Article:
Hernandez Alava, M., Wailoo, A., Grimm, S. et al. (2018) EQ-5D-5L versus EQ-5D-3L: the impact on cost-effectiveness in the United Kingdom. Value in Health, 21 (1). pp. 49-56. ISSN: 1098-3015

https://doi.org/10.1016/j.jval.2017.09.004

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

ABSTRACT 250 words (250 words max)

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 3L and 5Lfrom the same respondents. The EuroQoL dataset (n=3551) included patients with different diseases and a healthy cohort. The National Data Bank (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 UK 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 moving between instruments is not simply 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 effects[1]. Tariffs are anchored around 1 for full health and 0 for states considered equivalent to death, based on estimates from samples of the general population. For the 3L, these tariffs were based on a Time-Trade-Off (TTO) valuation method. In the UK, these tariffs range from 1 for full health to -0.594. 35% (84/ 243) of the health states are valued with a negative score. There is a gap between full health and the next level of impairment valued at 0.883. Tariffs for the 5L version are now available for England[2], Canada[3], Japan[4], Uruguay[5], the Netherlands[6] and Korea[7]. The valuation methods used for 5L used a combination of updated "lead time" TTO methods and Discrete Choice Experiments (DCE)[8]. In England, this has led to a smaller range of values (from 1 to -0.281), a smaller gap at the upper end of the distribution (0.951 is the next score after 1) and fewer values below 0 (5%, 153/3125).

EQ-5D is one of the most widely used instruments underpinning economic evaluations conducted in terms of cost-per Quality Adjusted Life Year (QALY), calculated from the tariff scores. It is therefore essential to understand the implications of 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/English tariffs. It should be noted that this 5L tariff could be subject to change as it progresses through the peer review process. We use two datasets where respondents 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 datasets.

The first was provided by the EuroQoL group (the EQG data). Between August 2009 and September 2010, the EuroQoL Group coordinated and partly funded a data collection study. Its main aim was to collect data on both versions of EQ-5D, the 3L and 5L, 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. 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, sampling methods ensured a wide range of severity across all the 5L and 3L dimensions.

The National Data Bank (NDB) for Rheumatic Diseases is a register of patients with rheumatoid disease, primarily recruited by referral from US and Canadian rheumatologists. Information supplied by participants is validated by direct reference to records held by hospitals and physicians (A minority of cases come by self-referral, with medical details obtained by NDB in the same way). Full details of the recruitment process are given by Wolfe and Michaud (2011)[9] . The EQ-5D responses and other patient-supplied data are collected by various means, primarily postal and web-based questionnaires completed directly by patients. Data collection began in 1998 and continues to the present, in waves administered in January and July of each year. In 2011, there was a switch from 3-level to the 5-level version of EQ-5D and both versions were included in the January 2011 wave. The NDB questionnaire is 27 pages long and it includes many general as well as RA specific questions. 5L and 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 3L and 5L[10]. Full details are provided

there. Responses to 3L and 5L are ordinal. The model reflects this in a system of 10 ordinal regressions, each of which is used to estimate the response level for one of the two versions of EQ-5D conditional on covariates (age and gender), arranged into the five health domains. The model reflects any tendency for an individual to give more or less positive responses across domains via a latent factor representing background response behaviour. A copula approach is used to specify the bivariate distribution of each 3L, 5L pair of responses. This captures the strong association between 3L and 5L responses within each health domain, without necessarily assuming that the strength of the association is the same in all parts of the health distribution – for example, someone who has experienced extreme pain may answer the pain questions in a more focused and coherent way than someone without experience of chronic pain. Five different copulas were examined (Gaussian, Clayton, Frank, Gumbel and Joe) which reflect different types and strengths of dependence at different parts of the distribution, with the data informing the most appropriate final choice of copula.

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.

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 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[11]. 3L was administered to patients at baseline, 6, 12, 18 and 24 months[12].
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[13]. 3L was completed at baseline, 3 and 8 months.
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. 3L was completed at 3 months.
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)[14]. 3L was administered at 3 and 12 months.
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[15] . Patients completed the EQ-5D at baseline, during clinic visits at weeks 3, 6, 9 and 12, then every 6 weeks for up to 1 year and then every 3 months until death.
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)[16,17]. It compared fludarabine, cyclophosphamide and rituximab (FCR), which is considered conventional frontline therapy, with fludarabine, cyclophosphamide, mitoxantrone and low dose rituximab (FCM-miniR). 3L was completed at baseline, after 3 cycles of therapy, at the end of therapy, 3 months after the end of therapy and then every 3 months after the end of therapy until 24 months post randomisation (i.e. at 6, 9, 12, 18 and 24 months post randomisation).

7) SHARPISH - The Self-Help and Relapse Prevention in Smoking for Health (SHARPISH) trial[18] sought to estimate the effectiveness and cost-effectiveness of self-help booklets versus a single leaflet to prevent smoking relapse in people who had stopped smoking for four weeks. 3L was administered at baseline, 2 months and 11 months post-randomisation.
8) WRAP – the Weight-Reduction Activity Programme (WRAP)[19] 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. 3L was administered at baseline, 3, 12 and 24 months.
9) CvLPRIT - The CvLPRIT (Complete- compared to Lesion-Only Revascularisation For Myocardial Infarction) trial[20] 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). 3L was administered immediately before discharge and at 12 months post-discharge.

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

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 3L and 5L and in the mobility and usual activities dimension of 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 (3L) and, pain/discomfort and anxiety/depression (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 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 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 3L and 5L. Within the EQG dataset, the distributional shapes for all dimensions of 5L are similar, displaying a decreasing profile corresponding to either shape (i) or (ii). The 3L distributions in the EQG dataset exhibit all four distributional shapes and appear more different across dimensions than in the 5 level version. The variation in shape highlights the need to use flexible model specifications which do not impose the same model structure across dimensions or datasets.

Figure 2 shows kernel estimates of the distributions of utility scores in both datasets. 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 5L, especially towards the top of the distribution. The number of individuals in full health is reduced by using 5L and the mode at the bottom of the distribution around the value of zero in the -3L distribution disappears in the distribution of 5L. The mean and median of 5L are higher than the corresponding mean and median of 3L in both datasets (see Table 1). The range of 5L is smaller as the worst state has a utility score of -0.281 compared to -0.594 of 3L.

Statistical model results

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. The Gaussian and Frank copulas are similar in that both allow for positive or negative dependence, symmetric in both tails, but the Frank form generates dependence weaker in the tails and stronger in the centre of the distribution. The Clayton copula allows only positive dependence, with strong left tail dependence and relatively weak right tail dependence; thus, if two variables are strongly correlated at low values but less so at high values, then the Clayton copula is a good choice. Therefore, in the EQG 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 3L and 5L in most dimensions. This is a test of the hypothesis that the underlying relationship between covariates and/or latent variable and EQ-5D is the same for 3L and 5L. Rejection of the hypothesis indicates 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, with one exception.

In COUGAR 02, there is 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 group[15]. 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 arm[15]) and did not outweigh the morbidity effect.

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

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

In those studies where the 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 EQ5D-3L to over £24k using 5L (EQG data) and over £30k using 5L (NDB data). CACTUS changes from a highly cost effective central estimate using 3L (£3058) to one that is more borderline (£23022) using 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 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). Consistent with our findings, a recent study that also used the EQG dataset reported that 5L leads to higher values overall and across all of the conditions health conditions in the EQG data[23].

In this sense, estimates of health gain from 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 observations given without recall of previous responses than the EQG studies. The NDB is also predominantly conducted in the English language. Whilst there is some evidence that the ranking of levels 4 and 5 ("severe" and "extreme" problems) may not be as expected in the English valuation study[8], this is less likely to be an issue affecting the descriptive system when respondents are provided with all five levels in their expected order. Therefore, we do not feel there is a rationale to prefer English speaking samples for the mapping work. Both datasets are also limited by their size and coverage of relevant health states. In the EQG data, only 119 of the possible 233 EQ-5D-3L utility values are observed. That figure is 83 for the NDB. We know that most of the 233 health states do appear in real patient records. For example, in UK 2010-2014 data for knee replacement procedures (n=320,000) we find 189 out of 233 possible utility values. There is a pressing need for well designed, large scale data collection to extend this work.

There are a number of implications for policy in the light of these results. Given the differences between 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. A single approach to mapping between 3L and 5L would aid consistent decision making. Additional data collection would also permit extended validation of the method and comparison against the EuroQoL "crosswalk" that provides a link between 5L responses and 3L[24]. Decision making bodies, like NICE in the UK, should endorse the use of either 3L or 5L and a set of methods that allow

evidence to be linked from one to the other. 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.

[1] Herdman M, Gudex C, Lloyd A, et al Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011; 20(10): 1727–1736.

[2] Devlin N, Shah K, Feng Y, Mulhern B, van Hout B.Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. OHE Research Paper 16/01. London: Office of Health Economics; 2016

[3] Xie F, Pullenayegum E, Gaebel K, Bansback N, Bryan S, Ohinmaa A, Poissant L, Johnson JA. A Time Trade-off-derived Value Set of the EQ-5D-5L for Canada. Med Care. 2016;54(1):98-105.

[4] Shiroiwa T, Ikeda S, Noto S, Igarashi A, Fukuda T, Saito S, Shimozuma K. Comparison of Value Set Based on DCE and/or TTO Data: Scoring for EQ-5D-5L Health States in Japan. Value Health. 2016 Jul-Aug;19(5):648-54.

[5] Augustovski F, Rey-Ares L, Irazola V, Garay OU, Gianneo O, Fern‡ndez G, Morales M, Gibbons L, Ramos-Go–i JM. An EQ-5D-5L value set based on Uruguayan population preferences. Qual Life Res. 2016;25(2):323-33. Erratum in: Qual Life Res. 2016 Feb;25(2):335.

[6] Versteegh MM, Vermeulen KM, Evers SM, de Wit GA, Prenger R, Stolk EA. Dutch Tariff for the Five-Level Version of EQ-5D. Value Health. 2016;19(4):343-52.

[7] Kim SH, Ahn J, Ock M, Shin S, Park J, Luo N, Jo MW. The EQ-5D-5L valuation study in Korea. Qual Life Res. 2016 Jul;25(7):1845-52.

[8] Feng Y, Devlin N, Shah K, et al. New Methods for Modelling EQ-5D-5L Value Sets: An Application to English Data, Office for Health Economics, available at: https://www.ohe.org/publications/new-methods-modelling-eq-5d-5l-value-sets-application-english-data (accessed on 23rd May 2017)

[9] Wolfe F and Michaud K. The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank. Rheumatology 2011;50:16-24.

[10] Hernandez Alava M, Pudney S. 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, Journal of Health Economics, in press.

[11] Choy EHS, Smith CM, Farewell V et al. Factorial randomised controlled trial of glucocorticoids and combination disease modifying drugs in early rheumatoid arthritis. Ann Rheum Dis 2008;67:65663.

[12] Wailoo A, Hernandez Alava M, Scott I, Ibrahim F, Scott D. Cost-effectiveness of treatment strategies using combination disease-modifying anti-rheumatic drugs and glucocorticoids in early rheumatoid arthritis. Rheumatology 2014

[13] 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.

[14] IMPROVE Trial Investigators. Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from IMPROVE randomised trial. BMJ 2014;348:f7661

[15] Ford HER, Marshall A, Bridgewater JA, et al 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

[16] Hillmen P, Milligan D, Schuh A et al. Results Of The Randomised Phase II NCRI Arctic (Attenuated dose Rituximab with ChemoTherapy In CLL) Trial Of Low Dose Rituximab In Previously Untreated CLL. Blood;2013, 122:1639

[17] Howard, DR, Munir, T, McParland, L et al. (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)

[18] 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 Institude for Health Research, 2015.

[19] 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

[20] Gershlick AH, Khan JN, Kelly DJ et al. "Randomized trial of complete versus lesion-only revascularization in patients undergoing primary percutaneous coronary intervention for STEMI and multivessel disease: the CvLPRIT trial." Journal of the American College of Cardiology 65.10 (2015): 963-972.

[21] Dolan, P. (1997). Modeling valuations for EuroQol health states. Medical Care, 35:1095–1108.

[22] 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.

[23] Mulhern B, Feng Y, Shah K, et al. Comparing the UK EQ-5D-3L and the English EQ-5D-5L Value Sets. OHE Research Paper 17/02. Available at https://www.ohe.org/publications/comparing-uk-eq-5d-3l-and-english-eq-5d-5l-value-sets (last accessed 20th June 2017)

[24] 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, Vol. 15: 708-15.

TABLES:

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

	EQG sample		**NDB sample**	
AGE				
Mean	51.23		63.32	
[95% confidence Interval]	[50.57, 51.89]		[62.99, 63.65]	
Median	54		64.13	
[95% confidence Interval]	[54, 56]		[63.78, 64.46]	
Standard Deviation	20.11		12.31	
Minimum	13		16.66	
Maximum	99		95.20	
Proportion female	0.53		0.81	
	EQ-5D-3L	***EQ-5D-5L***	***EQ-5D-3L***	***EQ-5D-5L***
UTILITY				
Mean	0.628	0.712	0.681	0.779
[95% confidence Interval]	[0.617, 0.639]	[0.703, 0.722]	[0.674, 0.688]	[0.773, 0.784]
Median	0.691	0.802	0.725	0.823
[95% confidence Interval]	[0.691, 0.725]	[0.792, 0.816]	[0.725, 0.727]	[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		
Mobility		
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***
Self-care		
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
Usual activities		
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**
Pain/discomfort		
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***

Anxiety/depression		
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					ICER				
	3L	**5L EQG**	**% change**	**5L NDB**	**% change**	**3L**	**5L EQG**	***% change***	**5L NDB**	***% change***
CARDERA 1	0.145	0.113	*-21.8%*	0.111	*-23.2%*	4648	5940	*27.8%*	6054	*30.3%*
CARDERA 2	0.084	0.075	*-10.4%*	0.077	*-8.0%*	13666	15252	*11.6%*	14846	*8.6%*
CARDERA 3	0.082	0.054	*-33.5%*	0.043	*-47.6%*	15929	23940	*50.3%*	30418	*91.0%*
Cactus	0.150	0.050	*-66.7%*	0.020	*-86.7%*	3058	9481	*210.0%*	23022	*652.8%*
Rain a	0.020	0.005	*-75.0%*	0.003	*-85.0%*	184700	738800	*300.0%*	1231333	*566.7%*
Rain b	0.051	0.021	*-58.8%*	0.021	*-58.8%*	294137	714333	*142.9%*	714333	*142.9%*
Improve	0.052	0.046	*-11.5%*	0.042	*-19.2%*	-44617*	-48113	*7.8%*	-54742	*22.7%*
Cougar 2	0.115	0.119	*3.5%*	0.118	*2.6%*	27180	26434	*-2.7%*	26484	*-2.6%*
Arctic	0.059	0.043	*-27.1%*	0.046	*-22.0%*	112193	162774	*45.1%*	152130	*35.6%*
Sharpish	0.000	-0.003	*NA*	-0.003	*NA*	NA**				
WRAP - CP12	0.062	0.047	*-23.7%*	0.039	*-36.2%*	1812	2373	*31.0%*	2840	*56.7%*
WRAP - CP52	0.044	0.044	*0.0%*	0.036	*-19.0%*	4305	4312	*0.2%*	5316	*23.5%*
CvLPRIT	0.020	0.010	*-52.5%*	0.009	*-53.0%*	21496	46761	*117.5%*	47521	*121.1%*

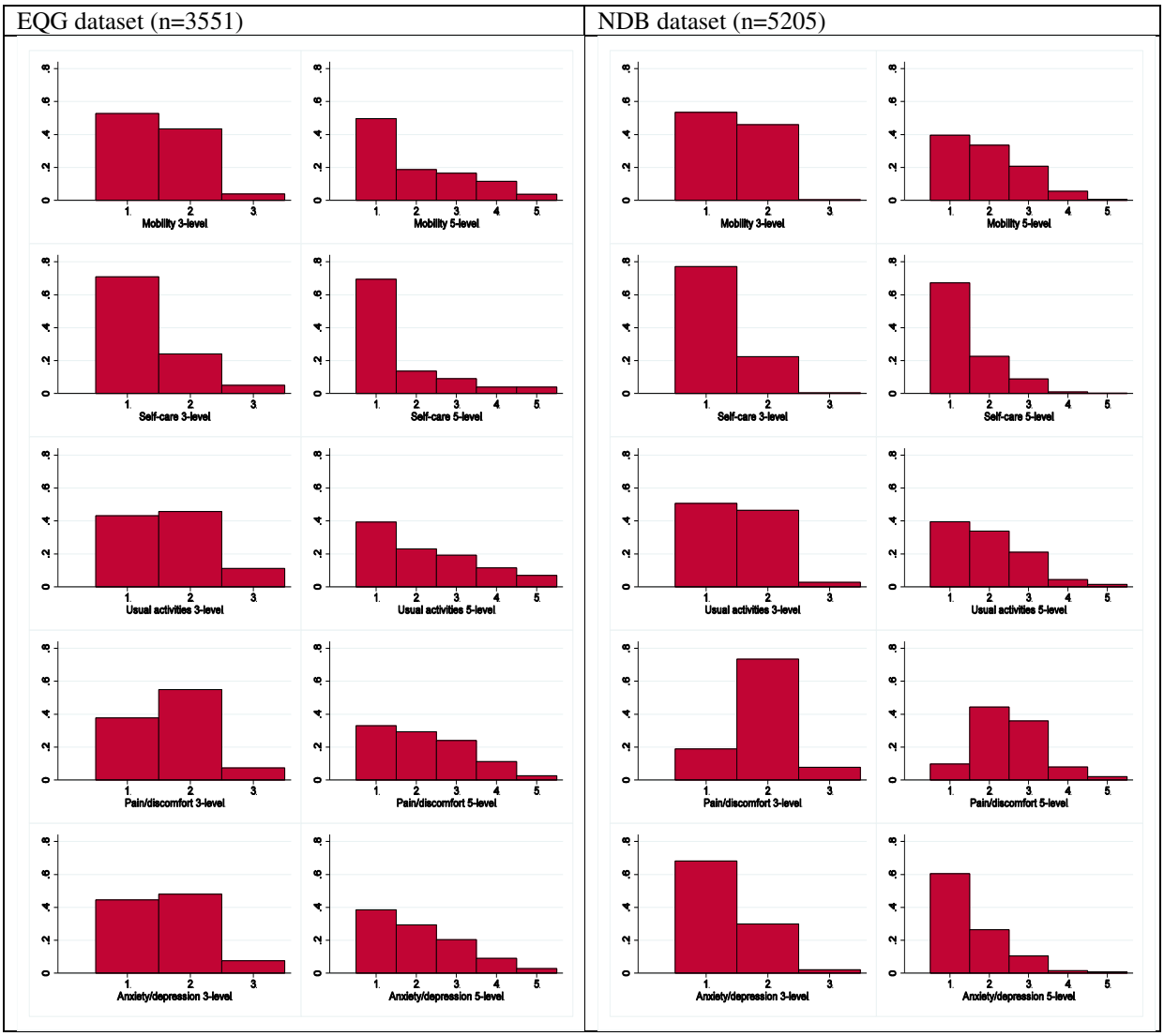
CARDERA 1 = MTX vs MTX + CS, CARDERA 2 = MTX vs MTX + PNS, CARDERA 3 = MTX + CS + PNS vs MTX

* 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



For the 3L, level 1=”no problems”, 2=”Some problems”, 3=”extreme problems/unable to do”.
For the 5L, level 1 = “no problems”, 2=”slight problems”, 3 = “moderate problems”, 4 = “severe problems”, 5 = “extreme problems / unable to do”

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

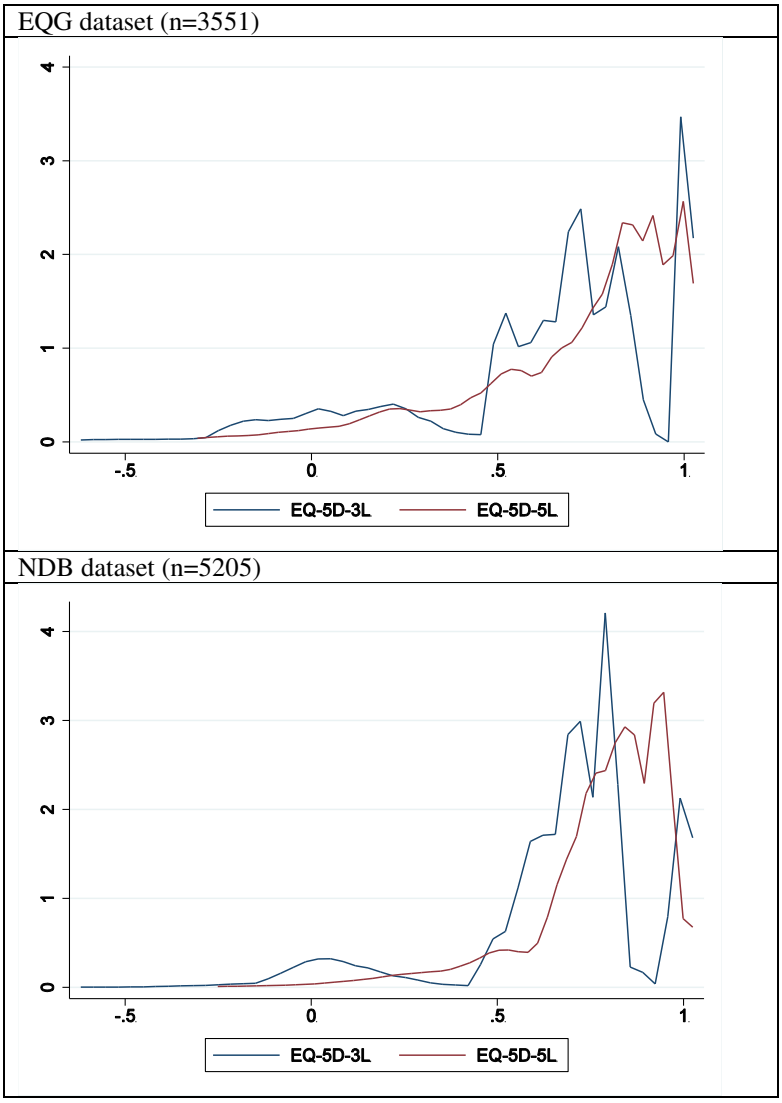
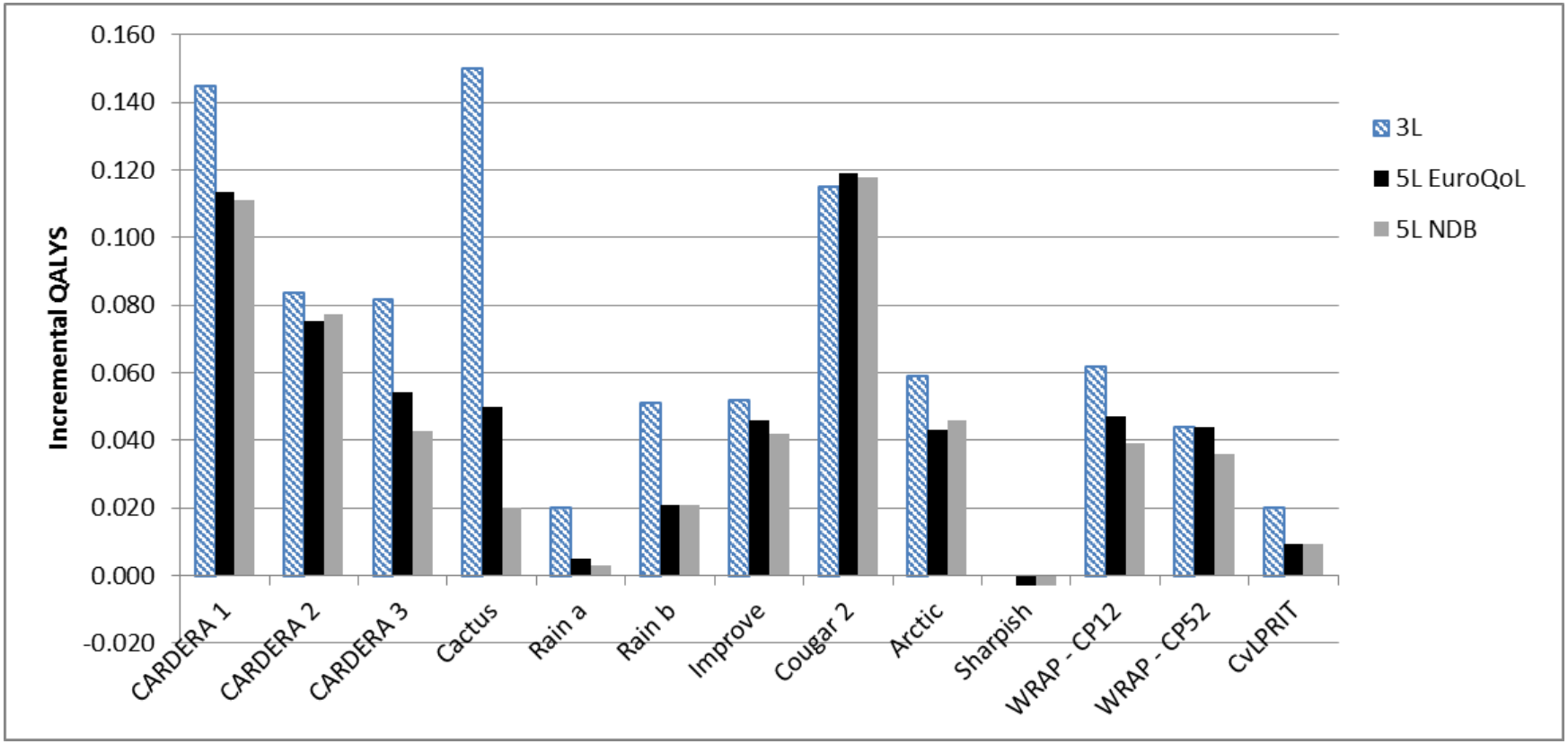


Figure 3: Histogram of incremental QALYs by 3L, 5L (EQG) and 5L (NDB) for all case studies



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