VALUE IN HEALTH 15 (2012) 708–715

Available online at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets

Ben van Hout, PhD[1], M.F. Janssen, PhD[2], You-Shan Feng, PhD[3], Thomas Kohlmann, PhD[3], Jan Busschbach, PhD[4], Dominik Golicki, MD[5], Andrew Lloyd, PhD[6], Luciana Scalone, PhD[7,8], Paul Kind, MPhil[9], A. Simon Pickard, PhD[10,]*

[1]*University of Sheffield, Sheffield, UK;* [2]*EuroQol Group, Rotterdam, The Netherlands;* [3]*Institute for Community Medicine, University of Greifswald, Greifswald, Germany;* [4]*Medical Psychology and Psychotherapy, Erasmus University, Rotterdam, Viersprong, Halsteren, The Netherlands;* [5]*Medical University of Warsaw, Warsaw, Poland;* [6]*Oxford Outcomes, Oxford, UK;* [7]*Research Centre on Public Health, University of Milano – Bicocca, Milan, Italy;* [8]*CHARTA Foundation, Milano, Milan, Italy;* [9]*Centre for Health Economics, University of York, York, UK;* [10]*Center for Pharmacoeconomic Research, College of Pharmacy, University of Illinois at Chicago, Chicago, IL, USA*

A B S T R A C T

Background: A five-level version of the EuroQol five-dimensional (EQ-5D) descriptive system (EQ-5D-5L) has been developed, but value sets based on preferences directly elicited from representative general population samples are not yet available. The objective of this study was to develop values sets for the EQ-5D-5L by means of a mapping ("crosswalk") approach to the currently available three-level version of the EQ-5D (EQ-5D-3L) values sets. **Methods:** The EQ-5D-3L and EQ-5D-5L descriptive systems were coadministered to respondents with conditions of varying severity to ensure a broad range of levels of health across EQ-5D questionnaire dimensions. We explored four models to generate value sets for the EQ-5D-5L: linear regression, nonparametric statistics, ordered logistic regression, and item-response theory. Criteria for the preferred model included theoretical background, statistical fit, predictive power, and parsimony. **Results:** A total of 3691 respondents were included. All models had similar fit statistics. Predictive power was slightly better for the nonparametric and ordered logistic regression models. In considering all criteria, the nonparametric model was selected as most suitable for generating values for the EQ-5D-5L. **Conclusions:** The nonparametric model was preferred for its simplicity while performing similarly to the other models. Being independent of the value set that is used, it can be applied to transform any EQ-5D-3L value set into EQ-5D-5L index values. Strengths of this approach include compatibility with three-level value sets. A limitation of any crosswalk is that the range of index values is restricted to the range of the EQ-5D-3L value sets.

Keywords: EQ-5D, mapping, preference-based measures, quality of life, utilities.

Introduction

As a generic preference-based measure of health, the EQ-5D questionnaire has many applications that aid decision making in health [1]3. The standard format of the EQ-5D descriptive health classifier system developed by the EuroQoL Group consists of five dimensions of health, each with three levels of problems (EQ-5D-3L, the 3L hereon). Over the past 20 years, value sets for the 3L health classifier system have been developed for many countries around the world [2].

The EuroQol Group has recently introduced a 5-level EQ-5D questionnaire (EQ-5D-5L, the 5L hereon) that expands the range of responses to each dimension from three to five levels [3] There is an extensive literature to support the validity and reliability of the 3L in many conditions and populations [4–9] There is, however, some evidence of limited sensitivity/responsiveness of the 3L to changes in health, in part due to ceiling and floor effects [10,11] Preliminary studies indicated that a 5L version improves upon the properties of the 3L measure in terms of reduced ceiling and floor effects, increased reliability, and improved ability to discriminate between different levels of health [11–13].

Studies that directly elicit preferences from representative general population samples to derive value sets for the 5L using a harmonized protocol are under development in a number of countries. It will take time, however, for these studies to be completed and results disseminated. In the interim, the EuroQoL Group coordinated a study that coadministered both the three-level and five-level versions of the EQ-5D questionnaire to facilitate the examination of various statistical approaches to estimating value sets for the 5L. Thus, the objective of this study was to examine different approaches to deriving value sets for the 5L utilizing currently available 3L value sets and recommend a crosswalk that would generate values for the 5L.

Methods

Data

Respondents completed both the 3L and the 5L in six countries: Denmark, England, Italy, the Netherlands, Poland, and Scotland.

* *Address correspondence to:* A. Simon Pickard, Center for Pharmacoeconomics Research, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Room 164, MC 886, Chicago, IL60612.

E-mail: pickard1@uic.edu.

1098-3015/$36.00 – see front matter

doi:10.1016/j.jval.2012.02.008

The official EQ-5D-5L language version for each country was used. Different subgroups were targeted, and in most countries, a screening protocol was implemented to capture a broad spectrum of health across the EQ-5D dimensions for both the 5L and 3L descriptive systems. The screening protocol was operationalized as follows. First, conditions were identified that would provide varying levels of problems on each dimension based on existing data sets and literature (e.g., stroke and rheumatoid arthritis for problems with mobility, depression and personality disorder for problems related to anxiety/depression). Second, after data were collected from approximately 100 patients with the selected condition, the frequency distributions for each dimension were examined. If only a limited range of responses to the various levels described by each system were endorsed, a screening question was added to filter out relatively healthy patients less likely to report any problems. The severity assurance protocol was followed in all countries except Italy, which did not administer a severity screening protocol for patients with liver disease. The 5L was administered first, followed by the visual analogue scale and a number of demographic questions, and finally the 3L. A previous study showed that when respondents scored the 3L first, there was a tendency to avoid the in-between levels 2 and 4 of the 5L, and therefore all respondents scored the 5L first [11].

Measures

The 3L version of the EQ-5D questionnaire is the standard version that has been used in hundreds of clinical trials and methodological studies published in the peer-reviewed literature [1]. It is a brief self-reported measure of generic health that consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each with three levels of functioning (e.g., no problems, some problems, and extreme problems). This health state classifier can describe 243 unique health states that are often reported as vectors ranging from 11111 (full health) to 33333 (worst health). Numerous societal value sets have been derived from population-based valuation studies around the world that, when applied to the health state vector, result in a preference-based score that typically ranges from states worse than dead (<0) to 1 (full health), anchoring dead at 0. In addition, the measure includes a visual analogue scale where health is rated on a scale from 0 (worse imaginable health) to 100 (best imaginable health). In developing the 5L, the dimensional structure of the EQ-5D questionnaire was retained and descriptors for the levels of each dimension were adapted to a five-level system based on qualitative and quantitative studies conducted by the EuroQol Group [3]. The labels for the 5L followed the format "no problems," "slight problems," "moderate problems," "severe problems," and "unable to"/"extreme problems" for all dimensions. For mobility, the description of "confined to bed" has been changed to "unable to walk about." In addition, for usual activities, the word "performing" has been changed to "doing" (UK version). Pilot studies investigating different preference-based elicitation techniques are currently being conducted for the 5L system to inform large-scale international valuation studies.

Modeling approaches

Methodologically, we identified two general approaches conducive to the development of a 5L crosswalk (i.e., a mapping approach that allows 5L index values to be calculated on the basis of a link between 5L dimension responses and 3L value sets) that were based on different paradigms for health measurement. The first approach utilized what we call direct and indirect methods to estimate the relationship between the 3L data and the 5L data. The second approach used psychometric scaling techniques that assume that the 3L and 5L response categories are indicators of a common underlying construct. Specifically, the first approach uses direct methods to "transfer to utility" or indirect "response mapping" techniques [14]. The direct method employs ordinary linear regression or related statistical techniques to directly "transfer" the 5L responses to the 3L preference-based index values. The indirect method requires multinomial regression or other techniques (e.g., ordered logistic regression [OLR]) suitable for predicting categorical responses to estimate the relationship between responses to the 3L and 5L descriptive systems.

The second approach, which used psychometric scaling techniques, assumes that the 3L and 5L response categories are indicators of a common underlying construct [15]. Psychometric scaling models can then be used to analyze the association between the underlying construct and the 3L and 5L responses. Given the parameter estimates from the scaling model, an algorithm can be derived for the assignment of scores to the 3L and 5L response categories. These scores indicate how the 5L categories correspond to those in the 3L system. Any model for the scaling of categorical responses can be used for this purpose, at least in principle.

Within these two approaches, four types of statistical models were explored to develop crosswalks from the 3L to the 5L. The first set of models used the direct method of linear regression (ordinary least squares) to examine the relationship between the 5L responses and 3L index-based scores. We used the UK value set based on the Dolan et al. [16] algorithm for this purpose, because it is by far the most used and cited. Because the UK algorithms include an interaction term for any level three response ("N3"), we tested variants for the 5L by using N2, N3, N4, and N5 terms. A final variant was a model with the logarithm of the sum score of all dimensions, to capture the decrease in preference value with increased worsening of the health state.

The second model was based on the indirect mapping method, where 3L responses were predicted from 5L responses, and probabilities associated with the 3L responses were applied to their index values to obtain 5L values. Simple nonparametric calculations based on the frequencies obtained when cross-tabulating the responses on the 3L and the 5L were used, that is, the proportions of the 3L level scores within each of the five 5L levels. This so-called nonparametric model leads, for each dimension and level of the 5L, to probabilities of being in each of the 3L levels. For each health state described by the 5L system ($n = 3125$), the probability of reporting each of the 243 3L health states was determined by taking the product of the corresponding probabilities. For instance, a respondent reporting the 5L health state vector 23245 and 12123 on the 3L system is the product of

1. the probability of level 1 on 3L-mobility given level 2 on 5L-mobility;
2. the probability of level 2 on 3L-self-care given level 3 on 5L-self-care;
3. the probability of level 1 on 3L-usual activities given level 2 on 5L-usual activities;
4. the probability of level 2 on 3L-pain/discomfort given level 4 on 5L-pain/discomfort;
5. the probability of level 3 on anxiety/depression given level 5 on 5L-anxiety/depression.

In total, 243 transition probabilities are generated. Note that in this model we did not allow for interaction between the dimensions. The 5L index value is then calculated by multiplying the 243 transition probabilities by their corresponding 3L index values, and subsequently summing them. This can be done for each 5L health state linked with each 3L health state. In this way, a 3125×243 matrix of transition probabilities was created.

Table 1 – Respondent characteristics.

Country	Population	n	% Female	Mean age (y)	Mean VAS (SD)	Mean EQ-5D-3L index value (SD)*
Denmark	Diabetes	230	46	52.4	75 (20)	0.78 (0.24)
	Orthopedic accident	94	34	37.8	79 (23)	0.63 (0.42)
	Rheumatoid arthritis	35	73	60.5	60 (25)	0.51 (0.32)
England	ADHD	69	54	34.3	63 (21)	0.59 (0.33)
	Arthritis	250	44	57.7	66 (20)	0.64 (0.23)
	Back pain	70	57	47.2	52 (19)	0.47 (0.28)
	COPD	125	37	60.8	57 (21)	0.56 (0.30)
	Depression	250	56	42.4	62 (21)	0.64 (0.30)
	Diabetes	45	58	50.8	69 (20)	0.72 (0.25)
	Myocardial infarction	75	27	56.7	63 (20)	0.64 (0.28)
	Parkinson's disease	32	44	49.8	66 (22)	0.46 (0.43)
	Stroke	85	39	57.4	53 (24)	0.52 (0.29)
Italy	Liver disease	426	31	56.0	70 (20)	0.80 (0.23)
Netherlands	Kidney dialysis	49	41	61.7	62 (21)	0.60 (0.37)
	Personality disorders	384	67	31.7	59 (18)	0.61 (0.27)
Poland	Stroke	529	49	69.9	52 (26)	0.38 (0.41)
	Student population	443	79	22.1	79 (16)	0.87 (0.14)
Scotland	Asthma	21	57	72.8	64 (18)	0.64 (0.24)
	Cardiovascular disease	176	54	71.4	60 (21)	0.54 (0.33)
	COPD	196	62	70.1	58 (21)	0.53 (0.34)
	Multiple sclerosis	15	53	63.9	52 (21)	0.47 (0.37)
	Parkinson's disease	5	60	63.0	41 (30)	0.25 (0.43)
	Rheumatoid arthritis	87	71	69.4	56 (22)	0.48 (0.34)
Overall		3691	53	51.5	64 (23)	0.62 (0.33)

ADHD, attention deficit/hyperactivity disease; COPD, chronic obstructive pulmonary disease; EQ-5D-3L, three-level version of the EuroQol five-dimensional questionnaire; VAS, visual analogue scale.
* Values based on UK value set [16].

This technique of calculating 5L values as a summation of 243 products of transition probabilities with 3L index values was also followed (as the final step) in the third and fourth models.

The third model, another instance of the indirect method, estimated transition probabilities by using a logistic regression model for ordered categories. OLR is an extension of standard lo-

Table 2 – Cross tabulation for EQ-5D-3L and EQ-5D-5L responses by dimension (consistent data set).

EQ-5D-3L	EQ-5D-5L				
	No problems	Slight problems	Moderate problems	Severe problems	Unable to
Mobility					
No problems	1782	119	16	1	4
Some problems	29	552	586	386	23
Confined to bed	1	1	4	30	112
Self-care					
No problems	2468	82	13	5	0
Some problems	43	408	313	109	6
Unable to	3	5	6	35	140
Usual activities					
No problems	1382	163	20	9	0
Some problems	42	661	656	274	15
Unable to	5	7	23	134	239
Pain/discomfort					
None	1126	211	21	6	2
Moderate	65	850	837	239	8
Extreme	1	4	19	159	82
Anxiety/depression					
None	1352	219	30	10	3
Moderate	45	841	692	164	6
Extreme	1	3	17	158	93

EQ-5D-3L, three-level version of the EuroQol five-dimensional questionnaire; EQ-5D-5L, five-level version of the EuroQol five-dimensional questionnaire.

Table 3 – In-sample (fit) and out-of-sample prediction (predictive power) for crosswalk methods (mean square error)*.

Data set/cohort	n	In-sample (fit)†					
		Direct method			Indirect method		
		Linear	Linear + log(sum)	Linear + N4 + N5	Nonparametric	OLR	OLR + interaction
Pooled	3691	0.015	0.014	0.014	0.014	0.014	0.013
Without COPD/asthma	3349	0.014	0.013	0.013	0.013	0.013	0.013
Without diabetes	3416	0.015	0.015	0.014	0.014	0.014	0.014
Without liver disease	3265	0.016	0.015	0.015	0.015	0.015	0.014
Without RA/arthritis	3319	0.014	0.013	0.013	0.013	0.013	0.013
Without cardiovascular disease	3440	0.014	0.014	0.013	0.013	0.013	0.013
Without stroke	3077	0.013	0.013	0.013	0.013	0.013	0.013
Without depression	3441	0.014	0.014	0.014	0.013	0.013	0.013
Without personality disorders	3307	0.014	0.013	0.013	0.013	0.013	0.013
Without students	3248	0.016	0.015	0.015	0.014	0.014	0.014
Data set/cohort	n	Out-of-sample (predictive power)‡					
		Direct method			Indirect method		
		Linear	Linear + log(sum)	Linear + N4 + N5	Nonparametric	OLR	OLR + interaction
COPD/asthma	342	0.021	0.021	0.020	0.020	0.020	0.020
Diabetes	275	0.008	0.007	0.007	0.007	0.007	0.008
Liver disease	426	0.006	0.006	0.006	0.005	0.005	0.006
RA/arthritis	372	0.019	0.019	0.020	0.018	0.018	0.018
Cardiovascular disease	251	0.020	0.019	0.019	0.019	0.019	0.018
Stroke	614	0.028	0.024	0.022	0.017	0.017	0.017
Depression	250	0.017	0.017	0.016	0.016	0.016	0.016
Personality disorders	384	0.024	0.021	0.023	0.021	0.021	0.023
Students	443	0.007	0.007	0.007	0.007	0.007	0.007
Mean		0.017	0.016	0.016	0.014	0.014	0.015

COPD, chronic obstructive pulmonary disease; OLR, ordered logistic regression; RA, rheumatic arthritis.

* The dependent variable for all direct models was the UK index value: independent variables for the linear model were the scores on the 5L dimensions (20 dummy variables for each of the levels on each of the dimensions indicating problems); for the linear plus log(sum) model, a variable was added that was the logarithm of the sum score of all dimensions; and for the linear + N4 + N5, two dummy variables were added that indicated any problems on level 4 (N4) or level 5 (N5) on any dimension. The dependent variables were the scores on the 3L dimensions for the indirect models: for the nonparametric and OLR models, the independent variables were the identical dimension scores on 5L (four dummy variables per dimension indicating problems), and for the OLR plus interaction model, the other 5L dimension scores were added (coded as 1, 2, and 3).

† All in-sample predictions were based on the consistent data set.

‡ The models for the nine population groups were based on the consistent data set. Out-of-sample predictions were based on the remaining data set including inconsistencies.

gistic regression in which the probability of a particular health state has a logistic link function to a set of explanatory variables. It is a special form of multinomial logistic regression in which the coefficients in the prediction function are identical for all categories of the dependent variable (which seems likely in this case). A variation of the OLR model was also explored that included interaction terms for the other dimensions.

The fourth model, based on the psychometric scaling approach, was an indirect method to obtaining values for the 5L. The partial credit model, an item-response theory (IRT)-based model, was used to define an underlying construct for each dimension as measured by the 3L and 5L systems [17,18]. Probabilities of response patterns are estimated along a continuous underlying variable for each pair of 3L and 5L items. Using this model, category-specific average person parameters are calculated and used to estimate the 5L index values according to an algorithm. This method has been previously explored as a methodological approach to deriving a crosswalk between the 3L and an experimental version of the 5L [19]. The model assumes the probability of responses to be normally distributed. So, for each score on the underlying variable there is a probability to be in one of the 3L states and in one of the 5L states. By integration over this underlying variable, estimates are obtained of the probability to be in any of the 3L scores given the 5L score. Finally, the technique of summating the 243 resulting products of transition probabilities with their corresponding 3L values was applied to calculate the 5L values.

Inconsistencies

An important issue we needed to resolve was the tension between using all data or to restrict the analysis to logically consistent responses. An example of a logical inconsistency would be a respondent who reports level 1 (no problems) on the 5L and level 3 (extreme problems) on the 3L. While such responses could be assumed random error, it was debatable whether to include them given decision rules can be implemented to identify inconsistent responses.

Problematically for developing a crosswalk, the value for 11111 on the 5L might be lower than 1 when including these responses,

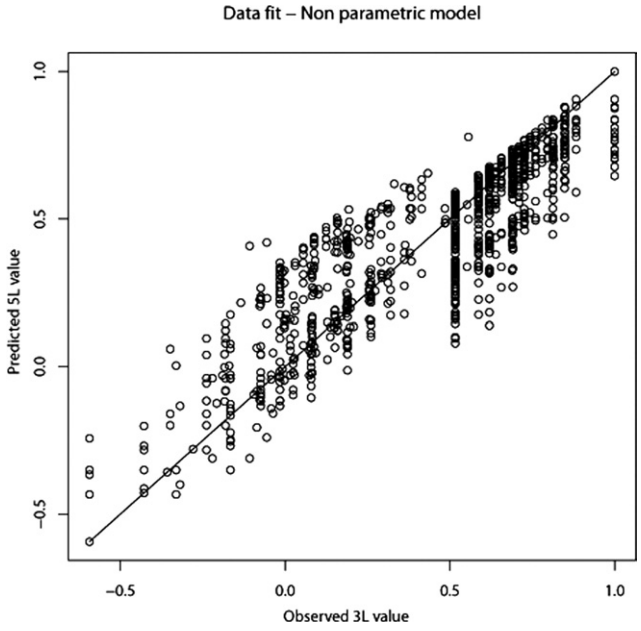

Fig. 1 – Data fit for final model.

counterintuitively truncating the range of values possible for the 5L system to less than the range of 3L values. For these reasons, we conducted analyses on the full data as well as excluded inconsistent responses. We then chose to exclude "inconsistent" responses to create a so-called consistent data set. The consistent data set was derived from logic rules intended to reduce the number of responses in crosswalk that appeared to be illogical response combinations to the 3L and the 5L. We defined all responses to be "inconsistent" when a 3L response corresponded to a 5L response that was two, three, or four levels away (e.g., 1 on 3L with 3 on 5L; or 2 on 3L with 1 on 5L).

Model selection

We applied four criteria to assess the performance of each of the models to recommend a preferred approach. First, the theoretical background of the various models was considered. There are some limitations to the direct and indirect methods that are known in advance of comparing their statistical performance. Indirect methods lead to a solution that is independent of the value set used, which is advantageous in that direct methods need completely new link functions for each value set. Only the weighted averages over the 243 states for each 5L value have to be recalculated when applying a new value set. Furthermore, the indirect method is modeling upon response behavior and therefore more closely follows the dimensional structure of the EQ-5D questionnaire.

The second and third criteria are statistical in nature: in-sample prediction (fit) and out-of-sample prediction (predictive power). Each model predicts 5L index values that can be compared to the observed 3L values. Here, fit was measured as the mean squared error (MSE) of the models on the (in-sample) pooled consistent data set. Predictive power was measured as the MSE of a number of out-of-sample predictions by using the following strategy. The data set was categorized into nine population subgroups, and the values resulting from the models within each population group were used to predict the values for the remainder of the data (out-of-sample). Inconsistencies were not excluded from the predictive samples (out-of-sample) when applying this approach. The fourth criterion was parsimony, which for our purposes was the model that was the least complicated and invoked the fewest assumptions when two approaches performed similarly.

A final consideration relates to a large gap in values between full health (11111) and the second best health state, a known criticism of the UK value set for the 3L. For the UK value set, this gap is 0.117 (1 minus the value for health state 11211, which is 0.883). We were interested in the extent to which each model reduced this gap in values using the 5L.

Results

In total, 3691 respondents completed both the 3L and the 5L. The overall cohort was 53% female and had a mean age of 51.5 ± 20 years. A mean (SD) visual analogue scale score of 64 (23) was observed, ranging from 41 (30) for Parkinson's disease to 79 (16) for the student sample. Mean (SD) index-based values were 0.62 (0.33), ranging from 0.25 (0.43) for Parkinson's disease to 0.87 (0.14) for the student population. For the purposes of modeling, respondents were classified into nine subgroups: chronic obstructive pulmonary disease/asthma (n = 342), diabetes (n = 275), liver disease (n = 426), rheumatoid arthritis/arthritis (n = 372), cardiovascular disease (n = 251), stroke (n = 614), depression (n = 250), personality disorders (n = 384), and students (n = 443) (Table 1).

The number of missing values ranged from 26 (0.70%) on self-care (5L) to 45 (1.22%) on pain/discomfort (3L). A total of 522 inconsistencies were found, distributed across 426 respondents. Cross-tabulations of responses to the 3L and the 5L, resulting from the full sample (including inconsistent responses), show that a broad spectrum of levels of health on each dimension was reported by the participants (Table 2).

The in-sample prediction (fit) and out-of-sample prediction (predictive power) produced similar results across the various models (Table 3). Results ranged from an MSE of 0.013 for OLR plus interaction to 0.015 for the linear model. Generally, the indirect methods performed slightly better than the direct methods. There was considerable variation across subsamples, from an MSE of 0.007 for the student sample (all models) to 0.028 for respondents with stroke (linear model: Table 3, bottom). Note that the IRT model could not be performed on the consistent data set. However, the IRT-based model performed equally well compared with other models when using the full data set (data not shown). Note that little is gained by allowing interactions between dimensions in the OLR model.

Plots of observed (3L) and predicted (5L) values based upon the linear and nonparametric models are shown in Figure 1. Figure 1 illustrates that the 5L values based on the models tended to underpredict 3L observed values on the upper end of the scale and overpredict values on the lower end of the scale. For the results of the OLR model, responses to each dimension on the 5L appear as bar graphs on the x-axis, which represents the level of severity of the trait/dimension (x^1b), as shown in Figure 2. The probability of endorsing level 1 (black line), level 2 (green line), or level 3 (red line) on the 3L system for a given level of the trait is represented by the three lines. As shown in Figure 2, the probability of endorsing level 3 in the 3L system is always lower than the probability of endorsing level 5 on the 5L system. Alternatively stated, Figure 2 illustrates that level 5 on the 5L system represents more extreme health problems than does level 3 on the 3L system, and conversely that level 1 on the 5L system is healthier than level 1 on the 3L system.

The gap between full health (11111) and the next best health state was reduced to the greatest extent when using the linear model. The reduction was 0.038; 0.049 with any level 4 (N4) and/or any level 5 (N5) included and 0.043 when including the logarithm of the summed score. In contrast, the gap was reduced by only 0.022 when using the OLR model (0.030 with interactions terms) and by 0.023 when using the nonparametric model.

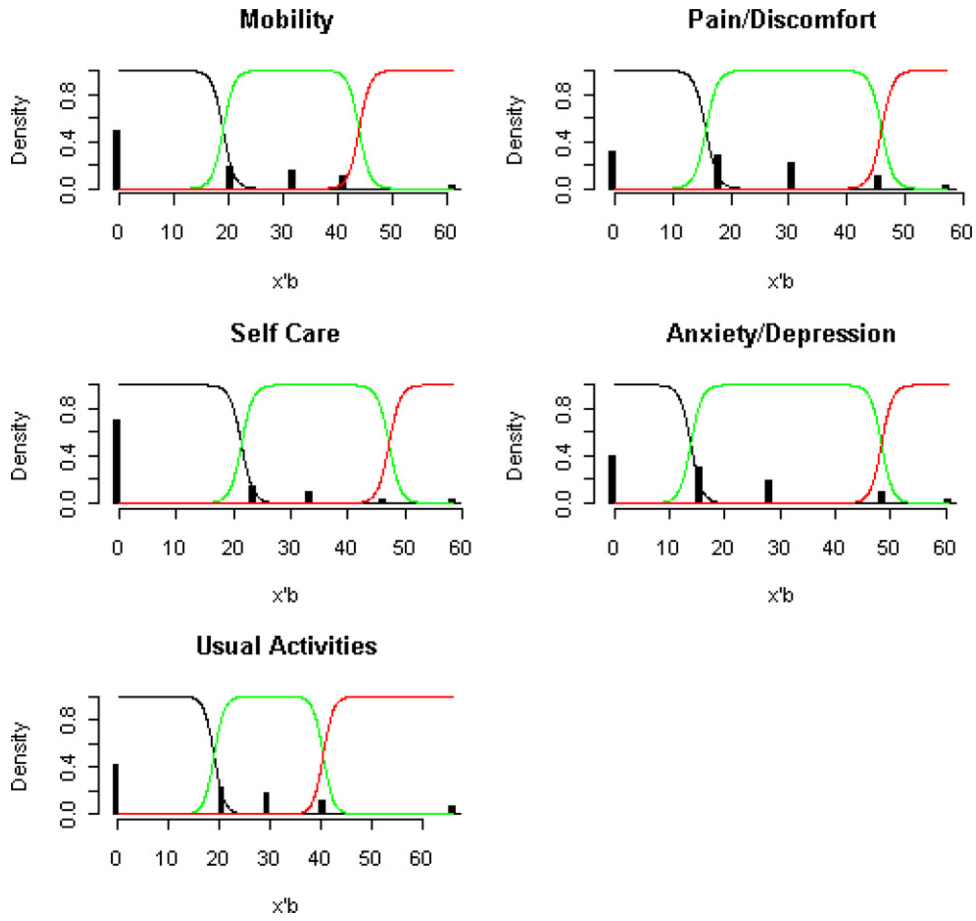

Fig. 2 – Fit for ordered logistic regression (OLR) by dimension (consistent data set). Black line: Probability of level 1; green line: probability of level 2; red line: probability of level 3; x1b: level of trait.

In regard to fit and predictive power, all models produced similar results. When considering theoretical background, indirect methods are preferred because the resulting models are independent of the value set used. Following the final criterion of parsimony, the nonparametric indirect mapping model was recommended for obtaining 5L values.

Because each 5L value for the nonparametric model was based on a summation of 243 products of transition probabilities with their corresponding 3L index values, we cannot show direct parameter estimates for the final 5L model, as was the case for the 3L value sets. To give an example of the actual 5L values for the final model, Table 4 shows mean observed 3L values with standard errors and 5L index values based on the nonparametric model for a selection of the most frequently occurring health states (UK value set).

Discussion

The objective of this study was to explore various methods that could be used to estimate value sets for health states defined by the EQ-5D-5L and to recommend a specific crosswalk. We employed criteria that are often used in studies that seek to estimate values or utilities for health-related quality-of-life measures, including the theoretical basis, model fit, predictive power, and parsimony/simplicity. The various approaches produced similar results on several of the criteria, and ultimately we preferred the intuitive appeal and transparency of the nonparametric model, and importantly, its "value set free" ability to estimate 5L value sets by using any 3L value set. While direct linear regression estimation using index values for EQ-5D-3L health states has the advantage of being technically simple, it is value set dependent. In contrast, the indirect method seeks an association between the two health state classification systems and yields solutions that are structurally independent of the value sets used to compute index values. This approach has been applied previously to build a crosswalk between the EQ-5D-3L and the short-form-12 item questionnaire (SF-12), although in that study all dimensions of the EQ-5D questionnaire and all items of the SF-12 were mapped [20].

In recent years, in the absence of value sets directly elicited from large samples representative of the general population, various disease-specific and generic measures have mapped descriptive systems onto established utility-based generic measures such as the EQ-5D questionnaire [21]. One of the major limitations of mapping items from one measure to another to estimate a utility-based summary score is the difference in content coverage. In this respect, the present study is well suited to a mapping approach because the dimensions of the EQ-5D-3L and the EQ-5D-5L are identical.

In selecting the "best" model there are many criteria that could be adopted. All other things being equal, the criterion of parsimony is a guiding principle in the sense that it enhances transparency and aids in the interpretation of scores. In this respect, the nonparametric model appears to be the most suitable approach, because it is easy to operationalize and produced prediction errors that hardly differed from the other models. When considering theoretical rigor, the OLR model was desirable in the sense that com-

Table 4 – Mean observed 3L and 5L index values based on the nonparametric model for frequently occurring health states (UK value set).

Health state	n	Observed 3L value	SE	5L index value	MSE
11112	209	0.890	0.006	0.879	0.007
11113	58	0.844	0.009	0.848	0.005
11121	143	0.840	0.010	0.837	0.014
11122	138	0.782	0.008	0.767	0.009
11123	54	0.769	0.013	0.749	0.009
11131	28	0.796	0.009	0.796	0.002
11211	28	0.907	0.016	0.906	0.007
11212	35	0.818	0.011	0.837	0.004
11213	23	0.826	0.010	0.819	0.002
11221	41	0.817	0.012	0.795	0.006
11222	67	0.730	0.008	0.736	0.005
11223	28	0.726	0.010	0.721	0.003
11324	21	0.428	0.051	0.501	0.057
21111	28	0.911	0.020	0.877	0.012
21121	50	0.765	0.011	0.767	0.006
21122	24	0.734	0.019	0.708	0.009
21221	55	0.723	0.008	0.735	0.005
21222	48	0.642	0.017	0.679	0.015
21231	28	0.707	0.007	0.710	0.001
21232	27	0.626	0.009	0.654	0.003
22222	14	0.578	0.030	0.592	0.012
22332	13	0.540	0.019	0.560	0.003
31333	17	0.614	0.006	0.620	0.001
32331	17	0.599	0.008	0.604	0.001
33333	15	0.509	0.025	0.516	0.009
43433	14	0.383	0.069	0.378	0.061
43443	17	0.267	0.067	0.206	0.076
55544	15	−0.337	0.044	−0.352	0.027

MSE, mean squared error; SE, standard error.

plementary dimensions were taken into account and it also provided good predictions. The IRT model may be the most elegant model with its acknowledgment of the latent and continuous scale underlying each dimension, and provides a rich source of information about the strengths and weaknesses of the descriptive systems that are insightful but not directly relevant to the goals of this article. IRT-based models were incompatible with the consistent data set because the model identifies parameters based on variation in responses. By excluding the variation, the parameters that distinguish between the likelihood to be in one state or another cannot be estimated.

We selected the nonparametric indirect model because it was simple, demonstrated fit statistics, and had predictive power comparable to that of the more complex models. Brazier et al. [21] conducted a comprehensive review of mapping studies and reported a range of root mean square error from 0.084 to 0.2 for a total of 119 within-sample models over 30 studies. Our MSE for the nonparametric model of 0.014 equals to a root mean square error of 0.12, which lies in the lower half of the reported range. Although we illustrated our results by using the UK value set [16], value sets were calculated for many country-specific value sets. These 5L value sets can be obtained from the EuroQol Web site at www.euroqol.org along with an Excel file that enables users to easily calculate the 5L index values from their 5L dimension scores.

This study has several limitations, some of which are common to mapping studies. First, mapping is data dependent, and so the selection of respondents can influence the calibration of values. For this reason, the data collection phase was designed to facilitate a wide range of levels of health across the different dimensions on the EQ-5D questionnaire in a large number of respondents from different counties. A second limitation relates to restrictions on the range of scale possible for 5L values when mapping to 3L value sets. Specifically, respondents who categorized themselves as 55555 when using the 5L can report no worse than 33333 when using the 3L, yet it is possible for them to report 23333 without being classified as inconsistent. For this reason, a crosswalk-based approach limits the value of 55555 to be no lower than that of 33333. This limitation places an artificial floor effect on the values of the 5L that contrasts with research showing that a five-level system actually broadens the measurement continuum and would be expected to result in lower values when compared with a three-level system [13,19]. The decision to base the crosswalk on the consistent data set utilized decision rules to minimize the influence of illogical response combinations that, because of the weights contributed by those responses, would mitigate the benefits of a scale based on a 5L system. A third limitation is that 3L and 5L dimension scores were pooled from various countries, using different translations of the 5L descriptive system. There might have been cultural differences in how respondents from the various countries interpret the different 5L translations. The only way to deal with this problem would be to develop a crosswalk for each country separately. This was deemed unfeasible because of budget and time constraints. Furthermore, intercountry results from the United Kingdom and Spain showed that the 5L labels performed substantially similarly on the response scaling task [3]. A final limitation is that there might have been an ordering effect by always presenting the 5L first.

In the near future, valuation studies will be carried out to obtain direct valuations for the new EQ-5D-5L, which should address the limitations mentioned above. In absence of those valuation studies, scores for the 5L can be obtained by using the approach recommended in the present study. While there are limitations to the crosswalk-based approach, a notable strength of the recommended crosswalk is the ability to apply it to all existing 3L value sets. In addition, it has the advantage of compatibility with past scoring approaches to the 3L in the sense that no other aspects of the protocol for eliciting utilities have been modified.

Acknowledgments

The authors thank Nancy Devlin, Paul Swinburn, and Maciej Niewada for their contributions to the study implementation. Views expressed in the article are those of the authors alone.

Source of financial support: This research was supported in part by the EuroQoL Group. Data collection in England was funded by the Department of Health Policy Research Programme grant PRP 070-0065. Data collection in Italy was funded by the Center for Health Associated Research and Technology Assessment Foundation and supported by the Italian hepatitis patients' organization EpaC Onlus.

REFERENCES

[1] Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001;33:337–43.
[2] Szende A, Oppe M, Devlin NJ. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Dordrecht, The Netherlands: Springer, 2007.
[3] 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.
[4] Pickard AS, Wilke CT, Lin HW, et al. Health utilities using the EQ-5D in studies of cancer. Pharmacoeconomics 2007;25:365–84.
[5] Janssen MF, Lubetkin EI, Sekhobo JP, et al. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med 2011;28:395–413.
[6] Pickard AS, Wilke C, Jung E, et al. Use of a preference-based measure of health (EQ-5D) in COPD and asthma. Respir Med 2008;102:519–36.

[7] Dyer MT, Goldsmith KA, Sharples LS, et al. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes 2010;8:1–13.
[8] Johnson JA, Pickard AS. Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada. Med Care 2000;38:115–21.
[9] Johnson JA, Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. Qual Life Res 1998;7:155–66.
[10] Pickard AS, De Leon MC, Kohlmann T, et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care 2007;45:259–63.
[11] Janssen MF, Birnie E, Haagsma JA, Bonsel GJ. Comparing the standard EQ-5D three-level system with a five-level version. Value Health 2008; 11:275–84.
[12] Pickard AS, De Leon MC, Kohlmann T, et al. Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Med Care 2007;45:259–63.
[13] Janssen MF, Birnie E, Bonsel GJ. Quantification of the level descriptors for the standard EQ-5D three-level system and a five-level version according to two methods. Qual Life Res 2008;17:463–73.
[14] Mortimer D, Segal L. Comparing the incomparable? A systematic review of competing techniques for converting descriptive measures of health status into QALY-weights. Med Decis Making 2008;28:66–89.
[15] Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38(9, Suppl.):II28–42.
[16] Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095–108.
[17] Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
[18] Muraki E. A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 1992;16:159–76.
[19] Pickard AS, Kohlmann T, Janssen MF, et al. Evaluating equivalency between response systems: application of the Rasch model to a 3-level and 5-level EQ-5D. Med Care 2007;45:812–9.
[20] Gray AM, Rivero-Arias O, Clarke PM. Estimating the association between SF-12 responses and EQ-5D utility values by response mapping. Med Decis Making 2006;26:18–29.
[21] Brazier JE, Yang Y, Tsuchiya A, et al. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010;11:215–25.