The University Of Sheffield.

UNIVERSITY of York

EEPRU Policy Research Unit in Economic Methods of Evaluation in Health & Social Care Interventions

Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study.

30th September 2020

FINAL REPORT

Authors: Mónica Hernández Alava, Steve Pudney, Allan Wailoo

Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, UK

FUNDED BY NIHR | National Institute for Health Research

Funding

This research is funded by the National Institute for Health Research (NIHR) Policy Research Programme, conducted through the Policy Research Unit in Economic Methods of Evaluation in Health and Social Care Interventions, PR-PRU-1217-20401.

Acknowledgements

The views expressed are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or its arm's length bodies, or other UK government departments. Any errors are the responsibility of the authors. We would like to thank Fred Wolfe and Kaleb Michaud of FORWARD, the National Databank for Rheumatic Diseases, for continued permission to use their data. The EuroQoL group also kindly provided data for analysis. We would like to acknowledge the assistance of Donna Davis and Liz McLintock with the production of this report. Helpful comments on previous versions of this report were received from Laura Gray and Alan Lamb.

Contents

1. Introduction

There are two versions of the EQ-5D for measuring and valuing health in adults in the UK. The descriptive system for both comprises a classification that allows respondents to indicate their health state on five dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression. In the EQ-5D-3L version (3L from here), respondents indicate the degree of impairment on each dimension according to three levels (no problems, some problems, extreme problems). The three-level version of EQ-5D (3L) is the most widely used preference based measure in economic evaluations across the National Institute for Health and Care Excellence's (NICE's) guideline producing programmes. It is currently explicitly recommended for reference case analyses in the Guide to the Methods of Technology Appraisal (Section 5.3).[1]

A newer, five-level version of the instrument has been developed. EQ-5D-5L (5L from this point) includes five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems). There is no currently approved value set for the 5L in England. However, the 5L descriptive system is increasingly used in clinical studies. Therefore, there is a requirement to estimate health state utility values for the 5L system based on the 3L value set via "mapping".

We know that 3L and 5L cannot be treated as if they were equivalent[2] [3]. Therefore, some means of linking from responses individuals give to the 5L to the responses we would have seen had they filled in the 3L and the associated values currently needs to be used. In future, the ability to link from older 3L evidence to 5L will be required, should an acceptable 5L value set for England be produced.

Van Hout et al[4] provide an option for mapping, which we refer to henceforth as the "Crosswalk". It is the approach recommended in the NICE 2013 Methods Guide. Van Hout et al estimate 3L from 5L responses using a series of modified cross-tabulations of responses to the 3L and 5L instruments, for each dimension of health separately. The approach is based on data provided by the EuroQol Group (EQG).

Alternative mappings have been derived from a modelling approach developed by Hernández Alava and Pudney[5] for the NICE Decision Support Unit (DSU), using a multi-equation statistical model of the joint 3L and 5L responses. Originally estimated using a sample of patients with rheumatoid arthritis from FORWARD, the National Data Bank for Rheumatic Diseases, the model was subsequently re-estimated using the same EQG dataset as van Hout et al. (see Wailoo et al[6]). Results from the model have been made available to cost effectiveness analysts through pre-programmed commands for popular software including Stata, R and Excel. Comparisons between both the methods and results of the Crosswalk and DSU approaches were reported in an earlier DSU report.[7]

We now report on an updated analysis for mapping from 5L to 3L, using the same methods developed by Hernández Alava and Pudney for the DSU but based on a new dataset. It is important to note that these

methods treat the 3L and 5L responses symmetrically and therefore provide the added functionality of mapping from 3L to 5L, which may be an important future option should a suitable 5L value set be developed for use in England. The approach has the advantage that mappings from 3L to 5L and 5L to 3L are logically consistent, which will not generally be true if separately-derived 3L→5L and 5L→3L mappings are used.[8]

It has previously been noted that both the FORWARD and EQG datasets have significant limitations for the purposes of mapping between the two variants of EQ-5D.[7] Because of this, the updated analysis is based on a new data collection exercise designed to overcome acknowledged limitations in existing mapping datasets. When comparing the new dataset with existing ones, our main focus is on the comparison with the EQG dataset, since the FORWARD alternative is confined to a single disease area.

We first set out the features and limitations of the EQG dataset, followed by a description of the new data collection exercise (described here as the EEPRU dataset). We provide key details of the respondent sample and compare the basic features of the EQG and EEPRU data.

Second, we report the results of the updated mapping model, estimated from the EEPRU data.

Third, we compare various aspects of fit for the updated model (and the earlier FORWARD and EQG-based versions), compared to the Crosswalk.

2. Methods

2.1. The EQG dataset

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 3L and 5L versions of EQ-5D, to compare them in terms of their measurement properties and to generate an interim value set for 5L using a mapping (or cross-walk) approach. Published papers provide little detail of the EQG dataset, in part because it was formed as a combination of add-ons to several separate studies in different countries (see Janssen et al[9] and van Hout et al[4]). The questionnaire introduced the 5 level version of EQ-5D first, followed by a few background questions (age, gender, education, etc), then the 3 level version of EQ-5D, the EQ-5D visual analogue scale, a set of five dimension-specific rating scales and finally the WHO (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/centres and patient recruitment agencies. All countries used paper and pencil questionnaires, except England, which used an online version. In all countries except Italy, a screening protocol was used to ensure a wide range of severity across all the 5L and 3L dimensions.

The new data collection exercise (referred to here as the EEPRU study) was motivated by four limitations of the EQG survey for the purpose of mapping in an English NHS context. (Most, but not all, of these drawbacks apply equally to the FORWARD dataset. For example, there is a very large degree of separation between the 3L and 5L instruments in the FORWARD survey).

1. *Sample size and coverage*

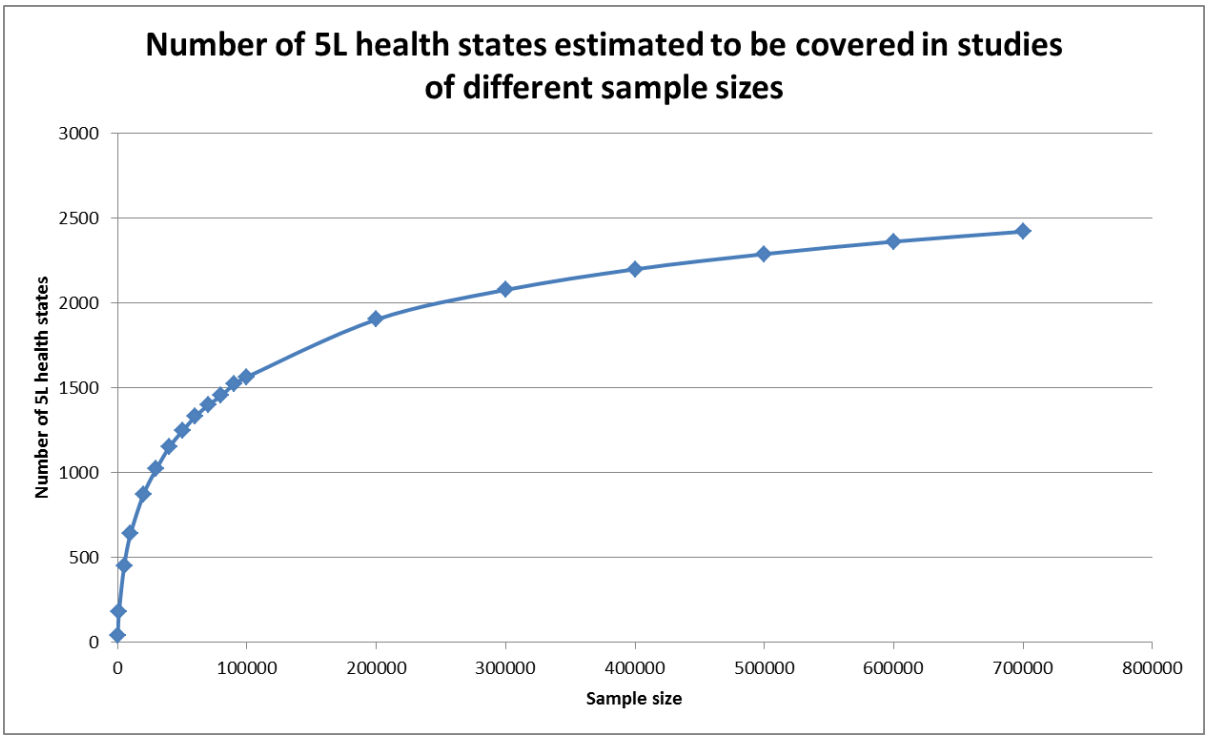
The EQG study provides 3691 responses. Relative to the numbers of combinations of 3L and 5L, this is small and means that any mapping model will inevitably require a proportionally large amount of extrapolation beyond what has been observed.

The 3L instrument describes $3^5 = 243$ logically possible health states, while the 5L instrument describes $5^5 = 3{,}125$ possible states. There are therefore $15^5 = 759{,}375$ possible 3L-5L combinations. Although we might expect a significant proportion of those combinations to be rare in practice, it is clear that a large sample survey is required to adequately represent the range of combinations likely to be encountered in practice.

To decide on the appropriate sample size, we used an existing large population survey to examine the proportion of the 3,125 5L combinations that are reported in practice. The General Practitioner Patient Survey (GPPS) is a large scale cross sectional study supported by NHS England, which records the 5L version of EQ-5D. We obtained data from 2012 to 2015 and used all observations with a complete 5L response (n = 792,571), yielding observations of 2,464 of the 3,125 possible health states. We drew random

subsamples (without replacement) of varying sizes from the data, with 10 repetitions each. The mean number of distinct 5L health states by sample size is displayed in Figure 1. The clear conclusion is that a large sample size is required for good coverage of 5L health states – for example, a sample size of 5,000 would be expected to observe only 450 5L health states, 18% of those observed in the whole GPPS sample.

Figure 1: Coverage of 5L health states and sample size



2. *Ambiguity and comparability of language*

The EQG study was international, with several components conducted in non-English speaking countries, and consequently may not be generalisable to the English NHS setting. This is a particular concern in relation to the labels used for levels 4 and 5 of the 5L instrument, where there may be ambiguity in the minds of some respondents about the ordering of "severe problems" (level 4) and "extreme problems" (level 5). The impact of this potential ambiguity may be less in the context of mapping compared to valuation studies, since the descriptions are presented to respondents in the intended order of severity. Nevertheless, language differences both here and in any other area may influence the observed relationship between 3L and 5L responses. This raises doubts about the validity of data for use in the UK setting but collected using other language variants of EQ-5D where the potential for ambiguity of "severe" vs "extreme" may be different or even non-existent, or other language impacts may be relevant.

3. *Placement within questionnaire*

The premise underpinning mapping is that the responses to each instrument are independent of the presence of the other in the same survey. The validity of this assumption is likely to depend on the degree of separation within the survey of the two instruments, and possibly also on the mode of administration. Responses to the 3L version encountered later in the EQG questionnaire may be contaminated in some

way by recollection of responses to the earlier 5L version if the two instruments are nearly contiguous. The degree of separation between the 5L and 3L instruments in the EQG survey was relatively small so there is a risk that 3L responses may be distorted (relative to what would be observed in a 3L-only survey).

4. *Question ordering*

The EQG survey presented all respondents with the 5L instrument before they encountered the 3L instrument. It has been established in a randomised experiment[10], undertaken as a pilot specifically to inform survey design for mapping, that the ordering of the two variants can have a material influence on the responses that are given.

2.2. New data collection

A new data collection exercise was undertaken designed to overcome some of the limitations of the EQG dataset. The first objective was to achieve a much larger sample – the study was designed to achieve 50,000 useable responses. Referring to Figure 1 above, it was envisaged that this would improve the coverage of 5L health states to more than half the number yielded by the GPPS.

Second, the study was conducted entirely in English from samples of respondents in the UK.

Third, the degree of separation between the two variants was maximised within the constraints of the survey instrument. Respondents would see variant 1 (3L or 5L) followed by the EuroQoL Visual Analogue Scale (VAS), then a series of questions (age group, sex, family circumstances, educational achievement, existing medical conditions, use of medication, caring responsibilities, life and health satisfaction), followed by the second variant (5L or 3L) and a repeat of the VAS instrument. The inclusion of the VAS and its repetition was a requirement of the EuroQoL group.

Fourth, the ordering of 3L/5L was randomised, so that sequencing effects could be assessed and neutralised as far as possible.

The survey was conducted online, following piloting in the *Understanding Society* "Innovation Panel Study[10]" (IP) which used a mixture of web interview (CAWI) and computer-assisted self-interviewing (CASI) in a sample of almost 3,000 individuals. Building on that independent pilot, data was collected by OnePoll using existing UK polling panels during April 2020, and anonymised responses were provided to the research team. Ethics approval for the study was granted by the University of Sheffield.

Members of the OnePoll panel are typically highly engaged. They are paid to complete surveys and are removed if there is evidence that they are not providing considered responses. There were no responses that we were prepared to deem logically inconsistent: we did not impose any requirement on how the 3L and 5L instruments could be answered. Therefore, no responses were ruled out as invalid in our survey. .

Whilst the OnePoll panel is generally considered to be representative of the UK population in many measurable aspects, it should be noted that we were not seeking a representative sample of UK opinions. Rather, the concern was to obtain sufficient observations across the EQ-5D severity range. Oversampling

of those in ill health was considered, but the preferred option was to use a sample from the overall panel that would be large enough to cover an appropriately wide range of health states.

2.3. Modelling methods

Full details of the modelling methods are reported in Hernández Alava and Pudney[5]. The approach is based on a joint statistical model of the ten EQ-5D responses (five at 5L, five at 3L), using a multi-equation ordinal regression framework. Three extensions were used to enhance the flexibility of the model: a copula specification to allow differing degrees of correlation between the 3L and 5L responses at the upper and lower extremes of health; a normal mixture residual distribution to give flexibility in the distributional form of responses; and a common factor to capture correlations in responses across the five dimensions of EQ-5D. Given the very large sample size, estimation of these complex models is computationally intensive and time-consuming. The tools we provide to allow analysts to implement the results from these models are simple to use.

The large sample with a high proportion of healthy individuals could prompt further extensions of the model specification (e.g. incorporation of latent healthy/unhealthy classes, or some specific measurement error process) but, given the time available, we have not pursued these here. The results reported in section 3 exploit the entire sample for estimation, without editing or over-riding the data in any way.

3. Results

3.1. Characteristics of the response sample.

A sample of n = 49,999 responses was received and included in the analysis. Summary information is provided in

Table ***1***.

The sample is well balanced between males and females. All respondents were over 18 years. The largest of the seven age categories was 35-44 years, but there was good spread across age groups between 25 and 74 years. Almost 2,000 responses were obtained from people aged 75 years or over. On average, sample members are well-educated: the most common level of educational attainment was undergraduate degree. Despite being designed as a general-population survey, there is extensive coverage of people with impaired health: 35% of respondents reported an existing diagnosed medical condition and 52% reported taking some type of medication.

The large EEPRU sample means that coverage of both 3L and 5L health states is greatly expanded relative to the EQG study. 90% of the possible 3L health states were observed in the EEPRU study (compared to 51% in the EQG study) and 43% of possible 5L health states were observed (compared to 11% for EQG). Although less than a tenth the size of the GPPS study referred to in section 2.1, the EEPRU dataset covers over half the number of health states observed in GPPS, which is very close to the calculation used to guide the choice of a 50,000 sample size.

Since the EEPRU sample is not specifically targeted on people with health problems, it might be expected to have low coverage of poor health states. This is not so. For example, define a poor 3L health state as any state where the sum of the 3L domain indicators (the 3L 'misery index', ranging from 5 to 15) is 9 or more; and define a poor 5L health state to be one where the 5L misery index (ranging from 5 to 25) is 11 or more.[1] The proportionate coverage of those states is still high: 88% for 3L and 39% for 5L.

[1] The cutoffs were chosen to give similar sample small proportions of poor states for 3L and 5L (roughly 14% of the EEPRU sample in each case).

Table 1: Respondent characteristics

Characteristic		EEPRU data		EQG data	
		n	%	*n*	%
Sex	*Male*	24,309	48.62	1,740	47.33
Education	*Secondary Education (GCSE/O-Levels)*	10,194	20.39		
	Post-Secondary Education (College, A-Level)	8,678	17.36		
	Vocational Qualification (Diploma, Cert) the responses to levels 3 and 5	8,908	17.82		
	Undergraduate Degree (BA, BSc etc.)	13,011	26.02		
	Postgraduate Degree (MA, MSc etc.)	6,564	13.13		
	Doctorate (PhD)	1,607	3.21		
	None of the above	1,037	2.07		
Family status	*Single + no children*	9,366	18.73		
	Single + I have children	2,004	4.01		
	In a relationship + no children	3,600	7.2		
	In a relationship + I have children	2,204	4.41		
	Cohabiting + no children	2,203	4.41		
	Cohabiting + I have children	2,217	4.43		
	Married + no children	5,121	10.24		
	Married + I have children	18,704	37.41		
	Divorced + no children	749	1.5		
	Divorced + I have children	2,357	4.71		
	Widowed + no children	366	0.73		
	Widowed + I have children	1,108	2.22		
Health	*Pre-existing diagnosed condition*	17,688	35.38		
	Taking medication	26,098	52.2		
Acting as carer	*Living in*	7,076	14.15		
	Not living in	8,569	17.14		
Age	*<18*			60	1.63
	18-24	4,210	8.42	588	15.93
	25-34	8,729	17.46	311	8.43
	35-44	9,767	19.53	325	8.81
	45-54	8,640	17.28	528	14.31
	55-64	9,167	18.33	772	20.92
	65-74	7,519	15.04	611	16.55
	75+	1,967	3.93	496	13.44
EQ-5D-3L	*Coverage among all 243 health states*	219	90.12	123	50.62
	Coverage among 192 'poor' health states [1]	168	87.50	88	45.83
EQ-5D-5L	*Coverage among all 3,125 health states*	1,341	42.91	330	10.56
	Coverage among 2,878 'poor' health states [2]	1,126	39.12	302	10.49

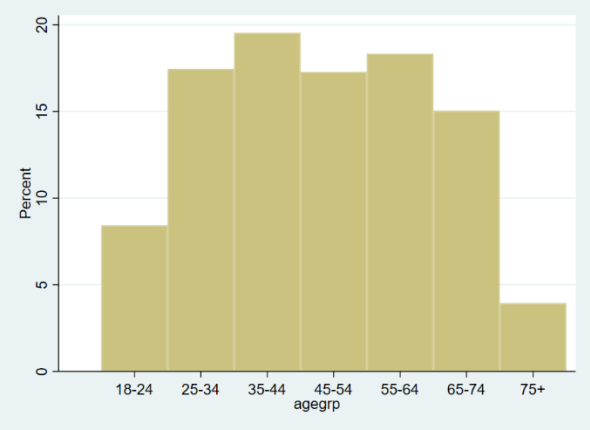
[1] 'Poor' 3L states are those where the five domain indicators sum to 9 or more. [2] 'Poor' 5L states are those where the domain indicators sum to 11 or more.

The age distribution of the EEPRU and EQG data is shown in Figure 2. In the EEPRU data, there are roughly equal sample proportions in each of the ten-year age groups from 25 to 74 years, with smaller numbers in the 18-24 and over 74 age groups. The EQG shows more variation in these same age bands,

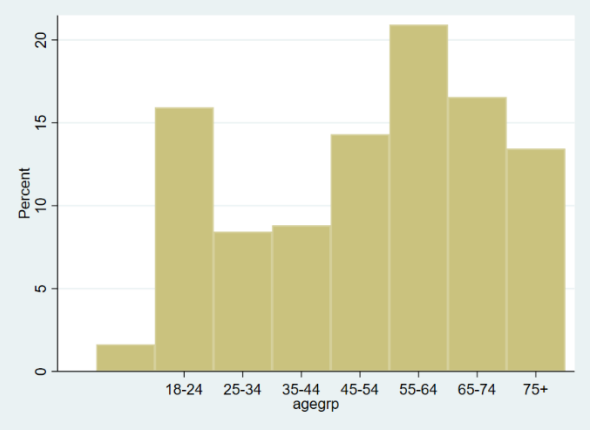
and much higher sample proportions in the youngest and oldest groups. However, the larger size of the EEPRU sample means that there are considerably more observations on the under-25s and over-74s than there are in the EQG sample.

Figure 2: Distribution of age in EEPRU and EQG datasets

a) EEPRU



b) EQG



The distribution of responses to the 3L and 5L instruments are displayed in Figures 3 and 4 for the EEPRU and EQG studies respectively. In the EEPRU dataset, all domains follow the same pattern, for both 3L and 5L: the largest proportions of respondents are in level 1 (no impairment/problems) with decreasing proportions as the degree of severity increases. This demonstrates the requirement for large sample sizes since, proportionally, the responses to levels 3 and 5 of the 3L and 5L instruments are small.

The EQG data shows a different distribution for 3L responses in particular, which may be due to the targeting of individuals with specific conditions. In the mobility domain, there are almost as many responses at level 2 (“Some problems”) as level 1 (“no problems”) and, in the “usual activities”, “pain/discomfort” and “anxiety/ depression” domains, level 2 responses are more frequent than level 1.

Figure 3: Distribution of responses to the 5L and 3L instruments in the EEPRU study

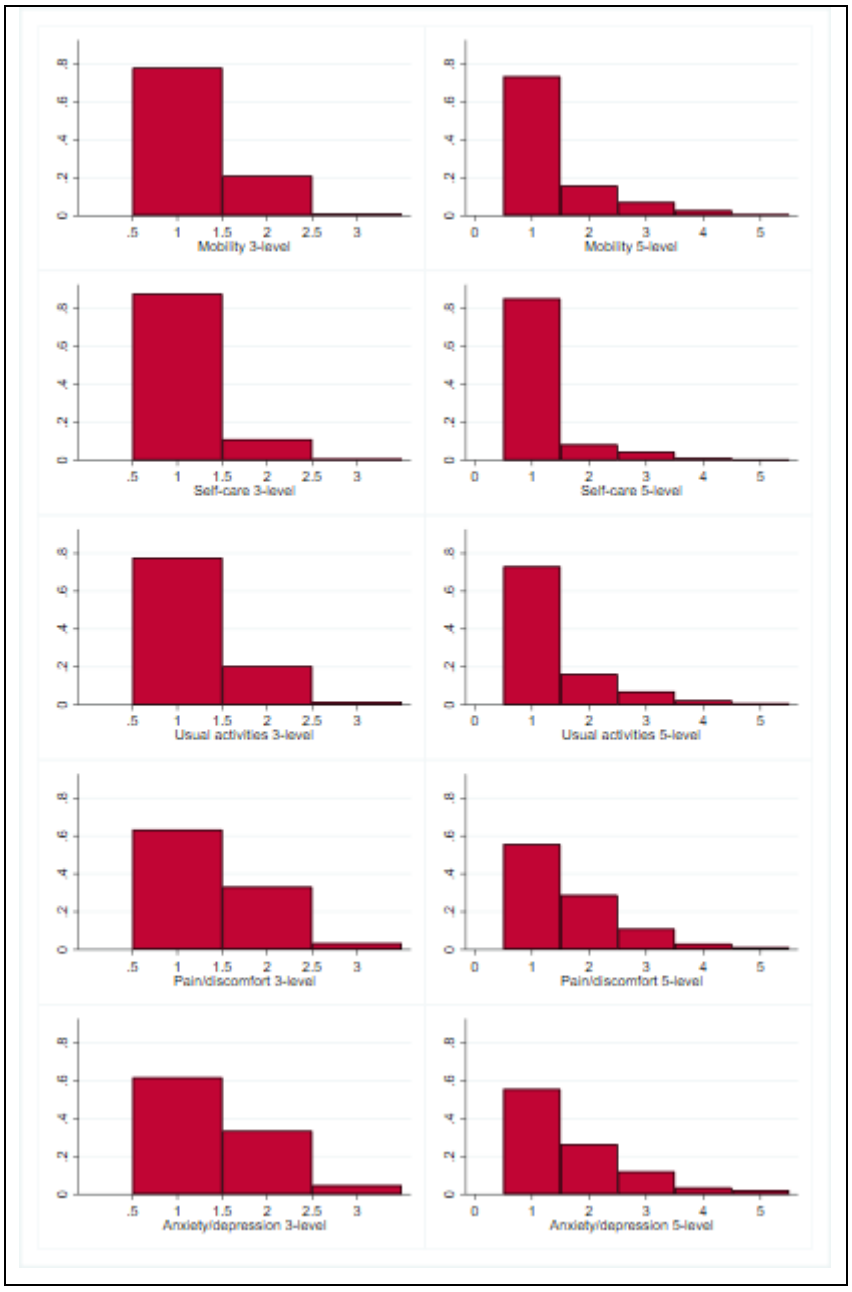


Figure 4: Distribution of responses to the 5L and 3L instruments in the EQG data

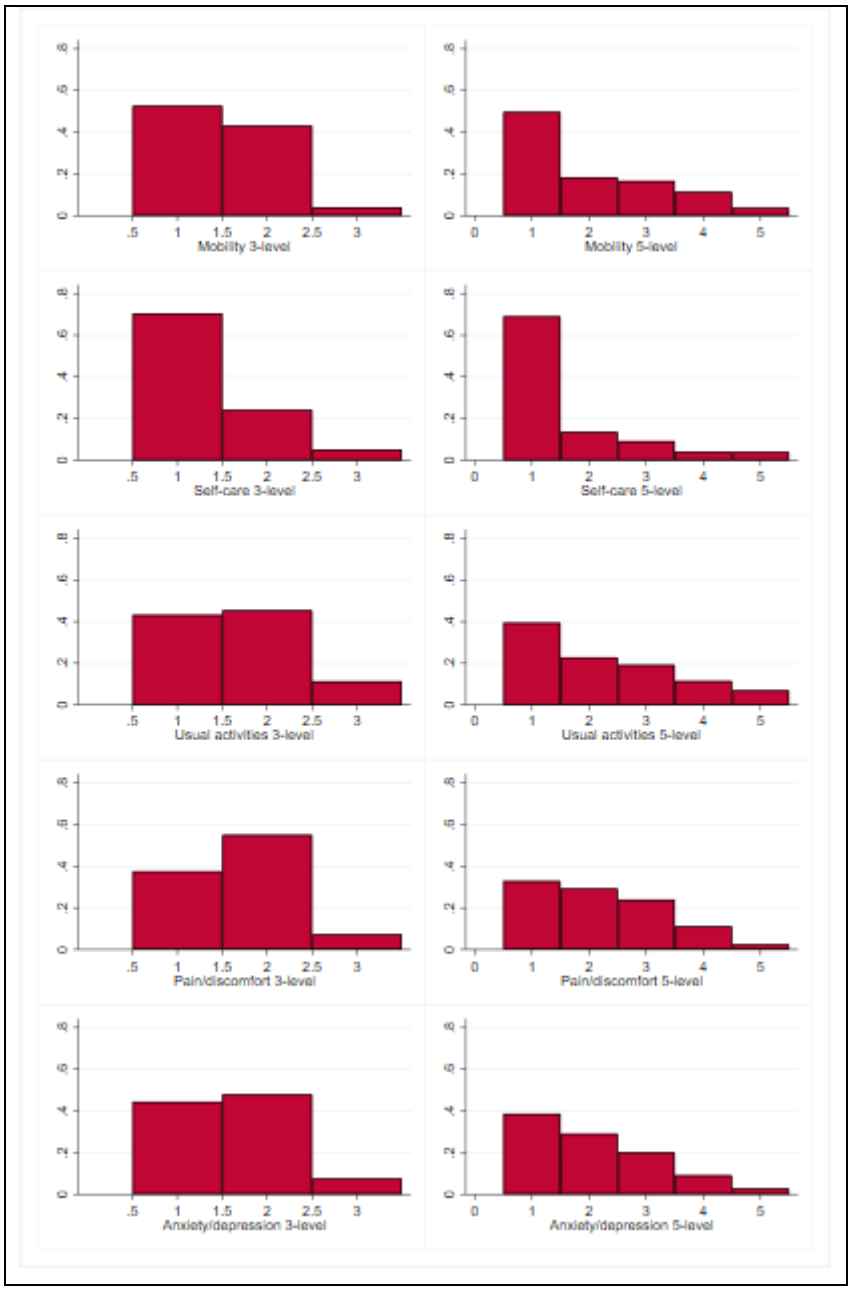


Table 2 shows that the EEPRU study yields larger numbers of cases in all domains and all response categories, for both 3L and 5L, than the EQG data. This is particularly important for the most severe categories of impairment. For both 3L and 5L, the "self care" domain has the fewest responses at maximal impairment ("unable to wash or dress self"); there are 427 vs 189 of these rare responses at 3L level 3 for the EEPRU vs. EQG dataset; in the same domain, there are 250 vs 146 5L responses at the extreme level 5. In other domains, there are up to ten times as many observations of extreme categories in the EEPRU sample compared to the EQG sample.

Table 2: Counts of responses to 3L and 5L by category

		3L		5L	
		EEPRU	**EQG**	**EEPRU**	**EQG**
Mobility					
	1	39,011	1922	36,727	1812
	2	10,488	1576	7,926	672
	3	500	148	3,569	606
	4			1,410	417
	5			367	139
Self-care					
	1	43,994	2568	42,689	2514
	2	5,578	879	4,172	495
	3	427	189	2,289	332
	4			599	149
	5			250	146
Usual activities					
	1	38,875	1574	36,576	1429
	2	10,256	1648	8,203	831
	3	868	408	3,575	699
	4			1,196	417
	5			449	254
Pain					
	1	31,698	1366	27,871	1192
	2	16,584	1999	14,415	1065
	3	1,717	265	5,499	877
	4			1,607	404
	5			607	92
Anxiety/depression					
	1	30,856	1614	27,784	1398
	2	16,756	1748	13,267	1063
	3	2,387	272	6,029	739
	4			1,771	332
	5			1,148	102

Overall, the conclusion is that a large general population survey like the EEPRU study offers an effective way of achieving good coverage of a wide range of health states. Whatever method is subsequently used

for mapping between 3L and 5L (including the modified cross-tabulation crosswalk[4], or DSU model-based response mapping[5], or direct utility mapping[11]), the new EEPRU dataset appears to offer a good foundation.

3.2. Models and predictions

Henceforth, we refer to mappings derived from the Hernández Alava and Pudney model estimated from three different datasets as DSU(FORWARD), DSU(EQG) and DSU(EEPRU). Previous work for the DSU[7] reported a comparison of the Crosswalk and the DSU(EQG) and DSU(FORWARD) models. Here we provide an update on those analyses, comparing the Crosswalk with the DSU (EEPRU) model; some further comparisons with the DSU(EQG) and DSU(FORWARD) models are given in the appendix. We would encourage readers to consult the earlier DSU report for greater detail on all aspects of this study, including interpretation of results, prior to reading this current report.

Meaningful comparisons of model fit and predictive accuracy are difficult to make for two reasons. One is that any measure of in-sample fit to data that were used to estimate one model but not another would bias the comparison in favour of the former. We address that problem by comparing mappings using two datasets that were not used in estimating the DSU or Crosswalk models.

A second difficulty is that the structure of the available FORWARD, EQG, IP and EEPRU datasets may bear little relation to the patient profile in any particular cost-effectiveness analysis. If, for example, a mapping model fits poorly at the lower extreme of health states and if such states happen to be rare in (say) the FORWARD dataset, then that model may appear to do better than other models from the viewpoint of the FORWARD sample, but nevertheless be less reliable when used in an application to a patient group with generally poor health outcomes. There is no complete solution to this problem (except in the unlikely event that one model predicts better than others throughout the health spectrum), since one mapping could be best for some cost-effectiveness studies, while other mappings perform better in other applications.

Here, we use the FORWARD and IP datasets for sample comparisons of the Crosswalk (estimated from EQG data) and the new DSU model (estimated from EEPRU data), since they are out-of-sample for both models. However, neither is ideal; FORWARD is a disease specific registry that may not be generalisable to the general population or other disease areas, and the IP suffers from a small sample size and relatively sparse coverage of poor health states. Differences in model fit in these datasets may be specific to their settings and not indicative of how the models would perform in other samples more representative of cost-effectiveness studies.

Table 3 gives some simple summary measures of predictive accuracy in the FORWARD and IP datasets. Both mappings produce results that are reasonable accurate on average across each of the datasets, with the Crosswalk doing slightly better in terms of mean error (ME), mean absolute error (MAE) and root mean squared error (RMSE), in both datasets. However, the differences are small, with 95% confidence intervals overlapping in all cases.

Table 3: Summary measures of error in predicted 3L utility scores, with 95% confidence intervals

Mapping	Mean error	Mean Absolute Error	Root Mean Square Error
FORWARD dataset			
Crosswalk	0.009	0.094	0.149
	[0.004, 0.013]	[0.091, 0.097]	[0.144, 0.154]
DSU(EEPRU)	0.010	0.100	0.150
	[0.005, 0.014]	[0.097, 0.103]	[0.144, 0.155]
IP dataset			
Crosswalk	0.002	0.055	0.108
	[-0.003, 0.007]	[0.050, 0.059]	[0.099, 0.117]
DSU(EEPRU)	0.007	0.064	0.109
	[0.002, 0.012]	[0.059, 0.068]	[0.100, 0.118]

95% bootstrap confidence intervals (500 replications)

When we look at the results in more depth, the comparison becomes much less straightforward. Table 4 shows the same summary fit measures in the FORWARD data for demographic subgroups. The Crosswalk assumes no differences between men and women in their reporting of health via 3L and 5L, whereas the DSU model allows for demographic differences. For women, the Crosswalk does better in terms of ME and MAE but the RMSE comparison is mixed. For men, the results are much more mixed.

Table 4: Summary measures of error in predicted 3L utility scores by age and sex in the FORWARD data

Demographic group	N	%	Mean error		Mean absolute error		Root mean square error	
			Cross-walk	DSU (EEPRU)	Cross-walk	DSU (EEPRU)	Cross-walk	DSU (EEPRU)
female < 25	25	0.48	0.011	0.009	0.132	0.138	0.179	0.171
female (25-34]	111	2.13	0.019	0.025	0.116	0.126	0.177	0.171
female (35-44]	252	4.84	0.004	0.016	0.099	0.111	0.148	0.157
female (45-54]	708	13.6	0.015	0.025	0.099	0.109	0.153	0.152
female (55-64]	1,300	24.98	0.017	0.025	0.098	0.107	0.151	0.149
female (65-74]	1,186	22.79	0.005	0.011	0.087	0.101	0.141	0.144
female >74	628	12.07	0.002	0.013	0.089	0.107	0.143	0.151
male < 25	1	0.02	-	-	-	-	-	-
male (25-34]	5	0.1	0.039	0.031	0.063	0.065	0.094	0.091
male (35-44]	19	0.37	0.010	0.041	0.090	0.129	0.145	0.183
male (45-54]	123	2.36	-0.032	-0.015	0.131	0.135	0.205	0.208
male (55-64]	303	5.82	0.011	0.009	0.091	0.098	0.140	0.137
male (65-74]	335	6.44	0.008	0.012	0.080	0.093	0.127	0.132
male > 74	209	4.02	-0.003	0.000	0.096	0.109	0.170	0.167

Shaded cells indicate better prediction accuracy in the FORWARD sample

Both the Crosswalk and the DSU model use response mapping – in other words they produce probabilistic predictors of the 3L health description, from which expected utility scores can be calculated. Their predictive performance therefore depends on the accuracy with which they predict the underlying 3L health state. Figures 5 and 6 examine predictive performance for these health states. It is important to do this rather than relying solely on the overall sample fit measures for utility scores in Tables 3 and 4, because cost-effectiveness applications may involve different combinations of health states than are found in the FORWARD and IP datasets.

Figure 5 shows differences between the predicted mean probability of being in each of the three 3L response categories and the corresponding sample proportions of those categories in the FORWARD and IP datasets, for each of the five health domains separately (these are marginal rather than joint analyses). A positive difference indicates under-prediction and a negative value indicates over-prediction, while a value close to zero signifies close alignment to the data. Using the FORWARD dataset as the basis for evaluation, the DSU(EEPRU) prediction errors are relatively close to zero in the domains of mobility, self-care and usual activities at all levels. However, the domains of pain and anxiety/depression show different and rather mixed patterns. For pain, the EEPRU model performs noticeably worse than the Crosswalk at levels 1 and 2, tending to under-predict "no pain" and over-predict "some pain", but is better than the Crosswalk at extreme pain level 3. Comparison of the Crosswalk and DSU(EEPRU) in the anxiety /depression domain shows a small advantage for the Crosswalk at all three levels, with both mappings displaying moderately large mean errors at levels 1 and 2.

Using the IP dataset, the picture that emerges is rather different. Mean errors are smaller for both mappings than they are in the FORWARD data, so there is not much to choose between them. It is striking that the results for pain and anxiety/depression are quite different from those in the FORWAED data, and mostly favour the DSU(EEPRU) mapping. The striking differences between the FORWARD and IP patterns in Figure 5 underline the important point that, in practice, no mapping can be regarded as superior for all applications.

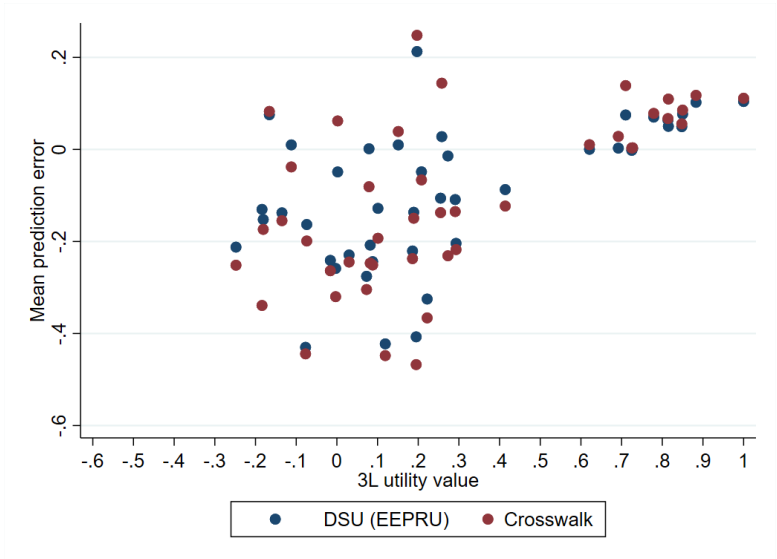
Figure 5: Difference between observed and predicted probabilities for Crosswalk and DSU(EEPRU) mapping models in FORWARD and IP datasets.

(a) FORWARD dataset (b) IP dataset

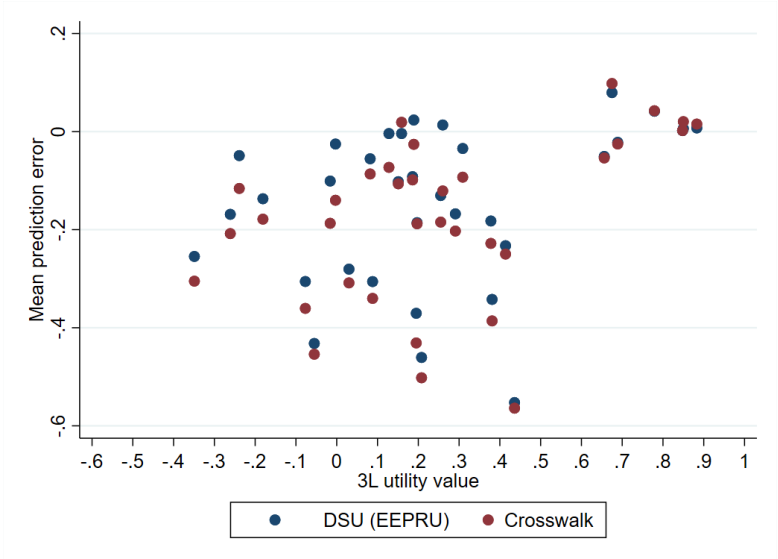
Figure 6 compares the Crosswalk and DSU(EEPRU) utility score predictions for each of the distinct 3L health states observed in the FORWARD and IP datasets. For each of these 3L health states, we calculate the mean mapped utility score for individuals reporting that state, and construct the mean error as the difference between the tariff value for that state and the mean prediction. The FORWARD data generates 83 distinct 3L health states (defined here as states with distinct utility scores); of these, the DSU mapping gives lower mean utility error for 43 states, while the Crosswalk gives lower mean error for 40 states. The IP data generates 61 distinct 3L health states, of which the DSU mapping gives lower mean utility error for 34 states, and the Crosswalk gives lower mean error for 27 states.

Figure 6 shows the pattern of Crosswalk and DSU(EEPRU) predictions for these states. The message that seems to emerge here is that there is no very clear pattern: there is no obvious cluster of points at which one mapping clearly dominates the other.

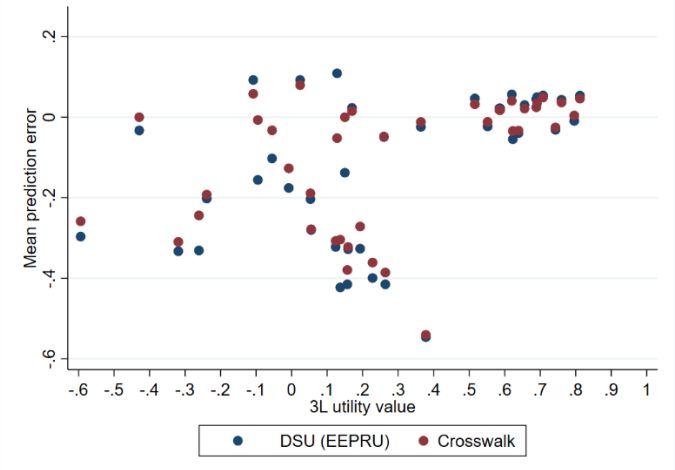
Figure 6: Mean prediction errors of 3L utility scores in the FORWARD and IP data



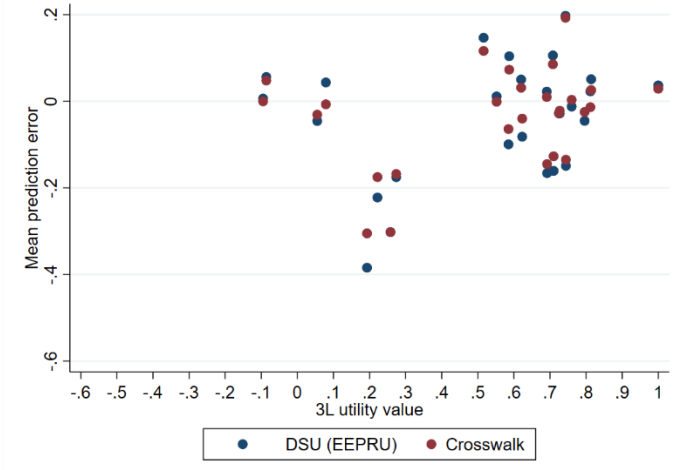
FORWARD data: 3L health states where DSU(EEPRU) has lower MAE (43 states)



IP data: 3L health states where DSU(EEPRU) has lower MAE (34 states)



FORWARD data: 3L health states where Crosswalk has lower MAE (40 states)



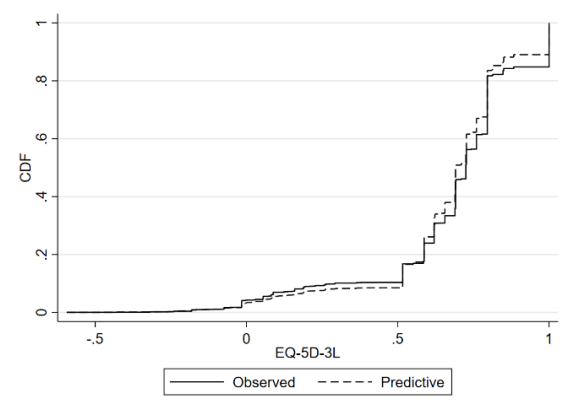
IP data: 3L health states where Crosswalk has lower MAE (27 states)

In many situations, where analysts are using patient level data to inform the utility values they will use in a cost-effectiveness analysis, Monte Carlo simulation models can be used, making random draws of EQ-5D from a predicted or estimated distribution. This is the case for economic evaluations alongside clinical trials and also where individual patient data is used to estimate health utility values for health states in a decision model. For such analyses, the key concept is the cumulative distribution function (cdf), which gives the probability of a 3L utility level less than or equal to any given value[12]. The cdf is a function that rises from 0 at the lowest possible utility level to 1 at the highest possible level. It can be estimated empirically from observed utility scores where they are available, or it can be constructed from an estimated mapping model conditional on observed 5L states. Figure 7 shows the empirical and predicted cdfs for the FORWARD and IP datasets.

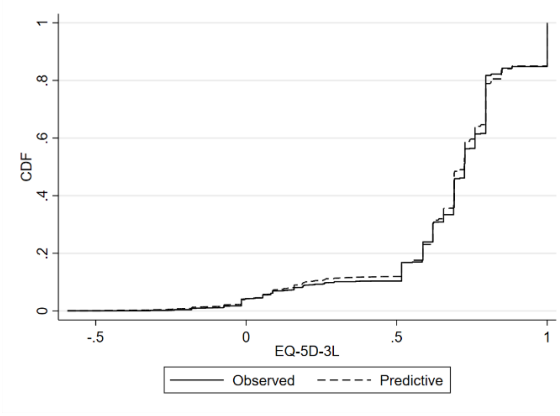
The DSU(EEPRU) model shows good performance in these out of sample tests. There is good alignment with the data between values of 0.5 and 1.0 (full health), especially in the IP sample. There is slight over-estimation for utility values below 0.5 down to about 0.1. Below these values there is very close alignment but note that the vast majority of the data (around 90%) lies above the value 0.5 in both datasets.

Between 0.5 and 1.0, the Crosswalk performs noticeably worse in both the datasets. Between 0.1 and 0.5 there is slight underestimation that is more apparent in the FORWARD dataset. The Crosswalk cdf deviates from the data more than the DSU(EEPRU) cdf, but the difference is not large.

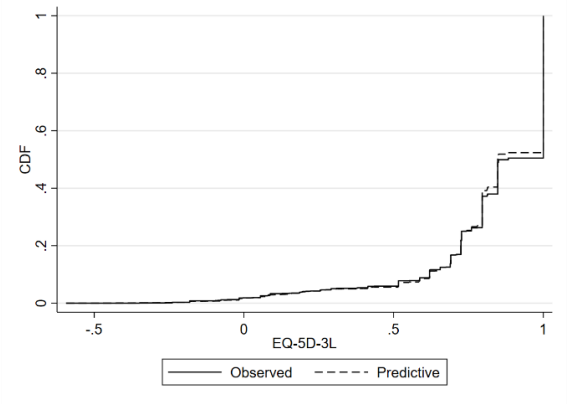
Figure 7: Observed and predicted cumulative distribution functions in the FORWARD and IP datasets, for the Crosswalk and DSU(EEPRU) mappings



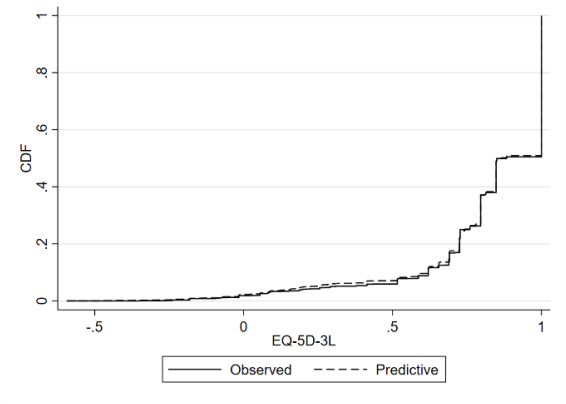
FORWARD dataset: Crosswalk



FORWARD dataset: DSU(EEPRU)



IP dataset: Crosswalk



IP dataset: DSU(EEPRU)

3.3. Understanding differences between model results

We undertook further investigation of model results in an attempt to identify those situations where the Crosswalk and DSU(EEPRU) model might lead to large differences, and to understand the differences between models, datasets and methods of model assessment more clearly. These two approaches are very different in the way they estimate EQ-5D-3L. The DSU(EEPRU) mapping is based on a statistical model, drawing on the EEPRU data in a very flexible framework. The Crosswalk is, in large part, based on assumptions about the relationship between 3L and 5L. Specifically, in any domain, it does not rely on data where a 5L response is at either level 1, 3 or 5. These are assumed to equate to levels 1, 2 and 3 on the 3L instrument with complete certainty. The effect of this is to reduce the prediction error to zero for those observations that align with this view, at the expense of possibly large errors for cases which do not conform to the assumption.

These differences can be illustrated by examining changes to scores in one domain at a time, and comparing model predictions against the observed data from the EQG, FORWARD, IP and EEPRU datasets. We consider four examples of this, starting from 5L health states 11111 (uniform full health) and 33333

(uniform mid-range health). We then vary the mobility level from 1 to 5 and similarly vary the pain domain, for illustrative purposes.

The first set of four columns in Table 5 gives the mean 3L utility score for individuals reporting the relevant 5L health state, in the EQG, FORWARD, IP and EEPRU datasets. They show that, in no dataset do all individuals responding 11111 on the 5L instrument also respond 11111 on the 3L. Hence the mean prediction from a statistical model that does not 'overrule' the data must be below 1 for such cases. The same holds true for the health states 31111, 51111, 13333, 33333, 53333, 11131, 11151, 33313, 33353: there is only 100% consistency in two cases, where the number of relevant observations is too small for the result to be meaningful. Moreover, the conflict is substantial in some cases – for example in states 31111, 13333 and 51111, there is compliance with the Crosswalk assumptions in only 26%, 20% and 6% of cases respectively in the EEPRU data. Note also that there is some evidence of non-monotonicity in the lower part of the 5L response scale – for example, in the EEPRU data, the mean observed 3L utility score rises from 0.821 to 0.930 as we go from the group responding 31111 ("moderate" mobility problems) to the group responding 51111 ("extreme" mobility problems). This may be evidence of some confusion among respondents about the interpretation of "severe" and "extreme" difficulties or, at least, that the relationship is less straightforward than that assumed in the design of the instrument. It should be noted that none of the other datasets have observations for the states 41111 or 51111 to provide more insight into this issue.

The final two columns of Table 5 give the expected 3L utility score predicted by the Crosswalk and DSU(EEPRU) mapping respectively (note that, for the latter, we hold age and sex fixed at 50 and male respectively, so it is not a mean of individual-specific predictions). The most striking aspect of the Crosswalk predictions is the very large difference between the predictions at level 4 and level 5. For example, the difference in Crosswalk 3L utility predictions between 5L states 41111 and 51111 is 0.81-0.34 = 0.47, which is a huge utility decrement for a deterioration from "severe" to "extreme", and far larger than the differences between other levels. Similar large differentials ranging from 0.32 to 0.48 are observed for deteriorations in mobility from state 43333 to 53333 and in pain from states 11141 to 11151 and 33343 to 33353. This is a worrying feature of the Crosswalk, especially in view of the scope for confusion over the interpretation of the labels "severe" and "extreme".

The DSU(EEPRU) predicted utility scores are more faithful to the data and do not display the same large drops from "severe" to "extreme" states. The DSU(EEPRU) mapping also reproduces the non-monotonicity evident in the sample, when the base state is 11111 and mobility and pain are varied beyond level 3.

Table 5: Predicted and mean observed 3L utility scores for illustrative health states

State	Mean 3L utility by 5L health state in four datasets				Predicted 3L utility	
	EQG	FORWARD	IP	EEPRU	Cross-walk	DSU (EEPRU) †
Mobility domain						
11111	0.992	0.973	0.991	0.991	1.00	0.99
% consistent ‖	*95.1*	*87.4*	*96.1*	*96.4*		
21111	0.991	0.915	0.834	0.883	0.88	0.92
31111	0.849§	0.905	0.850§	0.821	0.85	0.91
% consistent ‖	*33.3*§	*33.3*	*100*§	*26.2*		
41111	-	-	-	0.950§	0.81	0.95
51111	-	-	-	0.930	0.34	0.95
% consistent ‖				*5.6*		
13333	-	0.585	-	0.450	0.59	0.48
% consistent ‖		*100*		*20.0*		
23333	0.545	0.568	-	0.435	0.53	0.45
33333	0.509	0.456	0.540§	0.490	0.52	0.43
% consistent ‖	*80.0*	*66.7*	*66.7*§	*67.2*		
43333	0.474	0.585	0.516	0.416	0.48	0.42
53333	0.516§	-	-	0.329§	0.00	0.38
% consistent ‖	*0.0*§			*50.0*§		
Pain domain						
11111	0.992	0.973	0.991	0.991	1.00	0.99
% consistent ‖	*95.1*	*87.4*	*96.1*	*96.4*		
11121	0.840	0.857	0.861	0.867	0.84	0.86
11131	0.796	0.781	0.806	0.796	0.80	0.78
% consistent ‖	*89.3*	*84.3*	*88.2*	*85.9*		
11141	0.796§	0.714§	0.778§	0.605	0.58	0.81
11151	-	0.055§	-	0.816§	0.26	0.83
% consistent ‖		*0.0*§		*25.0*§		
33313	-	-	-	0.626§	0.64	0.52
% consistent ‖				*50.0*§		
33323	0.434§	0.516§	-	0.451	0.54	0.50
33333	0.509	0.456	0.540§	0.487	0.52	0.43
% consistent ‖	*80.0*	*66.7*	*66.7*§	*67.1*		
33343	0.321	0.300	0.585§	0.251	0.30	0.18
33353	-0.016§	0.250§	-	-0.044	-0.02	0.05
% consistent ‖	*100*§	*50.0*§		*75.0*		

§ based on 5 or fewer observations; † covariates fixed at age = 50, sex = male; ‖ percentage reporting the 3L health state assumed by the Crosswalk.

4. Discussion

There is an important current need to be able to map from the EQ-5D-5L descriptive system to the 3L value set. For those situations where the 5L instrument has been collected in clinical studies and analysts therefore have access to the individual patient responses to the 5L descriptive system, the current NICE guidance suggests they should use the Crosswalk mapping function.

There are other situations where analysts will not have access to patient responses but need to map to 3L. Where one needs to map from a 5L health utility score or from a mean utility score from a sample of respondents, to the 3L, any of the DSU mappings based on the Hernández Alava and Pudney model provides this additional option.

In addition, in many situations analysts may wish to map from the 3L (either the descriptive system, a health utility or a mean health utility) to the 5L. If an acceptable 5L value set for England becomes available then this will be required as a means of linking historical evidence to the new 5L values, potentially for decades. The DSU mappings also provide this feature in a consistent way.

Therefore, there are features of importance to end-users of the Hernández Alava and Pudney mapping method that the Crosswalk is not designed to be capable of. In addition, there are major differences in the statistical methods used to produce these alternative approaches. These have been described in some detail previously.[7]

There are limitations in the data underpinning both the Crosswalk and earlier versions of the DSU model. The study reported here attempted to overcome these data limitations and use the DSU method to produce a new, more reliable means of mapping between 3L and 5L, in either direction, and from descriptive system or utility score.

The new data collection (the EEPRU study) has a respondent sample of 50,000. That is a factor of 13.5 times larger than the EQG dataset and results in much fuller coverage of 3L and 5L health states. In terms of coverage of 5L health states, analysis of subsamples from a very large existing 5L dataset showed there are rapidly diminishing marginal returns to sample expansion beyond a sample size beyond about 50,000, so the EEPRU design appears to be a good compromise between the competing demands of low cost and good coverage. The large sample size ensured that ample data were obtained from those reporting poor health, despite the absence of any screening to target adverse health states.

The EEPRU study was conducted entirely using the English variants of 3L and 5L, using a UK population sample, avoiding any potential bias in mapping estimates from the characteristics of non-English wording.

The survey instrument was designed to ensure substantial separation between the 3L and 5L variants of the instrument. Questions on age, sex, a range of other socioeconomic factors, pre-existing health conditions, caring responsibilities, were included *inter alia* and this was the same for all respondents.

Finally, it has been shown that the ordering of 3L / 5L in the instrument has a material impact on the observed responses. We therefore randomised the ordering respondents were presented with.

Limitations remain. It should be noted that the data collection period coincided with the Covid-19 pandemic outbreak and significant limitations to normal life in the UK. It is possible that these circumstances affected the observed data in some unknown way, although it seems unlikely that the relationship between 3L and 5L would be substantially affected by this.

The survey was conducted entirely online. Our pilot work suggested that online data collection would provide reliable results, although there is no formal way of assessing whether any pair of 3L and 5L responses are invalid (the provision of additional response categories and different labels may lead to changes in the nature of the perceived task). At this point it is important to note the differences between the DSU approach compared to the Crosswalk. The latter assumes only certain patterns of response pairs are valid and discards parts of the data that do not conform to this.

Comparisons of the performance of the models was undertaken out-of-sample, using the FORWARD and IP datasets. There are obvious caveats to all results found here. All measures of fit reflect different issues with more or less focus on larger or smaller errors. The FORWARD and IP datasets are certainly not representative of all patient populations of interest to cost-effectiveness analysts, and both surveys have some of the drawbacks that the EEPRU data collection sought to eliminate. They are not gold standard test beds for assessing different mapping approaches.

We find mixed evidence of the relative performance of the Crosswalk and DSU(EEPRU) mapping approaches in the FORWARD and IP datasets, with the Crosswalk performing slightly better in terms of whole-sample summary measures. However, differences are small in relation to the sampling variability in their estimates, with overlapping confidence intervals.

We also report plots of the direct empirical cumulative density function (cdf) of 3L utility scores, and the predicted analogues derived from the Crosswalk and the DSU(EEPRU) model. This shows slightly better performance for the DSU(EEPRU) model compared to the Crosswalk, particularly where most of the data lies, at 3L utility levels exceeding 0.5. Again, differences are fairly small.

We have attempted to identify particular domains or configurations of health states in which the mapping models perform particularly badly, but the picture that emerges is not a clear one and there is some inconsistency between the evidence from the FORWARD and IP datasets. For example, in the former, the DSU(EEPRU) models predicts the mean probability of each response level very well, and better than the Crosswalk, in the mobility, self-care and usual activities domains, but not for the pain dimension. In the IP data, on the other hand, that picture is almost reversed.

Finally, we report on specific combinations of health states and compare model predictions to observed data. Here we show how the Crosswalk can lead to vastly different results than any modelling approach that is fully based on the data. This is particularly clear in relation to health states involving level 1 (no

problems), level 3 or 5 in any of the five dimensions. For such states, the Crosswalk is based entirely on assumption and they do not reflect the mean responses from the observed data. This issue tends to be most important at level 5, where the Crosswalk generates very large decrements when moving from level 4 ("severe" difficulties) to level 5 ("extreme" difficulties). Such large decrements are not consistent with the observed data, nor with the magnitude of change between other levels within the domain and, in our view they may lack face validity. This should be borne in mind when considering use of the Crosswalk in any application in which very poor health states are likely to be encountered.

An interesting feature of both the EEPRU data and the DSU model estimated from it is that there is not always a strictly monotonic relationship between 5L and 3L. This may be a signal of limitations in the 5L descriptive system. There have been concerns raised about the potential ambiguity of levels 4 ("severe problems") and 5 ("unable to" or "extreme problems") of the 5L instrument, which may be the cause of this observation. Unlike valuation studies, where individuals see descriptions of health states that do not give any indication of the expected ordering of levels within each domain, mapping studies provide this context. This may mitigate the impact of wording ambiguity but it is unclear if it eliminates it.

Given the quite mixed picture we have found, it is difficult to give firm recommendations about the types of application to which each mapping approach is best suited. In the absence of a clear message, one possibility would be to conduct sensitivity checks by using alternative mappings in parallel. This would give some indication of the extent to which cost-effectiveness results might be driven by the choice of mapping approach rather than by the evidence itself. However, sensitivity analysis is less appealing from the practical policy point of view, since it complicates decision-making and introduces the risk that analysts might 'cherry pick' the results most favourable to their case.

The DSU mapping approach can be easily implemented by analysts. There is a freely available Stata command (*eq5dmap*) fully described in The Stata Journal.[11] In addition, the command has been translated into Excel and R commands. Updates will be made to these tools to reflect the new analysis based on the EEPRU dataset and all versions of the command will be freely available via the EEPRU and DSU websites.

Appendix: Additional results relating to the DSU(FORWARD) and DSU(EQG) mappings

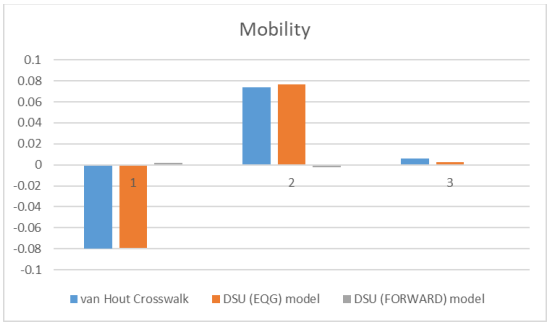
This appendix gives analogues of Table 3 and Figures 5 and 7, comparing the Crosswalk with the two earlier versions of the DSU mapping, based on the Hernández Alava and Pudney model estimated from FORWARD and EQG data.

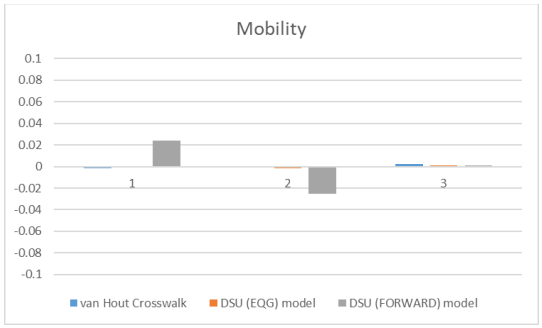
Table A1: Summary measures of error in predicted 3L utility scores, with 95% confidence intervals

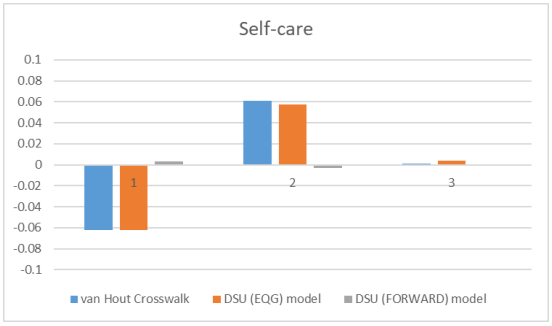
Mapping	Mean error	Mean Absolute Error	Root Mean Square Error
FORWARD dataset			
Crosswalk	0.009	0.094	0.149
	[0.004, 0.013]	[0.091, 0.097]	[0.144, 0.154]
DSU(FORWARD)	0.001	0.100	0.147
	[-0.004, 0.005]	[0.097, 0.103]	[0.142, 0.152]
DSU(EQG)	0.020	0.100	0.148
	[0.016, 0.024]	[0.096, 0.103]	[0.144, 0.153]
IP dataset			
Crosswalk	0.002	0.055	0.108
	[-0.003, 0.007]	[0.050, 0.059]	[0.099, 0.117]
DSU(FORWARD)	0.003	0.071	0.114
	[-0.003, 0.009]	[0.067, 0.075]	[0.106, 0.123]
DSU(EQG)	0.012	0.065	0.109
	[0.007, 0.018]	[0.061, 0.069]	[0.101, 0.117]

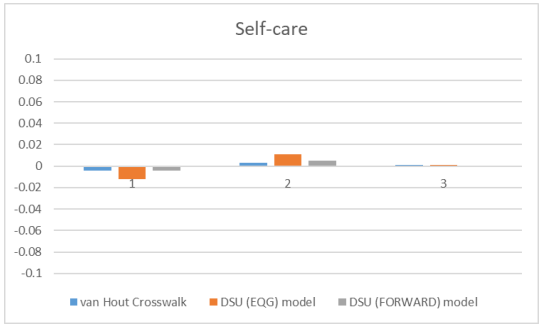
95% bootstrap confidence intervals (500 replications)

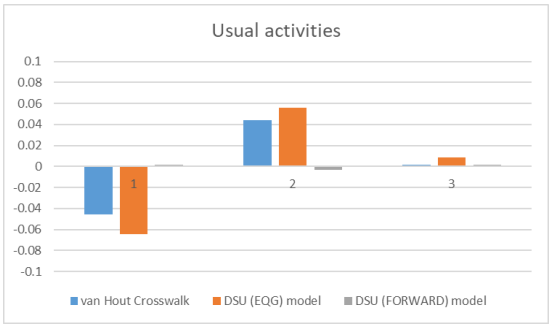
Figure A1: Difference between observed and predicted probabilities for Crosswalk and DSU(EEPRU) mapping models in FORWARD and IP datasets.

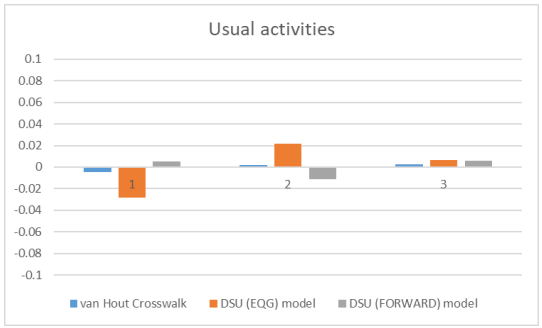


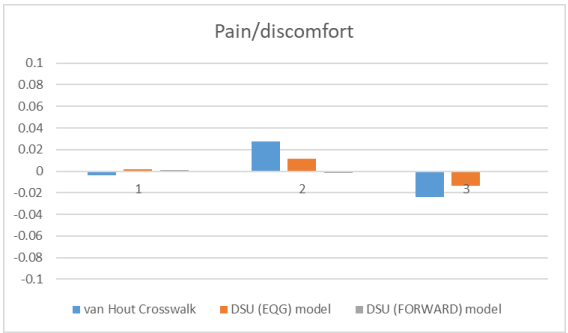


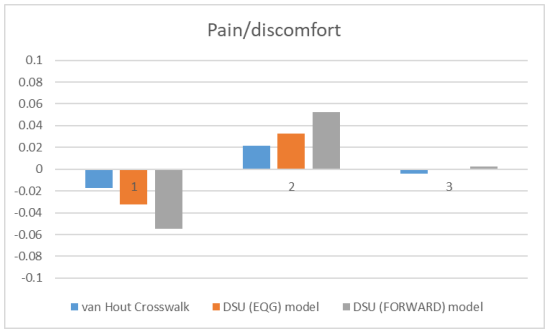


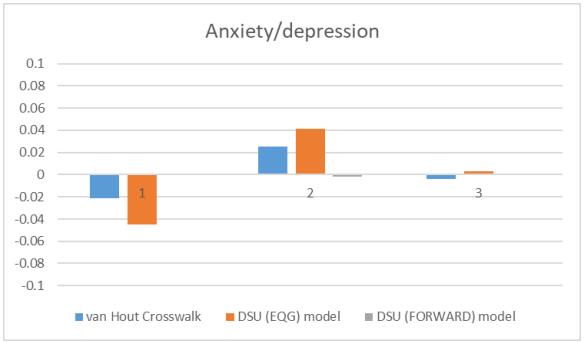


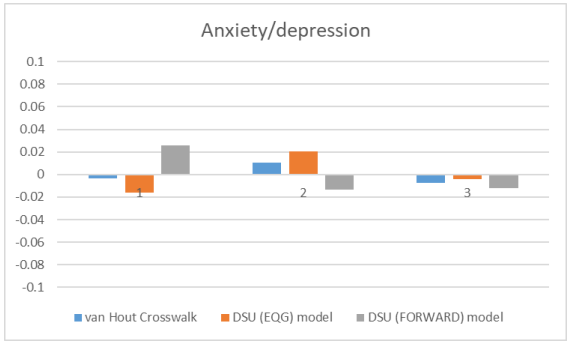








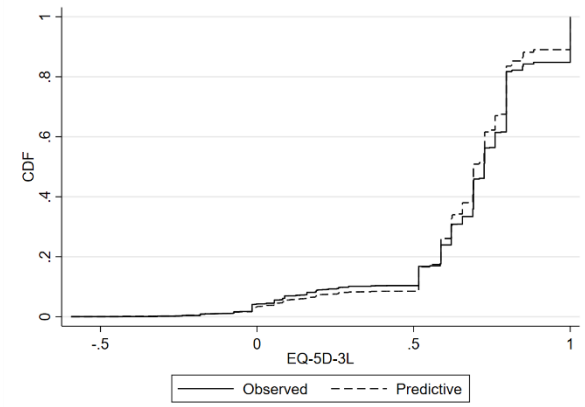




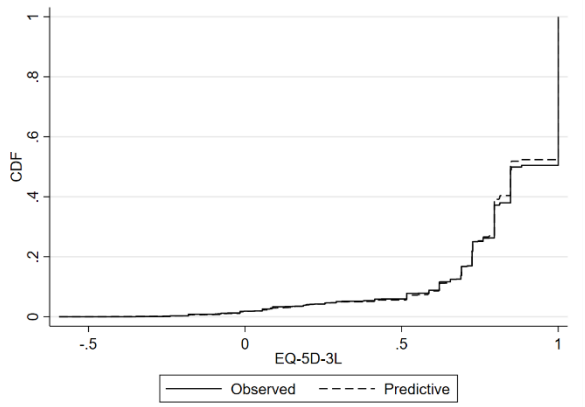
(a) FORWARD dataset

(b) IP dataset

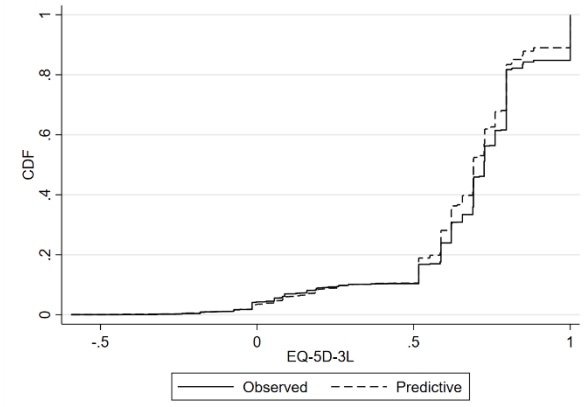
Figure A2: Observed and predicted cumulative distribution functions in the FORWARD and IP datasets, for the Crosswalk and DSU(EQG) and DSU(FORWARD) mappings



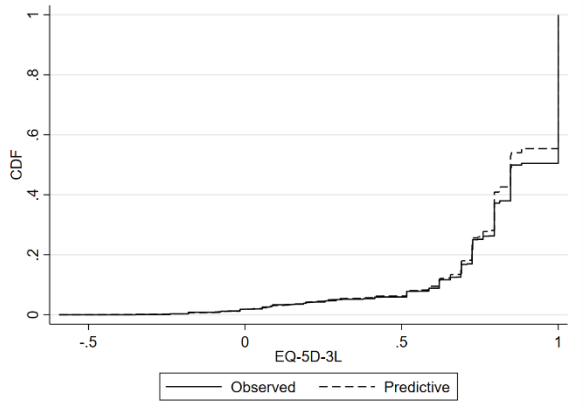
Crosswalk



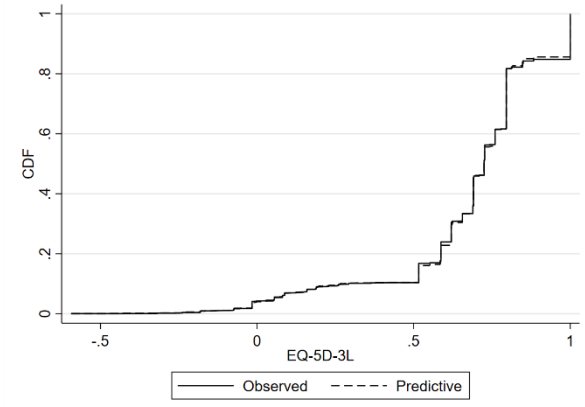
Crosswalk



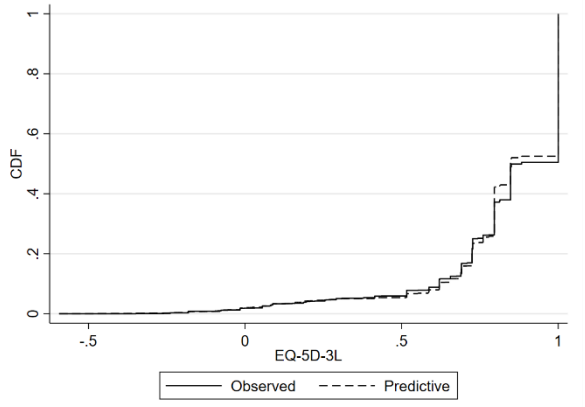
DSU(EQG)



DSU(EQG)



DSU(FORWARD)



DSU(FORWARD)

FORWARD dataset

IP dataset

References

[1] NICE (2013) Guide to the Methods of Technology Appraisal.

[2] Hernández Alava, M., Wailoo A., Grimm S., Pudney S., Gomes M., Sadique Z., Meads D., O'Dwyer J., Barton G., Irvine L. (2018) "EQ-5D-5L versus 3L: the impact on cost-effectiveness in the UK", *Value in Health*, 21:49-56.

[3] Pennington B, Hernández Alava M, Pudney S, Wailoo A. (2019) "The Impact of Moving from EQ-5D-3L to -5L in NICE Technology Appraisals", *Pharmacoeconomics*, 37(1):75-84.

[4] Van Hout B, Janssen MF, Feng Y et al. (2012) Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets, *Value in Health*, 15: 708-15.

[5] Hernández Alava M, Pudney S. (2017) 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. *Journal of Health Economics*, 55:139-152

[6] Wailoo A, Hernández Alava M, Grimm S, et al. (2017) "Comparing the EQ-5D-3L and 5L Versions. What are the Implications for Cost Effectiveness Estimates?" NICE DSU Report, available at http://nicedsu.org.uk/wp-content/uploads/2017/05/DSU_3L-to-5L-FINAL.pdf

[7] Hernández Alava, M, Wailoo A, Pudney S. (2017) Methods for Mapping Between the EQ-5D-5L and the 3L for Technology Appraisal. NICE DSU report available at http://nicedsu.org.uk/wp-content/uploads/2020/06/Mapping-5L-to-3L-DSU-report.pdf

[8] Lu, G., Brazier, J., and Ades, A. E. (2013). Mapping from disease-specific to generic health-related quality-of-life scales: A common factor model. *Value in Health*, 16:177-184.

[9] Janssen, M.F., Pickard, A.S., Golicki, D. et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study Qual Life Res (2013) 22: 1717. doi:10.1007/s11136-012-0322-4

[10] Hernández Alava, M and Pudney, S. (in submission) "Mapping between EQ-5D-3L and EQ-5D-5L. A survey experiment on the validity of multi-instrument data".

[11] Hernández-Alava, M. and Wailoo, A. (2015). Fitting adjusted limited dependent variable mixture models to EQ-5D, *Stata Journal* **15**, 737-750.

[12] Hernández Alava M, Wailoo A, Pudney S, Gray L & Manca A. (2020) Mapping clinical outcomes to generic preference-based outcome measures: development and comparison of methods. Health Technology Assessment, Vol.24(34)