Results

The initial specification had sex, age, and the square of age as covariates. The square of age was significant when the model was estimated with EQG data, but grossly insignificant when estimated with NDB data. The preferred specification for the EQG data set has age, age squared, and sex as covariates in all 10

## EQG data set (n = 3551)



## NDB data set (n = 5205)



**Fig. 2 – Smoothed empirical distribution functions of EQ-5D-3L and EQ-5D-5L in the EQG and NDB data sets. EQ-5D-3L, three-level EuroQol five-dimensional questionnaire; EQ-5D-5L, five-level EQ-5D; EQG, EuroQol Group; NDB, National Data Bank.**

ordinal regressions, whereas the model for the NDB data set excludes the square of age.

Table 2 presents the results for the two data sets. There are several differences between the models from the two data sets. The best-fitting model in the EQG data set chooses the same copula, Frank, in all dimensions of the EQ-5D. In contrast, the best-fitting model in the NDB data set selects a Gaussian copula for the mobility, usual activities, and pain/discomfort dimensions; a Clayton copula for the self-care dimension; and a Frank copula for the anxiety/depression dimension. The Gaussian and Frank copulas are similar in that both allow for positive or negative dependence, symmetric in both tails, but the Frank form generates dependence weaker in the tails and stronger in the center of the distribution. The Clayton copula allows only positive dependence, with strong left tail dependence and relatively weak right tail dependence; thus, if two variables are strongly correlated at low values but less so at high values, then the Clayton copula is a good choice. Therefore, in the EQG data set, the patterns of residual dependence between the 3L and 5L versions of the EQ-5D are similar across all dimensions, indicating symmetric dependence and weak dependence on the tails. In the NDB data set, a Frank copula was also selected for the anxiety/

depression dimension and the parameter of dependence was very similar to that estimated in the EQG data set. In contrast, the Gaussian copulas in the mobility, usual activities, and pain/discomfort dimensions indicate symmetric dependence as well but stronger dependence on the tails of the distribution than the Frank copula selected in the EQG data set. The copula chosen in the self-care dimension using the NDB data set, the Clayton copula, displays a very different pattern of dependence compared with the Frank copula chosen in the EQG data set. It exhibits asymmetric dependence on the tails, with strong dependence at lower values and weak dependence at high values.

There are significant statistical differences in the coefficients of the covariates and the latent factors between the 3L and the 5L versions in most dimensions. This is a test of the hypothesis that the underlying relationship between covariates and/or latent variables and the EQ-5D is the same for the 3L and 5L versions. Rejection of the hypothesis indicates that the effect of moving from the 3L to the 5L is not just a uniform re-alignment of the response levels. The only exception to this in both data sets is in the anxiety/depression and the self-care dimensions in the NDB data set.

### Cost-Effectiveness Results

Table 3 and Figure 3 report headline results for all the case studies. In almost all cases, the switch from the 3L to the 5L causes a decrease in the incremental QALY gain from effective health technologies. This is true whether the estimation of the 5L is based on the EQG or the NDB data, with one exception.

In COUGAR-02, there is an increase in incremental QALYs as a result of shifting from the 3L to the 5L. The increase is small but is apparent for both versions of 5L estimates. In COUGAR-2, mortality is a very substantial driver of cost effectiveness. Median overall survival in the DXL + ASC group was 5.2 months (95% confidence interval [CI] 4.1–5.9) versus 3.6 months (95% CI 3.3–4.4) in the ASC-only group [9]. Here, the value of improved survival is greater because utility values are increased when using the 5L. It is worth noting that although the RAIN study also included patients with a substantial mortality rate (~25% mortality within 6 months), this was substantially lower than in COUGAR-02 (approximate 6-month mortality of 75% in the control group and 60% in the DXL arm [9]) and did not outweigh the morbidity effect.

The responses people give to the 5L instrument and the changed tariff have the combined effect of shifting mean utility scores further up the utility scale toward full health, and compressing them into a smaller range. Thus, improvements in quality of life tend to be valued less using the 5L instrument compared with the same clinical change measured with the 3L instrument.

In six of the nine reported comparisons, the incremental QALY gain is greater when measured using the 5L and the EQG data set compared with using the 5L and the NDB data set. One of the three remaining comparisons showed no difference.

In those studies in which the 5L (EQG) lowered incremental QALYs, the impact ranged from a reduction of 10.4% (CARDERA comparison of methotrexate with methotrexate plus prednisolone) to 75% (RAIN comparison of dedicated neurocritical care unit with combined neuro/general critical care unit). The comparable range when using mapping on the basis of NDB data was 8% (CARDERA as before) to 87% (CACTUS).

The impact of these changes on incremental cost-effectiveness ratios (ICERs) is also substantial in several cases. In CARDERA, the comparison of triple therapy with monotherapy with disease-modifying antirheumatic drug changes from approximately £16,000 using the 3L to more than £24,000 using the 5L (EQG data) and more than £30,000 using the 5L (NDB data). CACTUS changes from a highly cost-effective central estimate using the 3L (£3,058) to one that is more borderline (£23,022) using the 5L (NDB data). CvLPRIT changes

**Table 2 – Summary of final model results.**

	EQG	NDB
Log likelihood	–23,891.83	–33,621.04
No. of parameters	78	68
Observations	3551	5205
Type of mixture in copula	Single mixture	Single mixture
Dimension-specific		
<i>Mobility</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	7.12 <sup>‡</sup>	11.86 <sup>†</sup>
Equality of coefficients (latent factor)	8.37 <sup>†</sup>	10.64 <sup>†</sup>
Equality of coefficients (covariates and factor)	12.19 <sup>†</sup>	26.49 <sup>†</sup>
<i>Self-care</i>		
Copula	Frank	Clayton
Equality of coefficients (covariates)	8.53 <sup>†</sup>	1.21
Equality of coefficients (latent factor)	3.68 <sup>‡</sup>	0.09
Equality of coefficients (covariates and factor)	9.39 <sup>‡</sup>	1.35
<i>Usual activities</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	3.29	0.67
Equality of coefficients (latent factor)	5.62 <sup>†</sup>	8.24 <sup>†</sup>
Equality of coefficients (covariates and factor)	0.04 <sup>†</sup>	9.11 <sup>†</sup>
<i>Pain/discomfort</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	0.57	34.36 <sup>†</sup>
Equality of coefficients (latent factor)	9.36 <sup>†</sup>	19.99 <sup>†</sup>
Equality of coefficients (covariates and factor)	11.95 <sup>†</sup>	50.74 <sup>†</sup>
<i>Anxiety/depression</i>		
Copula	Frank	Frank
Equality of coefficients (covariates)	5.60	4.94
Equality of coefficients (latent factor)	1.23	1.94
Equality of coefficients (covariates and factor)	7.08	6.19

EQG, EuroQol Group; NDB, National Data Bank.  
<sup>\*</sup> P = 0.10.  
<sup>†</sup> P = 0.01.  
<sup>‡</sup> P = 0.05.

from an ICER of just more than £20,000/QALY to in excess of £45,000/QALY when using either estimate of the 5L health utility. Other case studies demonstrate changes in cost effectiveness that may not span boundaries of typically cited cost-effectiveness thresholds but are, nevertheless, very substantial.

## Conclusions

We have shown that the 3L and 5L versions can produce substantially different estimates of cost effectiveness in a series of case studies spanning different health conditions, severities, and health technologies. Technologies that improve quality of life have those benefits valued more highly, in terms of health utility, when using the 3L instrument compared with the 5L instrument. This is because of the combined effect of the changed descriptive system and how individuals respond to it compared with the 3L (which we demonstrated is not the same across each health dimension) and the changed valuation system. The result is that, in almost all cases, it is estimated that the ICER of a clinically effective technology would be higher (i.e., becomes less cost-effective) if the 5L instrument had been used in place of the 3L instrument. When the cost effectiveness of a technology is substantially driven by mortality rather than by morbidity gains, the impact of shifting the 5L may lower ICERs (improve cost effectiveness). Consistent with our findings, a recent study that also used the EQG data set reported that the

5L leads to higher values overall and across all the health conditions in the EQG data set [23].

In this sense, estimates of health gain from the 3L and 5L are not consistent with each other. There is not a simple proportional adjustment that can be made to reconcile differences between the 3L and the 5L. Changes do not impact equally across the distribution of health and therefore different technologies are affected to a different degree by the shift from one instrument to another.

It is feasible to adjust 3L evidence to its 5L equivalent, as has been done in this article. The validity of this approach is, in part, dependent on the data on which it is based. We have demonstrated this method in two separate data sets and shown that they give substantially different results. Further investigation of the reasons for these differences is required. In particular, the NDB includes only patients with rheumatoid disease and may not be generalizable to other populations. Nevertheless, the design of the NDB questionnaires included much more separation between the completion of the 3L and the 5L and may, therefore, offer observations given without recall of previous responses than the EQG studies. The NDB study is also predominantly conducted in English. Although there is some evidence that the ranking of levels 4 and 5 (“severe” and “extreme” problems) may not be as expected in the English valuation study [2], this is less likely to be an issue affecting the descriptive system when respondents are provided with all five levels in their expected order. Therefore, we do not feel there is a rationale to prefer English-speaking samples for the mapping work. Both data sets are also limited by their size

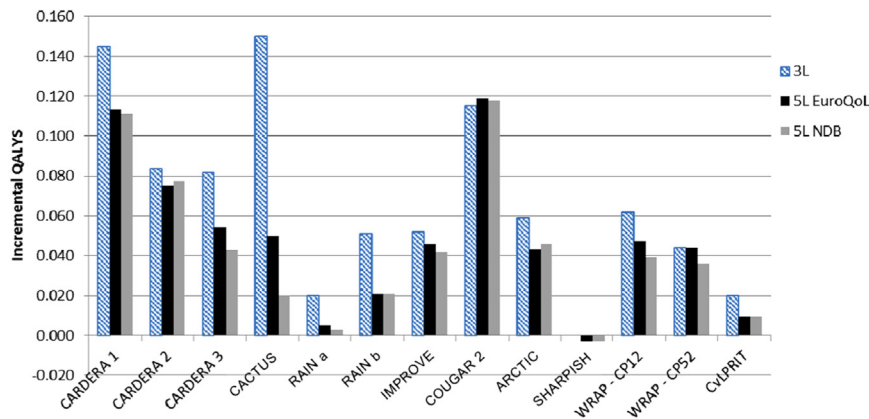
**Table 3 – Incremental QALYs and ICERs for 3L, 5L (EQG), and 5L (NDB) across all case studies.**

Study	Incremental QALYs					ICER				
	3L	5L (EQG)	% Change	5L (NDB)	% Change	3L	5L (EQG)	% Change	5L (NDB)	% Change
CARDERA 1	0.145	0.113	-21.8%	0.111	-23.2%	4648	5940	27.8%	6054	30.3%
CARDERA 2	0.084	0.075	-10.4%	0.077	-8.0%	13,666	15,252	11.6%	14,846	8.6%
CARDERA 3	0.082	0.054	-33.5%	0.043	-47.6%	15,929	23,940	50.3%	30,418	91.0%
CACTUS	0.150	0.050	-66.7%	0.020	-86.7%	3058	9481	210.0%	23,022	652.8%
RAIN a	0.020	0.005	-75.0%	0.003	-85.0%	184,700	738,800	300.0%	1,231,333	566.7%
RAIN b	0.051	0.021	-58.8%	0.021	-58.8%	294,137	714,333	142.9%	714,333	142.9%
IMPROVE	0.052	0.046	-11.5%	0.042	-19.2%	-44,617*	-48,113	7.8%	-54,742	22.7%
COUGAR-02	0.115	0.119	3.5%	0.118	2.6%	27,180	26,434	-2.7%	26,484	-2.6%
ARCTIC	0.059	0.043	-27.1%	0.046	-22.0%	112,193	162,774	45.1%	152,130	35.6%
SHARPISH	0.000	-0.003	NA	-0.003	NA	NA†				
WRAP-CP12	0.062	0.047	-23.7%	0.039	-36.2%	1812	2373	31.0%	2840	56.7%
WRAP-CP52	0.044	0.044	0.0%	0.036	-19.0%	4305	4312	0.2%	5316	23.5%
CvLPRIT	0.020	0.010	-52.5%	0.009	-53.0%	21,496	46,761	117.5%	47,521	121.1%

Note. CARDERA 1 = MTX vs. MTX + CS; CARDERA 2 = MTX vs. MTX + PNS; CARDERA 3 = MTX + CS + PNS vs. MTX. 3L, three-level EuroQol five-dimensional questionnaire; 5L, five-level EuroQol five-dimensional questionnaire; ARTIC, Attenuated dose Rituximab with ChemoTherapy in Chronic lymphocytic leukemia; CACTUS, Cost-effectiveness of Aphasia Computer Treatment Compared to Usual Stimulation; CARDERA, Combination of Anti-Rheumatic Drugs in Early Rheumatoid Arthritis; CS, cyclosporine; CvLPRIT, Complete-compared to Lesion-Only Revascularization for Myocardial Infarction trial; EQG, EuroQol Group; ICER, incremental cost-effectiveness ratio; IMPROVE, Immediate Management of Patients with Rupture: Open Versus Endovascular Repair; MTX, methotrexate; NA, not available; NDB, National Data Bank; PNS, prednisolone; QALY, quality-adjusted life-year; RAIN, Risk Adjustment in Neurocritical care; SHARPISH, Self-Help and Relapse Prevention in Smoking for Health; WRAP, Weight-Reduction Activity Program.  
 \* In the IMPROVE study, the technology of interest (endovascular aneurysm repair) was cost-saving.  
 † Incremental QALYs near 0 meant that the calculation of the ICER may be misleading and was therefore not reported.

and coverage of relevant health states. In the EQG data, only 119 of the possible 233 3L utility values are observed. That figure is 83 for the NDB data. We know that most of the 233 health states do appear in real patient records. For example, in the UK 2010 to 2014 data for knee replacement procedures (n = 320,000), we find 189 out of 233 possible utility values. There is a pressing need for well-designed, large-scale data collection to extend this work.

There are a number of implications for policy in the light of these results. Given the differences between the 3L and 5L instruments, consistency in decision making will be difficult to achieve. Consideration must be given to the value of any cost-effectiveness threshold (or thresholds) or other means for making adjustments between the two instruments. Mapping can help achieve this, and the copula-based method is a sophisticated development of ‘response mapping’ that obtains consistent and



**Fig. 3 – Histogram of incremental QALYs by 3L, 5L (EQG), and 5L (NDB) for all case studies. CARDERA 1 = MTX vs. MTX + CS, CARDERA 2 = MTX vs. MTX + PNS, CARDERA 3 = MTX + CS + PNS vs. MTX. CS; 3L, three-level EuroQol five-dimensional questionnaire; 5L, five-level EuroQol five-dimensional questionnaire; ARTIC, Attenuated dose Rituximab with ChemoTherapy in Chronic lymphocytic leukemia; CACTUS, Cost-effectiveness of Aphasia Computer Treatment Compared to Usual Stimulation; CARDERA, Combination of Anti-Rheumatic Drugs in Early Rheumatoid Arthritis; CS, cyclosporine; CvLPRIT, Complete-compared to Lesion-Only Revascularization for Myocardial Infarction trial; EQG, EuroQol Group; IMPROVE, Immediate Management of Patients with Rupture: Open Versus Endovascular Repair; MTX, methotrexate; NDB, National Data Bank; PNS, prednisolone; QALY, quality-adjusted life-year; RAIN, Risk Adjustment in Neurocritical care; SHARPISH, Self-Help and Relapse Prevention in Smoking for Health; WRAP, Weight-Reduction Activity Program.**

accurate results. A single approach to mapping between the 3L and 5L instruments would aid consistent decision making. Additional data collection would also permit extended validation of the method and comparison against the EuroQol “cross-walk” that provides a link between 5L responses and 3L responses [24]. Decision-making bodies, such as the National Institute for Health and Care Excellence in the United Kingdom, should endorse the use of either the 3L instrument or the 5L instrument and a set of methods that allow evidence to be linked from one to the other. The 5L instrument is increasingly being used in studies of clinical effectiveness, but this is unlikely to entirely replace existing evidence using the 3L instrument that will remain of relevance to many economic evaluations for many years to come.

## Acknowledgments

We thank Fred Wolfe and Kaleb Michaud for providing data from the NDB; the EuroQol Group for providing data and helpful feedback on a presentation at the International Society for Pharmacoeconomics and Outcomes Research, Vienna, 2016, in particular Paul Kind, Bas Jaansen, and Andrew Lloyd; and Jenny Dunn (School of Health and Related Research) for providing administrative support.

Source of financial support: This work was supported by a grant from the Medical Research Council Methodology Research Programme (grant no. MR/L022575/1) and the National Institute for Health and Care Excellence through its Decision Support Unit. The views represented are those of the authors alone.

## REFERENCES

- [1] Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011;20:1727–36.
- [2] Devlin N, Shah K, Feng Y, et al. Valuing health-related quality of life: an EQ-5D-5L value set for England. OHE Research Paper 16/01, Office of Health Economics London, 2016.
- [3] Xie F, Pullenayegum E, Gaebel K, et al. A time trade-off-derived value set of the EQ-5D-5L for Canada. *Med Care* 2016;54:98–105.
- [4] Shiroiwa T, Ikeda S, Noto S, et al. Comparison of value set based on DCE and/or TTO data: scoring for EQ-5D-5L health states in Japan. *Value Health* 2016;19:648–54.
- [5] Augustovski F, Rey-Ares L, Irazola V, et al. An EQ-5D-5L value set based on Uruguayan population preferences. *Qual Life Res* 2016;25:323–33.
- [6] Versteegh MM, Vermeulen KM, Evers SM, et al. Dutch tariff for the five-level version of EQ-5D. *Value Health* 2016;19:343–52.
- [7] Kim SH, Ahn J, Ock M, et al. The EQ-5D-5L valuation study in Korea. *Qual Life Res* 2016;25:1845–52.
- [8] Feng Y, Devlin N, Shah K, et al. New methods for modelling EQ-5D-5L value sets: an application to English data. Available from: <https://www.ohe.org/publications/new-methods-modelling-eq-5d-5l-value-sets-application-english-data>. [Accessed May 23, 2017].
- [9] Wolfe F, Michaud K. The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank. *Rheumatology* 2011;50:16–24.
- [10] Hernandez Alava M, Pudney S. Econometric modelling of multiple self-reports of health states: The switch from EQ-5D-3L to EQ-5D-5L in evaluating drug therapies for rheumatoid arthritis. *J Health Econ* 2017;55:139–52.
- [11] Choy EHS, Smith CM, Farewell V, et al. Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis. *Ann Rheum Dis* 2008;67:656–63.
- [12] Wailoo A, Hernandez Alava M, Scott I, et al. Cost-effectiveness of treatment strategies using combination disease-modifying anti-rheumatic drugs and glucocorticoids in early rheumatoid arthritis. *Rheumatology (Oxford)* 2014;53:1773–7.
- [13] Latimer NR, Dixon S, Palmer R. Cost-utility of self-managed computer therapy for people with aphasia. *Int J Technol Assess Health Care* 2013;29:402–9.
- [14] IMPROVE Trial Investigators. Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from IMPROVE randomised trial. *BMJ* 2014;348:f7661.
- [15] Ford HER, Marshall A, Bridgewater JA, et al. COUGAR-02 Investigators. Docetaxel versus active symptom control for refractory oesophagogastric adenocarcinoma (COUGAR-02): an open-label, phase 3 randomised controlled trial. *Lancet Oncol* 2014;15:78–86.
- [16] Hillmen P, Milligan D, Schuh A, et al. Results of the randomised phase II NCRI ARCTIC (Attenuated dose Rituximab with ChemoTherapy In CLL) trial of low dose rituximab in previously untreated CLL. *Blood* 2013;122:1639.
- [17] Howard DR, Munir T, McParland L, et al. Clinical effectiveness and cost-effectiveness results from the randomised, phase IIB trial in previously untreated patients with chronic lymphocytic leukaemia (CLL) to compare fludarabine, cyclophosphamide and rituximab (FCR) with fludarabine, cyclophosphamide, mitoxantrone and low dose rituximab (FCM-miniR): the Attenuated dose Rituximab with ChemoTherapy In CLL (ARCTIC) trial. *Health Technol Assess* 2017;21:1–374.
- [18] Blyth A, Maskrey V, Notley C, et al. Effectiveness and economic evaluation of self-help educational materials for the prevention of smoking relapse: randomised controlled trial. *Health Technol Assess* 2015;19:1–70.
- [19] Ahern AL, Aveyard PN, Halford JC, et al. Weight loss referrals for adults in primary care (WRAP): protocol for a multi-centre randomised controlled trial comparing the clinical and cost-effectiveness of primary care referral to a commercial weight loss provider for 12 weeks, referral for 52 weeks, and a brief self-help intervention [ISRCTN82857232]. *BMC Public Health* 2014;14:620.
- [20] Gershlick AH, Khan JN, Kelly DJ, et al. Randomized trial of complete versus lesion-only revascularization in patients undergoing primary percutaneous coronary intervention for STEMI and multivessel disease: the CvLPRIT trial. *J Am Coll Cardiol* 2015;65:963–72.
- [21] Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997;35:1095–108.
- [22] Devlin N, Shah K, Feng Y, et al. Valuing Health Related Quality of Life: An EQ-5D-5L Value Set for England. Technical Report 16.02. Sheffield, UK: Health Economics and Decision Science, University of Sheffield, 2016.
- [23] Mulhern B, Feng Y, Shah K, et al. Comparing the UK EQ-5D-3L and the English EQ-5D-5L value sets. OHE Research Paper 17/02. Available from: <https://www.ohe.org/publications/comparing-uk-eq-5d-3l-and-english-eq-5d-5l-value-sets>. [Accessed June 20, 2017].
- [24] Van Hout B, Janssen MF, Feng Y, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health* 2012;15:708–15.