



The
University
Of
Sheffield.

School of
Health
And
Related
Research

Health Economics & Decision Science (HEDS) Discussion Paper Series

EQ-5D- 5L versus 3L: the impact on cost-effectiveness.

Authors: Monica Hernandez Alava, Allan Wailoo, Sabine Grimm, Stephen Pudney, Manuel Gomes, Zia Sadique, David Meads, John O'Dwyer , Garry Barton, Lisa Irvine.

Corresponding author: Allan Wailoo

SCHARR, University of Sheffield,
Regent Court, 30 Regent Street, Sheffield, S1 4DA
Tel: (+44) (0)114 222 0736
Fax: (+44) (0)114 222 0710
Email: a.j.wailoo@sheffield.ac.uk

No. 17.02

Disclaimer:

This series is intended to promote discussion and to provide information about work in progress. The views expressed in this series are those of the authors. Comments are welcome, and should be sent to the corresponding author.

This paper is also hosted on the White Rose Repository: <http://eprints.whiterose.ac.uk/>

EQ-5D-5L versus 3L: the impact on cost-effectiveness.

Monica Hernandez Alava¹ BSc, MSc, PhD, Allan Wailoo BSc MA, MSc, PhD¹, Sabine Grimm BA, MSc, PhD², Stephen Pudney BSc MSc¹, Manuel Gomes BSc MSc PhD³, Zia Sadique MSc, PhD³, David Meads BA MSc, PhD⁴, John O'Dwyer MSc⁴, Garry Barton BA MSc PhD⁵, Lisa Irvine BA MSc⁵

¹ *School of Health and Related Research, University of Sheffield, UK*

² *Maastricht University Medical Centre, Maastricht, Netherlands.*

³ *London School of Hygiene and Tropical Medicine, London, UK*

⁴ *University of Leeds, UK*

⁵ *Norwich Medical School, University of East Anglia, UK*

Acknowledgments:

We wish to thank the following people: Fred Wolfe and Kaleb Michaud for providing data from the NDB; the EuroQoL group for providing data and providing helpful feedback on a presentation at ISPOR, Vienna, 2016: in particular Paul Kind, Bas Jaansen and Andrew Lloyd; Jenny Dunn, ScHARR, for providing administrative support.

Source of funding:

This work was supported by a grant from the Medical Research Council Methodology Research Programme under grant MR/L022575/1 and the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views represented are those of the authors alone.

Corresponding author:

Allan Wailoo
ScHARR
University of Sheffield
Sheffield
S1 4DA
Tel: 00 44 114 2220736
Email: a.j.wailoo@sheffield.ac.uk

ABSTRACT

Objectives

To model the relationship between EQ-5D-3L and EQ-5D-5L and examine how differences impact on cost-effectiveness in case studies.

Methods

We used two datasets that included both EQ-5D-3L and EQ-5D-5L from the same respondents. The EuroQoL dataset (n=3551) included patients with different diseases and a healthy cohort. The National Databank (NDB) dataset included patients with rheumatoid disease (n=5205). We estimated a system of ordinal regressions in each dataset using copula models, to link responses to the 3L instrument to 5L and its tariff, and vice versa. Results were applied to nine cost-effectiveness studies.

Results

Best-fitting models differed between EuroQoL and NDB datasets in terms of the explanatory variables, copulas and coefficients. In both cases the coefficients of the covariates and latent factor between -3L and -5L were significantly different, indicating that the two instruments are not a uniform realignment of the response levels for most dimensions. In the case studies, moving from 3L to 5L caused a decrease of up to 87% in incremental QALYs gained from effective technologies in almost all cases. ICERs increased, often substantially. Conversely, one technology with a significant mortality gain saw increased incremental QALYs.

Conclusion

5L shifts mean utility scores up the utility scale towards full health and compresses them into a smaller range, compared to -3L. Improvements in quality of life are valued less using 5L than with 3L. 3L and 5L can produce substantially different estimates of cost effectiveness. There is no simple proportional adjustment that can be made to reconcile these differences.

INTRODUCTION

The EQ-5D comprises a descriptive system of health-related quality of life and associated tariffs or “utility” scores. The descriptive system covers five dimensions of health: mobility, ability to self-care, ability to undertake usual activities, pain and discomfort, and anxiety and depression. The original version of EQ-5D allows respondents to indicate the degree of impairment on each dimension according to three levels (no problems, some problems, extreme problems). This is the EQ-5D-3L. A new version of the instrument, EQ-5D-5L, includes five levels of severity for each dimension (no problems, slight problems, moderate problems, severe problems, and extreme problems) with the intention of improving the instrument’s sensitivity and reducing ceiling effects¹. Tariffs for this 5L version are now available for England, Canada, Japan, Uruguay, the Netherlands and Korea.

EQ-5D is one of the most widely used instruments underpinning economic evaluations conducted in terms of cost-per Quality Adjusted Life Year (QALY). It is therefore essential to understand the implications of using the new 5L version of the instrument compared to the 3L version. This paper provides information on how the two versions of EQ-5D relate to each other, using the UK tariffs. This is done by utilising two reference datasets where patients filled in both 3L and 5L instruments. We estimate the joint distribution of responses to the two instruments. This model is then used in 9 cost-effectiveness studies to compare results when using directly observed 3L values with estimated 5L results.

METHODS

Data

We used two reference datasets.

The first was provided by the EuroQoL group (the EQG data). Conducted between August 2009 and September 2010, the EuroQoL Group coordinated and partly funded a data collection study. Its main aim was to collect data on both versions of EQ-5D, the 3L and 5L, to compare them in terms of their measurement properties and to generate an interim value set for EQ-5D-5L using a mapping (or cross-walk) approach. 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 6 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 was mainly collected through specialist hospitals/centres and patient recruitment agencies. All countries used paper and pencil questionnaires, apart from 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 EQ-5D-5L and EQ-5D-3L dimensions.

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

Statistical analysis

The aim is to estimate the relationship between the two instruments. Hernandez and Pudney have previously developed a flexible model which allows analysis of the joint responses to EQ-5D-3L and EQ-5D-5Lⁱⁱⁱ. The model is a system of ordinal regressions estimated jointly, incorporating a flexible copula mixture residual distribution. It is a type of response mapping model with all equations for the five health domains and two versions of the EQ-5D instrument estimated jointly. Thus, there are 10 ordinal regressions corresponding to the five dimensions of EQ-5D-5L and the five dimensions of EQ-5D-3L. Following the natural pairing of the dimensions in the two versions of EQ-5D, the 10 regressions are arranged in five groups. Each group corresponds to one EQ-5D dimension and contains an ordinal regression for EQ-5D-5L and another for EQ-5D-3L.

To capture the dependence between the two regressions in each dimension we use a copula representation. Copulas are very useful as they can generate a number of dependence structures. Different copulas were assessed in the analysis.

To complete the model, the five bivariate groups of regressions are linked by a latent factor which represents background response behaviour. Some respondents may have a tendency to give pessimistic assessments, while others tend to make light of their health problems. The common latent factor varying across individuals represents this type of heterogeneity, and has the effect of inducing correlation between all responses from the same individual.

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

Summing up, the multi-equation model described above allows for the discrete nature of responses to EQ-5D and uses a highly flexible mixture-copula specification of the underlying latent model. Importantly, the model does not impose the assumption that responses in the five dimensions of EQ-5D are statistically independent. For the purposes of this study, the advantage of a response mapping type model is that it allows a) the consistency of the responses to the two descriptive systems to be investigated and b) the implied differences in the utility values to be analysed. It, therefore, also enables investigation of the impact on economic evaluation decisions of moving from the 3- level version of EQ-5D to the new 5- level version.

Cost effectiveness case studies

We used the copula mapping models in nine cost-effectiveness case studies.

All were economic evaluations based on individual patient level data using EQ-5D-3L. We made a pragmatic decision in selecting case-studies. We sought collaborators who had previously completed suitable studies using

the 3L instrument and who were willing and able to replicate their study substituting predicted utility scores for 5L using a bespoke Stata command. Included studies were:

- 1) CARDERA - The Combination of Anti-Rheumatic Drugs in Early Rheumatoid Arthritis (CARDERA) trial was a double-blind, factorial designed, placebo-controlled randomized trial which compared the benefits of adding cyclosporine, high-dose step-down prednisolone or both to methotrexate monotherapy^{iv}.
- 2) CACTUS - The Cost-effectiveness of Aphasia Computer Treatment Compared to Usual Stimulation (CACTUS) pilot randomized controlled trial tested the feasibility of comparing self-managed computer therapy combined with usual stimulation (such as participation in normal language stimulation activities and support groups) to usual stimulation alone in people with aphasia^v.
- 3) RAIN - The Risk Adjustment in Neurocritical care (RAIN) trial compared a) Management in a dedicated neurocritical care unit versus a combined neuro/general critical care unit, and; b) 'Early' transfer to a neuroscience centre versus 'no or late' transfer, for patients who initially present at a non-neuroscience centre and do not require urgent neurosurgery, for patients with acute traumatic brain injury.
- 4) IMPROVE - The Immediate Management of Patients with Rupture: Open Versus Endovascular Repair (IMPROVE) trial compared either endovascular repair or open repair of ruptured abdominal aortic aneurysm (AAA)^{vi}.
- 5) COUGAR-02 - The COUGAR-02 randomised, controlled, open-labelled trial compared docetaxel chemotherapy plus active symptom control and active symptom control only in patients in the UK with advanced adenocarcinoma of the oesophagus, oesophagogastric junction, or stomach^{vii}.
- 6) ARCTIC - The Attenuated dose Rituximab with ChemoTherapy in CLL (ARCTIC) study was a multi-centre, randomised, controlled, open, phase IIB non-inferiority trial conducted in previously untreated patients with Chronic Lymphocytic Leukaemia (CLL)^{viii,ix}. It compared fludarabine, cyclophosphamide and rituximab (FCR), which is considered conventional frontline therapy, with fludarabine, cyclophosphamide, mitoxantrone and low dose rituximab (FCM-miniR).
- 7) SHARPISH - The Self-Help and Relapse Prevention in Smoking for Health (SHARPISH) trial^x sought to estimate the effectiveness and cost-effectiveness of self-help booklets versus a single leaflet to prevent smoking relapse in people who had stopped smoking for four weeks.

- 8) WRAP – the Weight-Reduction Activity Programme (WRAP)^{xi} was a multi-centre, non-blinded, three-arm parallel groups randomised controlled trial of two commercial weight loss programmes, compared to a brief intervention in overweight adults.
- 9) CvLPRIT - The CvLPRIT (Complete- compared to Lesion-Only Revascularisation For Myocardial Infarction) trial^{xii} randomised patients presenting with ST-segment elevation Myocardial Infarction (STEMI) with bystander stenosis to an infarct-only strategy (only treat the blocked artery which caused the heart attack) vs. complete revascularisation (treat the blocked artery and also treat any narrowed arteries which may cause heart attacks in future).

We use the UK/English value sets for the 3L and 5L versions of EQ-5D^{xiii,xiv}.

RESULTS

Datasets

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

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

- i. Decreasing profile with a dominant mode at the first category.

This distributional shape can be seen in the self-care dimension of both EQ-5D-3L and EQ-5D-5L and in the mobility and usual activities dimension of EQ-5D-5L in the EQG dataset and on the self-care and anxiety/depression of both versions of EQ-5D in the NDB dataset.

- ii. Decreasing profile with a heavier central section.

In the EQG dataset, the pattern can be seen in the mobility dimension (EQ-5D-3L) and, pain/discomfort and anxiety/depression (EQ-5D-5L). In the NDB dataset, the mobility and usual activities dimensions for both versions of EQ-5D exhibit this shape.

- iii. A strong mode in the centre of the distribution.

This shape can be found in the pain/discomfort dimension in EQ-5D-3L in the EQG dataset and in both versions of EQ-5D in the NDB dataset.

iv. A mode in the centre of the distribution and an almost as large first category.

This distributional shape is similar to shape (ii) in that they both exhibit a decreasing profile, but shape (iv) has less central concentration. This shape can only be found in the EQG dataset in the usual activities and anxiety/depression dimensions of EQ-5D-3L.

In the NDB dataset, both versions of EQ-5D display the same pattern within each dimension but different shapes across dimensions: shape (i) in both the self-care and anxiety/depression dimensions, shape (ii) in the mobility and usual activities dimension and shape (iii) in the pain/discomfort dimension. In contrast, in the EQG dataset only the self-care dimension shows the same shape of distribution in both EQ-5D-3L and EQ-5D-5L. Within the EQG dataset, the distributional shapes for all dimensions of EQ-5D-5L are similar, displaying a decreasing profile corresponding to either shape (i) or (ii). The EQ-5D-3L distributions in the EQG dataset exhibit all four distributional shapes and appear more different across dimensions than in the 5 level version.

Figure 2 shows kernel estimates of the distributions of utility scores in both datasets. EQ-5D-3L in both datasets exhibits the typical characteristics documented in the literature: a large mass of observations at 1 (full health), a gap of no observations between full health and the next feasible value (0.883) and a multimodal distribution. In both datasets, the distributions are smoother for EQ-5D-5L, especially towards the top of the distribution. The number of individuals in full health is reduced by using EQ-5D-5L and the mode at the bottom of the distribution around the value of zero in the EQ-5D-3L distribution disappears in the distribution of EQ-5D-5L. The mean and median of EQ-5D-5L are higher than the corresponding mean and median of EQ-5D-3L in both datasets (see Table 1). The range of EQ-5D-5L is smaller as the worst state has a utility score of -0.281 compared to -0.594 of EQ-5D-3L.

Table 1: Descriptive statistics in the EQG and NDB estimation samples

	EQG sample	NDB sample
AGE		
Mean [95% confidence Interval]	51.23 [50.57, 51.89]	63.32 [62.99, 63.65]
Median [95% confidence Interval]	54 [54, 56]	64.13 [63.78, 64.46]
Standard Deviation	20.11	12.31
Minimum	13	16.66
Maximum	99	95.20

	0.53		0.81	
	<i>EQ-5D-3L</i>	<i>EQ-5D-5L</i>	<i>EQ-5D-3L</i>	<i>EQ-5D-5L</i>
UTILITY				
Mean [95% confidence Interval]	0.628 [0.617, 0.639]	0.712 [0.703, 0.722]	0.681 [0.674, 0.688]	0.779 [0.773, 0.784]
Median [95% confidence Interval]	0.691 [0.691, 0.725]	0.802 [0.792, 0.816]	0.725 [0.725, 0.727]	0.823 [0.817, 0.829]
Standard Deviation	0.333	0.278	0.254	0.191
Minimum	-0.594	-0.281	-0.594	-0.226
Maximum	1	1	1	1
Number of health states [percentage out of possible health states]	123 [50.62]	660 [21.12]	86 [35.39]	524 [16.77]

Statistical model results

Five different copulas, Gaussian, Clayton, Frank, Gumbel and Joe, were assessed in the analysis. The initial specification had gender, age and the square of age as covariates. The square of age was significant when the model was estimated with EQG data, but grossly insignificant when estimated with NDB data. The preferred specification for the EQG dataset has age, age squared and gender as covariates in all ten ordinal regressions whereas the model for the NDB dataset excludes the square of age.

Table 2 summarizes the results for the two datasets. There are several differences between the models from the two datasets. The best fitting model in the EQG dataset chooses the same copula, Frank, in all dimensions of EQ-5D. In contrast, the best fitting model in the NDB dataset selects a Gaussian copula for the mobility, usual activities and pain/discomfort dimensions, a Clayton copula for the self-care dimension and a Frank copula for the anxiety/depression dimension. Therefore, in the EQG dataset the patterns of residual dependence between the 3- and 5- level versions of EQ-5D are similar across all dimensions indicating symmetric dependence and weak dependence on the tails. In the NDB dataset, a Frank copula was also selected for the anxiety/depression dimension and the parameter of dependence was very similar that estimated in the EQG dataset. In contrast, the Gaussian copula in the mobility, usual activities and pain/discomfort dimensions indicate symmetric

dependence as well but stronger dependence on the tails of the distribution than the Frank copula selected in the EQG dataset. The copula chosen in the self-care dimension using the NDB dataset, the Clayton copula, displays a very different pattern of dependence compared to the Frank copula chosen in the EQG dataset. It exhibits asymmetric dependence on the tails with strong dependence at lower values and weak dependence at high values.

There are significant statistical differences in the coefficients of the covariates and latent factor between EQ-5D-3L and EQ-5D-5L in most dimensions. This highlights that the effect of moving from 3 levels to 5 levels is not just a uniform realignment of the response levels. The only exception to this in both datasets is in the anxiety/depression dimension and in the self-care dimension in the NDB dataset.

Cost effectiveness results

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

There are two exceptions. In the WRAP study, we see that in the comparison of the 52-week programme (CP52) compared to the brief intervention, incremental QALYs increase, albeit very slightly, when using 5L (EQG) compared to 3L. Further investigation revealed that because utility was lower for intervention compared to control at 12 months, this caused the cumulative impact to lower total QALYs using 5L. The results of this case study were still consistent with the other case studies in that differences in utilities tend to be smaller using 5L than with 3L.

COUGAR 02 is the only other case study with an increase in incremental QALYs as a result of shifting from 3L to 5L. The increase is small but is apparent for both versions of 5L estimates. In COUGAR 2, mortality is a very substantial driver of cost effectiveness. Median overall survival in the DXL + ASC group was 5.2 months (95%CI 4.1–5.9) versus 3.6 months (3.3–4.4) in the ASC group¹⁸. Here, the value of improved survival is greater because utility values are increased using 5L. It is worth noting that whilst the RAIN study also included patients with a substantial mortality rate (approximately 25% mortality within 6 months) this was substantially lower than in COUGAR-02 (approximate 6-month mortality of 75% in the control group and 60% in the docetaxel arm¹⁸) and did not outweigh the morbidity effect.

The 5L instrument and tariff have the effect of shifting mean utility scores further up the utility scale towards full health, and compressing them into a smaller range. Thus, improvements in quality of life tend to be valued less using 5L than equivalent changes measured with 3L.

In six of the nine reported comparisons, the incremental QALY gain is greater when measured using EQ-5D-5L and the EQG dataset, compared to EQ-5D-5L and the NDB dataset. One of the three remaining comparisons showed no difference.

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

The impact of these changes on ICERs is also substantial in several cases. In CARDERA, the comparison of triple therapy compared to DMARD monotherapy changes from approximately £16k using EQ-5D-3L to over £24k using EQ-5D-5L (EQG data) and over £30k using EQ-5D-5L (NDB data). CACTUS changes from a highly cost effective central estimate using EQ-5D-3L (£3058) to one that is more borderline (£23022) using EQ-5D-5L (NDB data). CVLPRIT changes from an ICER of just over £20k per QALY to in excess of £45k per QALY when using either estimate of 5L health utility. Other case studies demonstrate changes in cost effectiveness that may not span boundaries of typically cited cost-effectiveness thresholds but are, nevertheless, very substantial.

CONCLUSIONS

We have shown that EQ-5D-3L and 5L versions can produce substantially different estimates of cost effectiveness in a series of case studies spanning different health conditions, severities and health technologies. Technologies that improve quality of life have those benefits valued more highly, in terms of health utility, when using the 3L instrument compared to 5L. This is because of the combined effect of the changed descriptive system and how individuals respond to it compared to 3L (which we demonstrated is not the same across each health dimension), and the changed valuation system. The result is that, in almost all cases, it is

estimated that the incremental cost effectiveness ratio of a clinically effective technology would be higher (i.e. becomes less cost-effective) if the 5L instrument had been used in place of the 3L. Where the cost effectiveness of a technology is substantially driven by mortality rather than morbidity gains, the impact of shifting the 5L may lower ICERs (improve cost-effectiveness).

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

It is feasible to adjust 3L evidence to its 5L equivalent, as has been done in this paper. The validity of this approach is, in part, dependent on the data on which it is based. We have demonstrated this method in two separate datasets and shown that they give substantially different results. Further investigation of the reasons for these differences is required. In particular, the NDB includes only patients with rheumatoid disease and may not be generalizable to other populations. However, the design of the NDB questionnaires included much more separation between the completion of 3L and 5L and may, therefore, offer more independent observations than the EQG studies.

There are a number of implications for policy in the light of these results. Given the differences between 3L and 5L consistency in decision making will be difficult to achieve. Consideration must be given to the value of any cost-effectiveness threshold (or thresholds) or other means for making adjustments between the two instruments. Mapping can help achieve this, and the copula-based method is a sophisticated development of “response-mapping” that obtains consistent and accurate results. 5L is increasingly being used in studies of clinical effectiveness, but this is unlikely to entirely replace existing evidence using 3L that will remain of relevance to many economic evaluations for many years to come.

-
- ⁱ 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(10): 1727–1736.
- ⁱⁱ Wolfe F and Michaud K. The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank. *Rheumatology* 2011;50:16-24.
- ⁱⁱⁱ Hernandez Alava M, Pudney S. (2016) “Copula-based modelling of self-reported health states An application to the use of EQ-5D-3L and EQ-5D-5L in evaluating drug therapies for rheumatic disease”, *Health Economics & Decision Science (HEDS) Discussion Paper Series*
- ^{iv} Choy EHS, Smith CM, Farewell V et al. Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis. *Ann Rheum Dis* 2008;67:65663.
- ^v Latimer NR, Dixon S, Palmer R. Cost-utility of self-managed computer therapy for people with aphasia. *International Journal of Technology Assessment in Health Care* 2013;29:4: 402–409.
- ^{vi} IMPROVE Trial Investigators. Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from IMPROVE randomised trial. *BMJ* 2014;348:f7661
- ^{vii} Ford HER, Marshall A, Bridgewater JA, et al on behalf of the COUGAR-02 Investigators Docetaxel versus active symptom control for refractory oesophagogastric adenocarcinoma (COUGAR-02): an open-label, phase 3 randomised controlled trial. *Lancet Oncol* 2014; 15: 78–86
- ^{viii} Hillmen P, Milligan D, Schuh A et al. Results Of The Randomised Phase II NCRI Arctic (Attenuated dose Rituximab with ChemoTherapy In CLL) Trial Of Low Dose Rituximab In Previously Untreated CLL. *Blood*;2013, 122:1639
- ^{ix} Howard, DR, Munir, T, McParland, L et al. (2015) Clinical effectiveness and cost-effectiveness results from the randomised, phase IIB trial in previously untreated patients with Chronic Lymphocytic Leukaemia (CLL) to compare fludarabine, cyclophosphamide and rituximab (FCR) with fludarabine, cyclophosphamide, mitoxantrone and low dose rituximab (FCM-miniR): the Attenuated dose Rituximab with ChemoTherapy In CLL (ARCTIC) trial. *Health Technology Assessment.* ISSN 1366-5278 (In Press)
- ^x Blyth, A, Maskrey, V, Notley, C, Barton, GR, Brown, JT, Aveyard, P, Holland, R, Bachmann, OM, Sutton, S, Leonardi Bee, J, Brandon, HT, and Song, F Effectiveness and economic evaluation of self-help educational materials for the prevention of smoking relapse: randomised controlled trial. . *National Institute for Health Research*, 2015.
- ^{xi} Ahern AL, Aveyard PN, Halford JC, Mander A, Cresswell L, Cohn SR et al. Weight loss referrals for adults in primary care (WRAP): Protocol for a multi-centre randomised controlled trial comparing the clinical and cost-effectiveness of primary care referral to a commercial weight loss provider for 12 weeks, referral for 52 weeks, and a brief self-help intervention [ISRCTN82857232]. 2014 Jun 18;14(1). 620. Available from: 10.1186/1471-2458-14-620
- ^{xii} Gershlick AH, Khan JN, Kelly DJ et al. "Randomized trial of complete versus lesion-only revascularization in patients undergoing primary percutaneous coronary intervention for STEMI and multivessel disease: the CvLPRIT trial." *Journal of the American College of Cardiology* 65.10 (2015): 963-972.
- ^{xiii} Dolan, P. (1997). Modeling valuations for EuroQol health states. *Medical Care*, 35:1095–1108.
- ^{xiv} Devlin, N., Shah, K., Feng, Y., Mulhern, B., and van Hout, B. (2016). Valuing health related quality of life: An EQ-5D-5L value set for England. Technical Report 16.02, *Health Economics & Decision Science*, University of Sheffield.

TABLES:

Table 1: Descriptive statistics in the EQG and NDB estimation samples

	EQG sample		NDB sample	
AGE				
Mean [95% confidence Interval]	51.23 [50.57, 51.89]		63.32 [62.99, 63.65]	
Median [95% confidence Interval]	54 [54, 56]		64.13 [63.78, 64.46]	
Standard Deviation	20.11		12.31	
Minimum	13		16.66	
Maximum	99		95.20	
Proportion female	0.53		0.81	
UTILITY	<i>EQ-5D-3L</i>	<i>EQ-5D-5L</i>	<i>EQ-5D-3L</i>	<i>EQ-5D-5L</i>
Mean [95% confidence Interval]	0.628 [0.617, 0.639]	0.712 [0.703, 0.722]	0.681 [0.674, 0.688]	0.779 [0.773, 0.784]
Median [95% confidence Interval]	0.691 [0.691, 0.725]	0.802 [0.792, 0.816]	0.725 [0.725, 0.727]	0.823 [0.817, 0.829]
Standard Deviation	0.333	0.278	0.254	0.191
Minimum	-0.594	-0.281	-.594	-0.226
Maximum	1	1	1	1
Number of health states [percentage out of possible health states]	123 [50.62]	660 [21.12]	86 [35.39]	524 [16.77]

Table 2: Summary of final model results

	EQG	NDB
Log-likelihood	-23891.83	-33621.04
Number of parameters	78	68
Observations	3551	5205
Type of mixture in copula	Single mixture	Single mixture
Dimension Specific		

<i>Mobility</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	7.12*	11.86***
Equality of coefficients (latent factor)	8.37***	10.64***
Equality of coefficients (covariates & factor)	12.19**	26.49***
<i>Self-care</i>		
Copula	Frank	Clayton
Equality of coefficients (covariates)	8.53**	1.21
Equality of coefficients (latent factor)	3.68*	0.09
Equality of coefficients (covariates & factor)	9.39*	1.35
<i>Usual activities</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	3.29	0.67
Equality of coefficients (latent factor)	5.62**	8.24***
Equality of coefficients (covariates & factor)	0.04**	9.11**
<i>Pain/discomfort</i>		
Copula	Frank	Gaussian
Equality of coefficients (covariates)	0.57	34.36***
Equality of coefficients (latent factor)	9.36***	19.99***
Equality of coefficients (covariates & factor)	11.95**	50.74***
<i>Anxiety/depression</i>		
Copula	Frank	Frank
Equality of coefficients (covariates)	5.60	4.94*
Equality of coefficients (latent factor)	1.23	1.94
Equality of coefficients (covariates & factor)	7.08	6.19
Statistical significance: * = 10%, ** = 5%, *** = 1%		

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

	Inc QALYs							
	3L	5L EQG	% change	5L NDB	% change	3L	5L EQG	%
CARDERA 1	0.145	0.113	-21.8%	0.111	-23.2%	4648	5940	
CARDERA 2	0.084	0.075	-10.4%	0.077	-8.0%	13666	15252	
CARDERA 3	0.082	0.054	-33.5%	0.043	-47.6%	15929	23940	
Cactus	0.150	0.050	-66.7%	0.020	-86.7%	3058	9481	
Rain a	0.020	0.005	-75.0%	0.003	-85.0%	184700	738800	
Rain b	0.051	0.021	-58.8%	0.021	-58.8%	294137	714333	
Improve	0.052	0.046	-11.5%	0.042	-19.2%	-44617*	-48113	
Cougar 2	0.115	0.119	3.5%	0.118	2.6%	27180	26434	
Arctic	0.059	0.043	-27.1%	0.046	-22.0%	112193	162774	
Sharpish	0.000	-0.003	NA	-0.003	NA	NA**		
WRAP - CP12	0.062	0.047	-23.7%	0.039	-36.2%	1812	2373	
WRAP - CP52	0.044	0.044	0.0%	0.036	-19.0%	4305	4312	
CvLPRIT	0.020	0.010	-52.5%	0.009	-53.0%	21496	46761	

CARDERA 1 = MTX vs MTX + CS, CARDERA 2 = MTX vs MTX + PNS, CARDERA 3 = MTX + CS + PNS vs MTX

* In the IMPROVE study, the technology of interest (endovascular aneurysm repair) was cost saving.

** Incremental QALYs near zero meant the calculation of the ICER may be misleading and was therefore not reported.

FIGURES

Figure 1: Response histograms for EQ-5D-3L and EQ-5D-5L in the EQG dataset and the NDB dataset

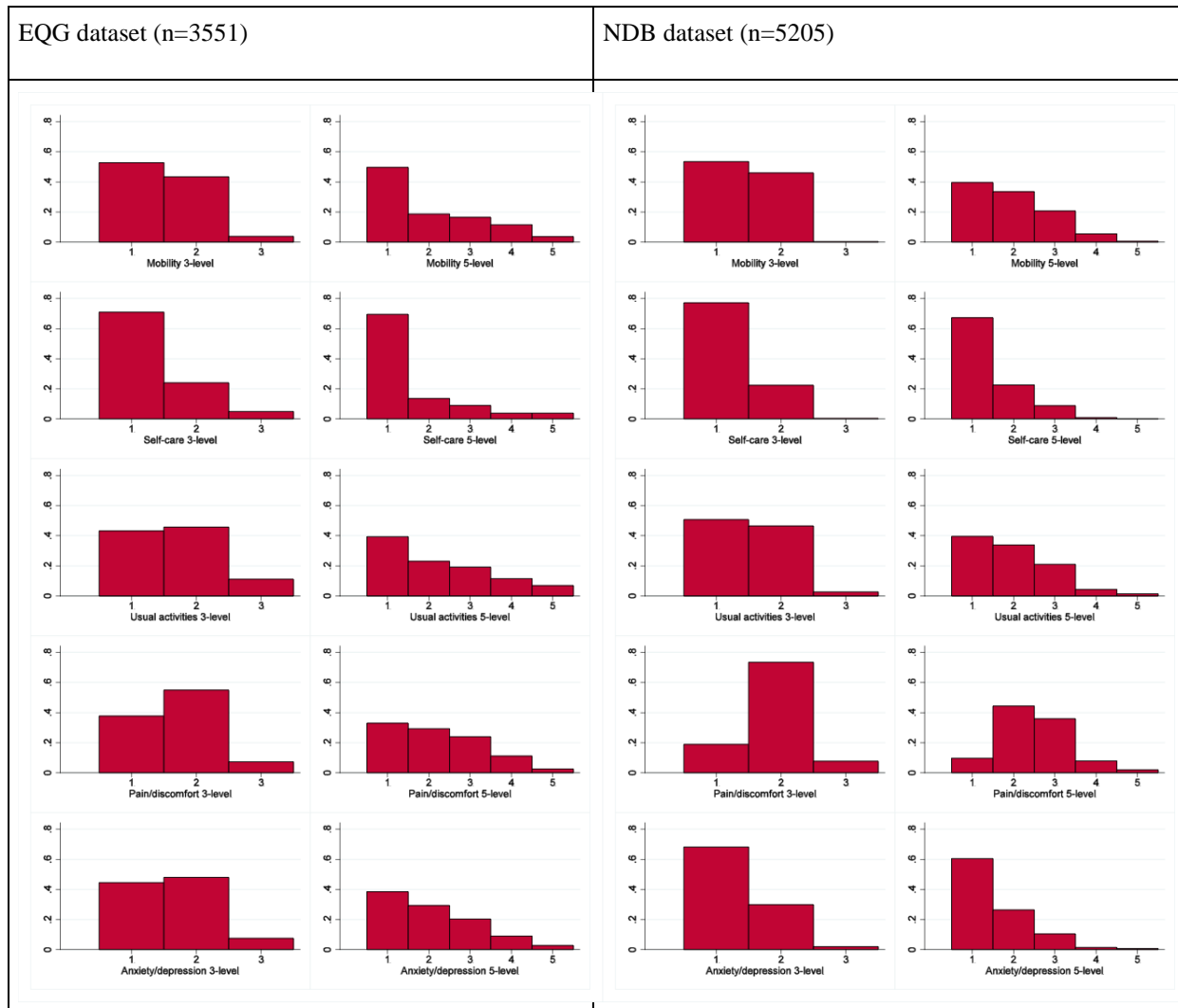


Figure 2: Smoothed empirical distribution functions of EQ-5D-3L and EQ-5D-5L in the EQG and NDB datasets.

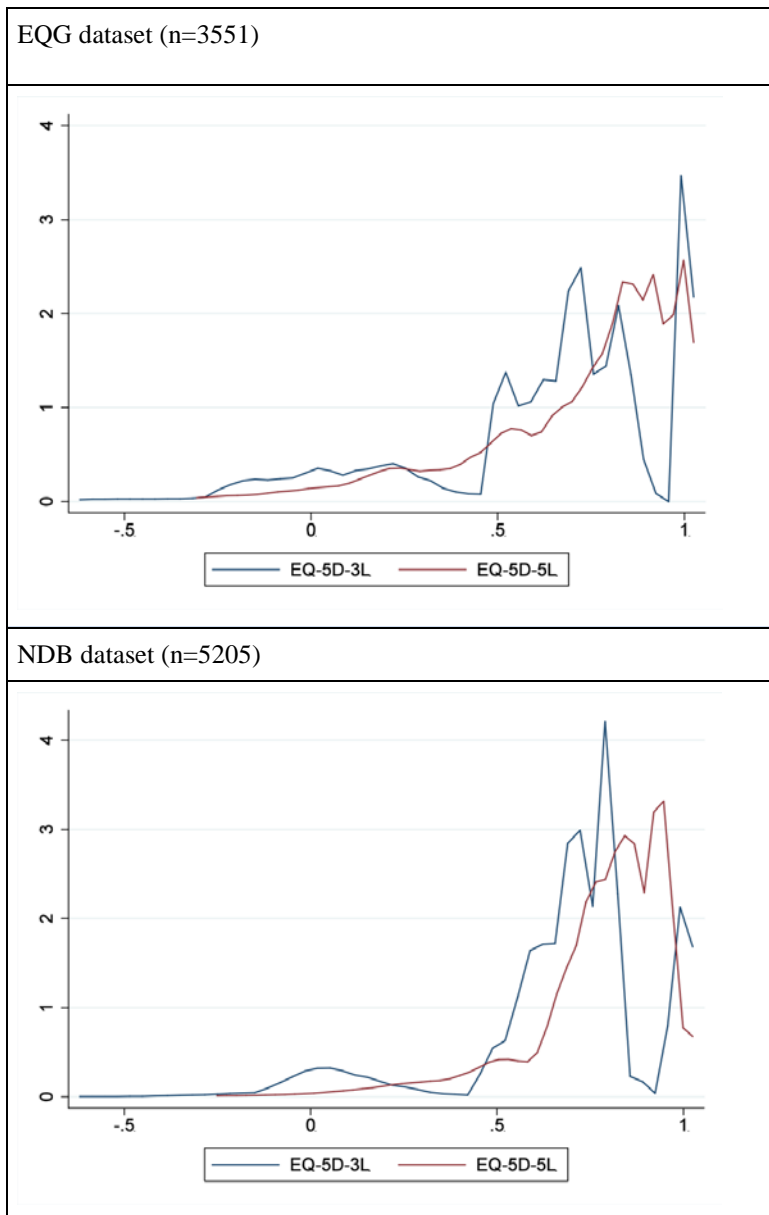
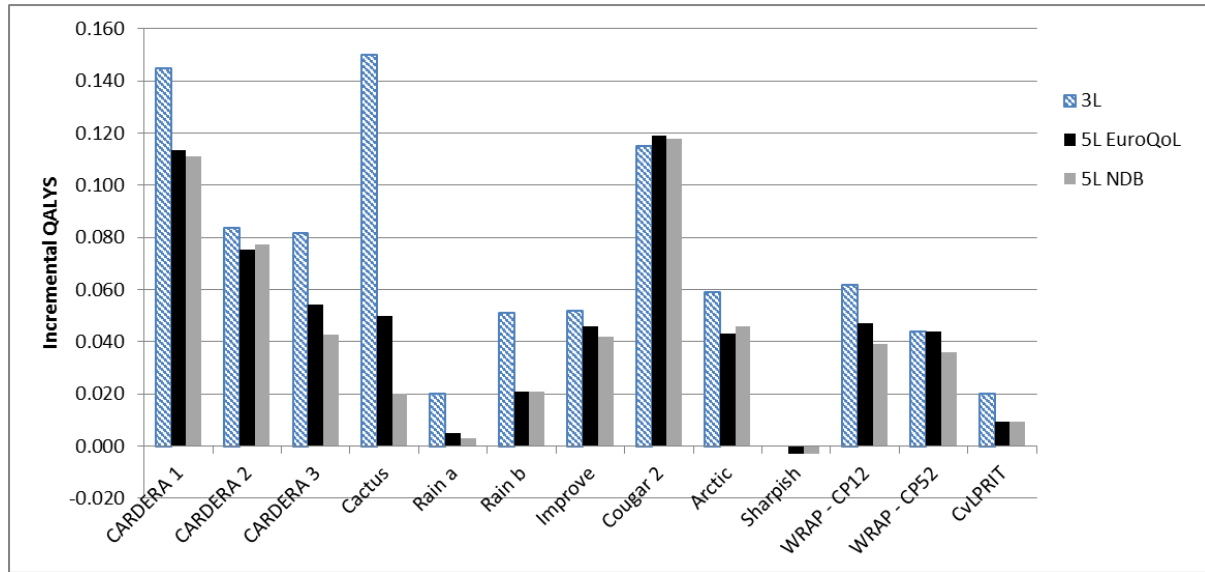


Figure 3: Histogram of incremental QALYs by 3L, 5L (EQG) and 5L (NDB) for all case studies



CARDERA 1 = MTX vs MTX + CS, CARDERA 2 = MTX vs MTX + PNS, CARDERA 3 = MTX + CS + PNS vs MTX

