COMPARING THE EQ-5D-3L AND 5L VERSIONS. WHAT ARE THE IMPLICATIONS FOR COST EFFECTIVENESS ESTIMATES?

REPORT BY THE DECISION SUPPORT UNIT

13th March 2017

Allan Wailoo[1], Monica Hernandez Alava[1], Sabine Grimm[2], Stephen Pudney[1], Manuel Gomes[3], Zia Sadique[3], David Meads[4], John O'Dwyer[4], Garry Barton[5], Lisa Irvine[5]

[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] Leeds Institute of Health Sciences, University of Leeds, UK

[5] Norwich Medical School, University of East Anglia, UK

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

Acknowledgements

We wish to thank the following people: Fred Wolfe and Kaleb Michaud for providing data from the NDB; the EuroQoL group also 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.

EXECUTIVE SUMMARY

Background

The NICE Guide to the Methods of Technology Appraisal expresses a preference for using the EQ-5D for adult populations to estimate the health related quality of life in adults which, in turn, are used to calculate the impact of different technologies in terms of Quality Adjusted Life Years (QALYs).

The EQ-5D comprises 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). This report is intended to provide information on how using 5L instead of 3L is likely to affect the results of economic evaluations, and highlight the implications of the findings for NICE.

Estimating the relationship between EQ-5D-3L and 5L

We used two reference datasets where patients filled in both 3L and 5L instruments. One was supplied by the EuroQoL group (EQG). Questionnaires were administered in six countries and included eight broad patient groups plus a healthy student population (n=3691). The second was provided by the National Databank for Rheumatic Diseases (NDB) from the January 2011 wave of questionnaires to patients of rheumatologists in the US and Canada (n=5311).

Our aim was to estimate the joint distribution of the responses to the two versions of EQ-5D, conditional only on age and gender, to provide a general model that could be applied widely. A flexible model has previously been developed by two of the co-authors (MH and SP) for mapping between 3L and 5L. 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 where the relationships between the two versions of EQ-5D are estimated jointly, so that mapping can, in principle be made consistently in either direction. Our implementation of this approach is based on much less restrictive assumptions than linear regression and its

extensions, and can be expected to be less vulnerable to specification error bias. The model was estimated using the EQG dataset and the NDB dataset but excluding all rheumatology specific outcomes as covariates, thus making the mapping usable in any patient group. The dependence between responses to the two variants of EQ-5D in each dimension was captured with a copula representation. Copulas are very useful as they can generate a number of dependence structures. We assessed five different copulas in the analysis.

In the final models, there were 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 case studies

Nine cost-effectiveness studies conducted alongside clinical trials were used as case studies. Each had existing analyses based on patient completion of the EQ-5D-3L instrument. In each case, we used the copula models to generate a revised analysis based on estimated 5L scores. We compared directly-observed 3L and estimated 5L (EQG and NDB) results.

The 5L instrument and associated tariff has the effect of shifting mean utility scores further up the utility scale towards full health, and compresses them into a smaller space. Thus, improvements in quality of life tend to be valued less using 5L than equivalent changes measured with 3L. In almost all cases, this means that a switch from 3L to 5L causes a decrease in the incremental QALY gain from effective health technologies and therefore technologies appear less cost-effective. This is true whether the estimation of 5L is based on EQG or NDB data. However, an important exception is where life extension is a substantial element of health gain, the ICER can reduce rather than increase.

Estimated incremental QALY gains reduced by up to 75% when moving from 3L to 5L (EQG dataset) or 87% (NDB dataset).

Discussion

The 3L and 5L versions of EQ-5D produce substantially different estimates of cost effectiveness. Improvement in quality of life will be measured as a greater health utility gain

with 3L than the same change using the 5L. This is because of the combined effect of differences in the way individuals respond to the changed descriptive system and the changed valuation system, compared to 3L. In this sense, 3L and 5L are not consistent with each other.

5L is already being used as the descriptive system in many ongoing clinical studies. Yet 3L will remain part of the relevant evidence base for many years, perhaps decades. This raises several challenges for decision-making, particularly where there is a need to ensure consistency between appraisals.

The use of either 3L or 5L with no adjustment to either, as if they were interchangeable, is not appropriate. Nor is there a simple proportional adjustment that can be made between 3L and 5L. Changes do not happen equally across the distribution of health and therefore different technologies are affected to different degrees by the shift from one instrument to another.

It is feasible to reliably adjust 3L evidence to 5L equivalent values, as has been done in this report. Whilst the model also allows translation of 5L to 3L, the performance is worse. There are also significant differences in utility estimates according to whether we estimate the expected 5L score using data from the EQG or from the NDB. Those differences were even more pronounced when we incorporated disease specific covariates to further improve the mapping model. This raises the possibility that future mapping between the instruments may be best performed using estimates based on disease-specific datasets, rather than a single generic mapping.

These findings have implications for recommendations NICE may make about its willingness to accept unadjusted utility values from the different EQ-5D instruments, how it may wish to specify any adjustments be made, and the cost-effectiveness threshold.

CONTENTS

TABLES

FIGURES

1. INTRODUCTION

The NICE Guide to the Methods of Technology Appraisal[1] expresses a preference for using the EQ-5D for adult populations to estimate the health related quality of life in adults. These estimates are used to calculate the impact of different technologies in terms of Quality Adjusted Life Years (QALYs) (NICE Methods Guide, 2013, 5.3.1). A single instrument is preferred by NICE in most situations because of the need to make decisions that are consistent across technologies, patient groups and disease areas. NICE recognises that different preference based instruments lead to different estimates of health utility and therefore one approach (EQ-5D) is recommended for the reference case. The guide uses the term "EQ5D" as shorthand for the 3 level version of the instrument (EQ5D-3L) as described in section 5.3.6.

The EQ-5D comprises 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). It is the 3L version that is the main focus for the NICE methods guide and submissions to date. However, the EuroQoL group have developed a new, five level 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). The 5L was produced with the intention of improving the instrument's sensitivity and reducing ceiling effects[2]. The NICE Methods Guide was written at a time when the descriptive system of the 5L instrument was available but no separate valuation had reported. The guide states:

> *"The EQ-5D-5L may be used for reference-case analyses. The descriptive system for the EQ-5D-5L has been validated, but no valuation set to derive utilities currently exists. Until an acceptable valuation set for the EQ-5D-5L is available, the validated mapping function to derive utility values for the EQ-5D-5L from the existing EQ-5D (-3L) may be used"(5.3.12)*

There is now an English valuation set for the EQ-5D-5L[3] (ref Devlin et al 2016). Whilst the 2013 Methods Guide implies that both 5L and 3L may be acceptable as reference case

analyses, little is known of the implications of such a decision or whether alternative approaches have merit.

The purpose of this report is to provide information on the likely implications of conducting analyses using 5L compared to 3L. Specifically, we use data from two separate studies where individuals completed both the 3L and 5L instruments simultaneously. We compare those responses statistically and then consider the implied differences in estimated health utility scores given the associated tariffs for the 3L and 5L. We develop statistical models that transform observed 3L responses to estimated 5L ones, and vice versa. Using these models, we apply the results in a series of case study cost-effectiveness analyses.

2. METHODS AND DATA

2.1. ESTIMATING THE RELATIONSHIP BETWEEN EQ-5D-3L AND 5L

Hernandez and Pudney[4] have previously developed a flexible model which allows analysis of the joint responses to EQ-5D-3L and EQ-5D-5L. After estimation, the model can be used for mapping between the health descriptions provided by the two instruments. The advantage of estimating a joint model is that this supports consistent mapping both ways, from the 3- to the 5-level version and vice versa. The original model was estimated using the National Data Bank for Rheumatic Diseases (NDB), a register of patients with rheumatic disease in the US and Canada.

The underlying model is a system of ordinal regressions with a flexible copula mixture residual distribution. Copulas are multivariate probability distributions. Their use here is based on the concept that an individual's responses to the 3L and 5L, within each of the five dimensions of health, will be correlated but that the degree and form of that correlation may vary across the spectrum of disease severity. Several different types of copula were tested with the preferred type for each health dimension being determined by the data.

The model uses an underlying latent factor, a means of joining the model across all health dimensions. This recognises that for each respondent, the responses given on each dimension of health will be correlated to each other.

Finally, the mixture approach allows error terms to be non-normal. In practice, this is important because it overcomes the problem of misspecification and leads to estimates that are much less likely to be biased.

Hernandez and Pudney's original work includes Rheumatoid Arthritis (RA) specific variables as covariates (for example the Health Assessment Questionnaire measure of functional disability) and therefore cannot be used for mapping between the different versions of EQ-5D in a non-RA disease area. Consequently, the model developed in Hernandez and Pudney was re-estimated here with two alternative approaches: a) using a different dataset covering a range of diseases across six different countries and b) using the NDB dataset but excluding all RA specific variables.

2.2. DATA

2.2.1. EuroQoL Group coordinated study (EQG)

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. A copy of the questionnaire used in Scotland can be found in the Appendix. 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. This dataset was used to develop a "crosswalk" between the EQ-5D-3L value set and the EQ-5D-5L descriptive system providing an interim EQ-5D-5L value set[5].

Published information on the data collection can be found in Janssen et al (2013)[6] and van Hout et al (2012)[5].

2.2.2. *The NDB dataset*

The 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)[7]. 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.

2.2.3. *Comparisons of the datasets*

The EQG and NDB datasets contain a total of 3691 and 5311 respondents respectively. Missing values on the analysis variables, 140 observations (3.79%) in the EQG and 106 observations (2%) in the NDB, leave final estimation samples of 3551 and 5205 respondents in the EQG and NDB datasets respectively.

In this section the distribution of age and gender, the EQ-5D responses, and the utility scores are compared across the datasets to establish their similarities and differences.

Table 1 compares the distribution of age across both samples. The EQG sample is younger with an average age of 51 versus 63 in the NDB sample and covers a larger age range. There is a big difference in the proportion of females in the samples. The EQG sample includes 53% of females whereas in the NDB dataset the proportion of females is much larger, 81%, reflecting the nature of rheumatic diseases.

Table 1: Descriptive statistics of age (years) in the EQG and NDB estimation samples

	EQG sample	NDB sample
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

2.2.3.1. EQ-5D response distributions

Figure 1 shows histograms of the response distributions for each dimension of the 3- and 5-level versions of EQ-5D in both datasets. There are differences both across the dimensions and between the datasets[1]. 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.

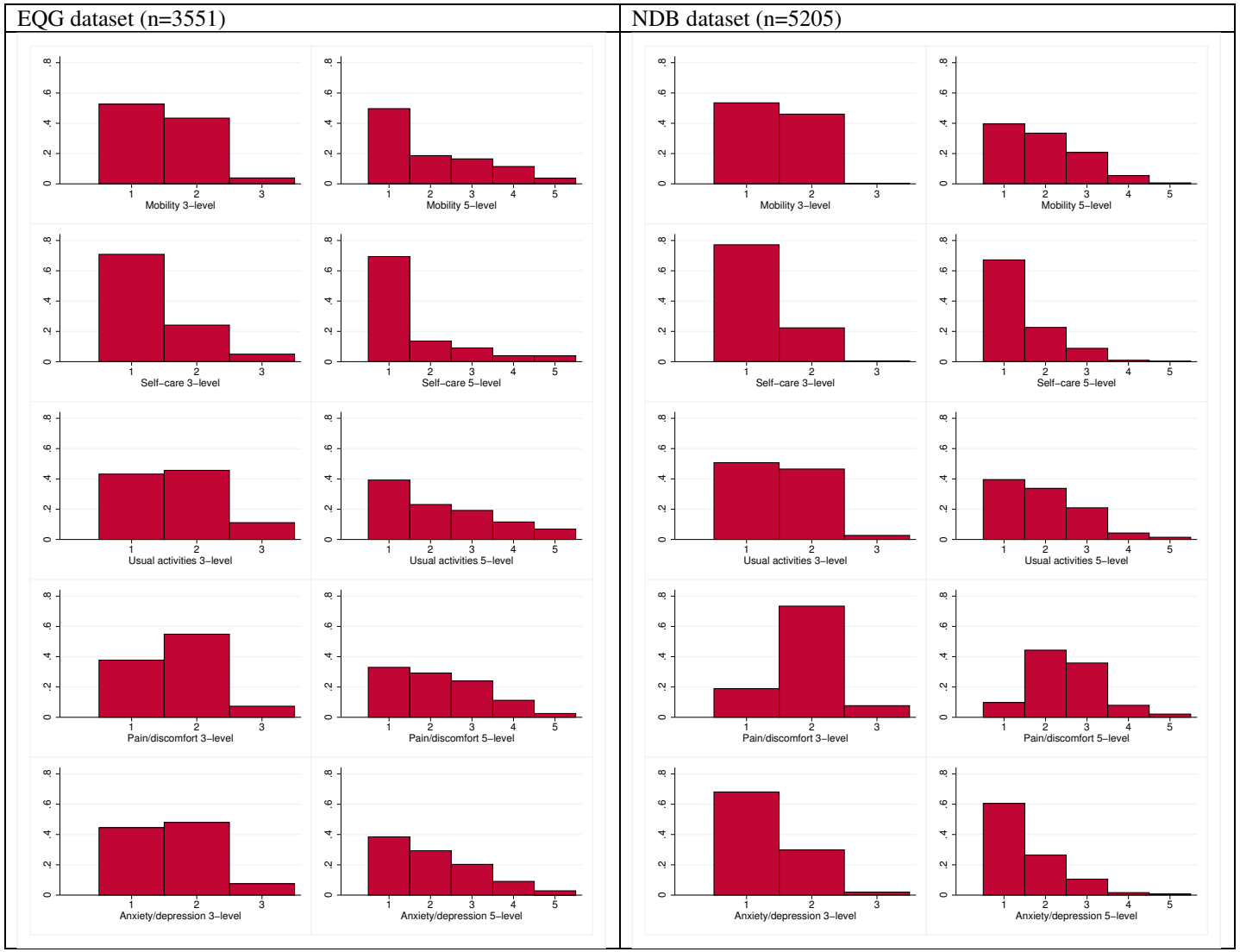
In the NDB dataset, both versions of EQ-5D display the same pattern within each dimension, but different shapes across dimensions: shape (i) in both the self-care and anxiety/depression dimensions, shape (ii) in the mobility and usual activities dimension and shape (iii) in the

[1] Note that standard statistical tests for equality of distributional forms across the NDB and EQG samples are not applicable since the datasets were not drawn by random sampling.

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.

The contrast in empirical distributions between the two datasets is not surprising. NDB relates to a population relatively homogeneous in age and medical condition, whereas EQG is heterogeneous in nationality and demographic and health characteristics.

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



2.2.3.2. Utility scores distributions

We use the value sets produced by Dolan (1997) and Devlin et al. (2016) for the 3L and 5L versions of EQ-5D. Figure 2 shows kernel estimates of the distributions of utility scores in both datasets. EQ-5D-3L in both datasets exhibit the typical characteristics documented in the literature. The 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 2). 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. The mean and median utility values are higher in the NBD dataset for both versions of EQ-5D indicating that the EQG sample has lower average health than the NDB. Both datasets span the full range of EQ-5D-3L but only EQG spans the full range of the EQ-5D-5L. The NDB dataset, being disease specific covers a lower proportion of different health states compared to the EQG dataset (see Table 2). Table T3 summarises the health state values for the two versions of EQ-5D in terms of their correlation with each other and with age and gender. They show high correlation between both versions of EQ-5D with the correlation in the EQG dataset being higher. This could be a consequence of the questionnaire design as the two versions of EQ-5D in the EQG questionnaire were very close together, only separated by some demographic questions. For each version of EQ-5D, both datasets display the same correlations between age, gender, and the utility values.

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

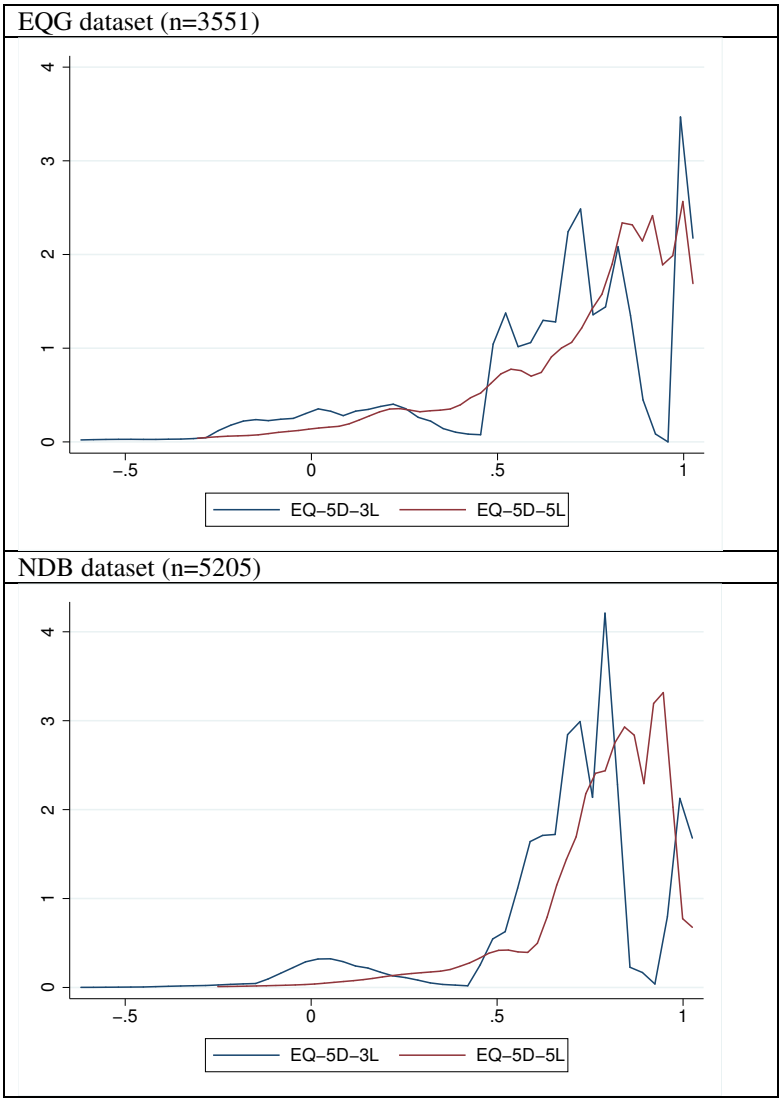


Table 2: Descriptive statistics of the utility values of EQ-5D-3L and EQ-5D-5L in the EQG and NDB datasets

	EQG dataset		***NDB dataset***	
	EQ-5D-3L	***EQ-5D-5L***	***EQ-5D-3L***	***EQ-5D-5L***
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	.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 3: Spearman correlations of EQ-5D-3L and EQ-5D-5L in the EQG and NDB datasets

	EQG dataset		NDB dataset	
	EQ-5D-3L	**EQ-5D-5L**	**EQ-5D-3L**	**EQ-5D-5L**
EQ-5D-3L	1	0.911	1	0.845
EQ-5D-5L	0.911	1	0.845	1
Female	-0.051	-0.072	-0.051	-0.072
Age	0.035	0.061	0.035	0.061

2.2.4. Model

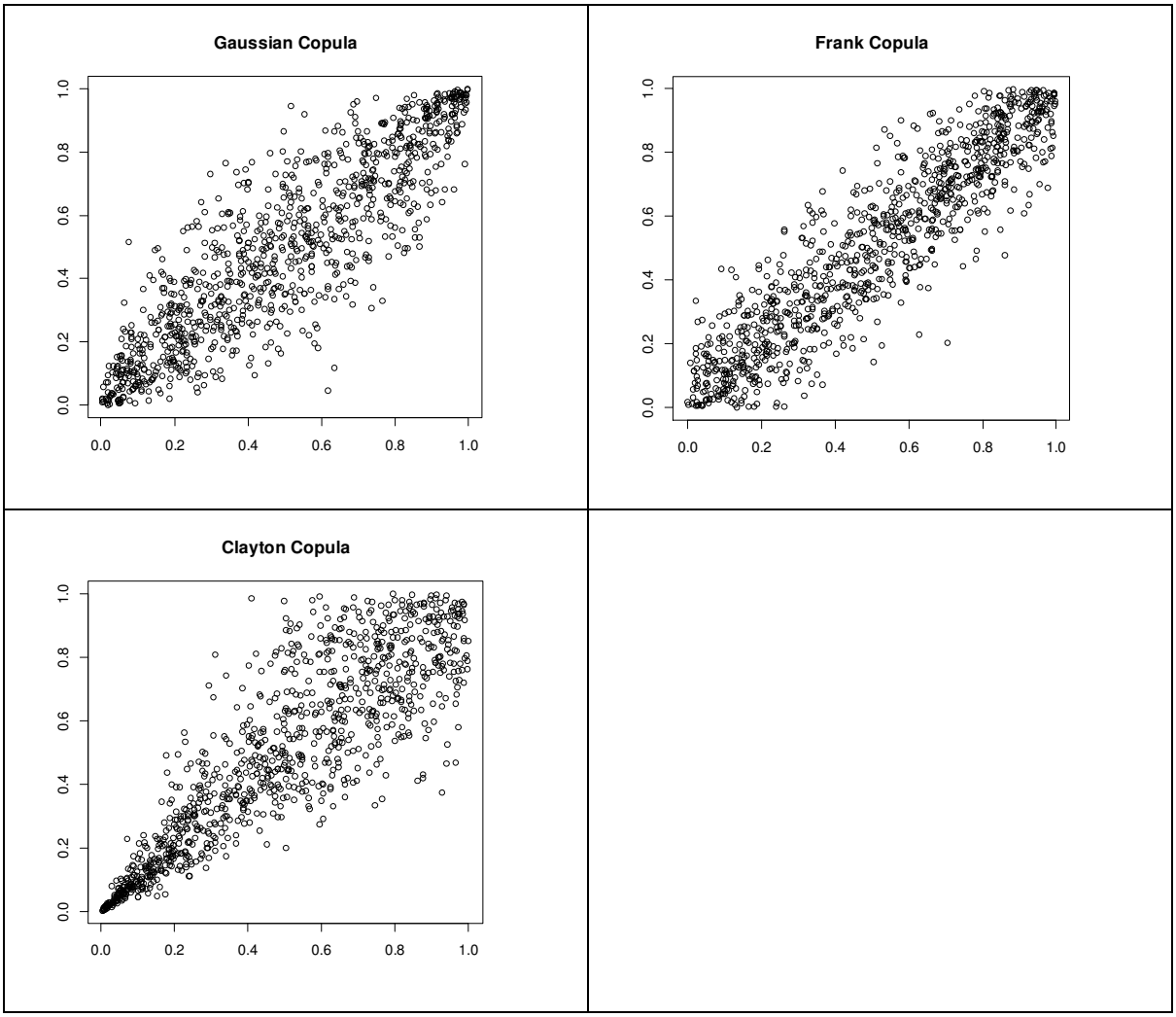
Hernandez and Pudney[4] developed a flexible model which allows mapping between EQ-5D-3L and EQ-5D-5L. The model is a system of ordinal regressions estimated jointly, incorporating a flexible copula mixture residual distribution. It is a type of response mapping model with all equations for the five 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. Five different copulas, Gaussian, Clayton, Frank, Gumbel and Joe, were assessed in the analysis. The Gaussian and the Frank copulas are similar in the sense that both of them allow for positive and negative dependence and dependence is symmetric in both tails. However, compared to the Gaussian copula, the Frank copula exhibits weaker dependence in the tails and dependence is strongest in the middle of the distribution. In contrast, the Clayton, Gumbel and Joe copulas allow only for positive dependence, and dependence in the tails is asymmetric. The Clayton copula exhibits strong left tail dependence and relatively weak right tail dependence. Thus, if two variables are strongly correlated at low values but not so correlated at high values, then the Clayton copula is a good choice. The Gumbel and Joe copulas display the opposite pattern with weak left tail dependence and strong right tail dependence. The right tail dependence is stronger in the Joe copula than in the Gumbel and thus the Joe copula is closer to the opposite of the Clayton copula.

Figure 3 illustrates the differences between the Gaussian, the Frank and the Clayton copulas. The parameters of the copulas are chosen in such a way that they all have approximately a

Kendall's rank correlation ("Kendall's tau") equal to 0.7. Comparing the Gaussian with the Frank copula, we see that the points are tighter together in the middle of the distribution in the Frank copula than in the Gaussian copula, and thus the dependence looks stronger in the middle of the distribution. In the tails, we can see that the points are closer together in the Gaussian copula than in the Frank copula, highlighting the weaker dependence in the tails of the Frank copula. Both the Gaussian and the Frank copula exhibit symmetric dependence, but the Clayton copula displays asymmetric dependence on the tails. Dependence is very strong on the left tail (all the points are tightly packed together) but there is very weak dependence on the right tail (the points are widely scattered).

Figure 3: Scatter plots of 1000 simulated draws from the Gaussian, Frank and Clayton copulas (Kendall's tau = 0.7)



Given the differences across the shapes of the distributions in the dimensions of EQ-5D depicted in Figure 1, it is expected that different copulas would be suited to different EQ-5D dimensions and to different datasets.

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.

3. MODELLING RESULTS

Our aim was to estimate the joint distribution of the responses to the two versions of EQ-5D conditional only on age and gender to provide a general model which can be applied widely. Additional disease specific variables in the model such as those used in Hernandez and Pudney (two-way mapping model) can generate improvements in model fit but at the cost of general applicability.

The best fitting joint models for each of the datasets are presented and discussed below. Our 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. Appendix Table 1 presents the full estimation results.

Table 4 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 to that estimated in the EQG dataset. In contrast, the Gaussian copula in the mobility, usual activities and pain/discomfort dimensions indicates 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 (see Figure 3).

There are also 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 a uniform realignment of the response levels. The only exceptions to this are in the anxiety/depression dimension (in both datasets) and in the self-care dimension (in the NDB dataset).

These differences, both in distributional form and in coefficients, cannot be interpreted in a simple unambiguous way, because the two datasets differ so much in their design and

empirical pattern. The NDB dataset is close to a census of the set of people with arthritic disease who are registered with certain branches of the health services in the USA and Canada, whereas the EQG is an assemblage of country-specific convenience samples most of which are targeted informally towards patients severely affected by specific diseases of various kinds. Compared to the NDB sample, the average severity reported by EQG respondents is substantially higher for anxiety and depression, moderately higher for self-care and usual activities, slightly higher for mobility, but substantially lower for pain. The difference in the optimal choices of copula mirrors this to some extent.[2] For example, the upper tail of the pain distribution is particularly salient for patients with arthritis, and for the NDB sample the Gaussian copula gives stronger tail dependence than the Frank copula does for the EQG sample. The statistical significance of coefficient differences follows a similar pattern, with particularly significant differences for coefficients in the pain, mobility and usual activities domains for the NDB sample, but less so for the EQG sample.

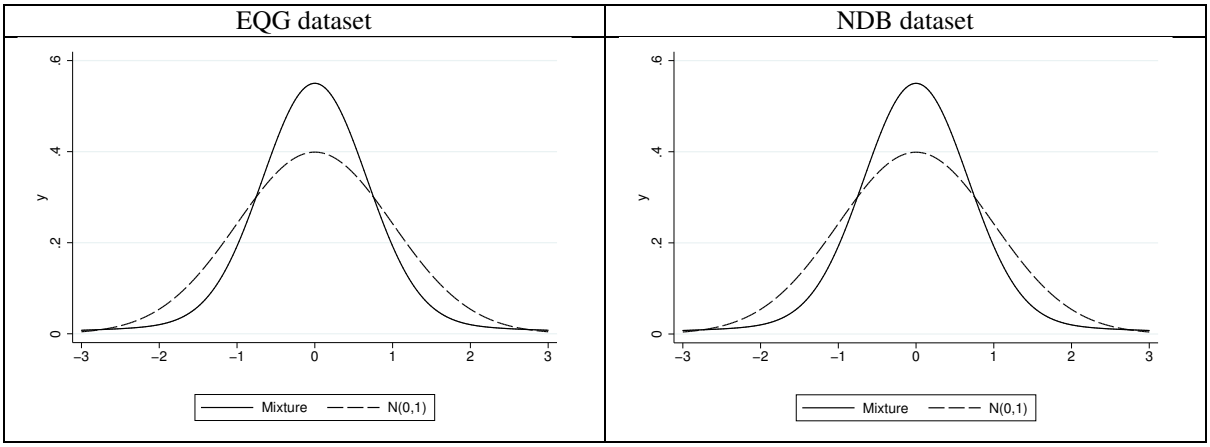
In both models, normality is rejected in favour of a single mixture distribution of two normal components. Figure 4 plots the mixture distributions for both models, compared with the standard normal distribution. The estimated mixtures for both datasets are very similar. They are composed of a dominant component just above zero and a second smaller and much more dispersed component centred around negative values. Therefore the residual mixture components in both datasets exhibit a much bigger central mode than the normal distribution and a slightly heavier left tail.

[2] The sample mean values of the 5L domain responses are: mobility 1.94 (NDB), 2.01 (EQG); self-care 1.45 (NDB), 1.60 (EQG); usual activities 1.95 (NDB), 2.24 (EQG); pain 2.49 (NDB), 2.21 (EQG); anxiety/depression 1.56 (NDB) 2.08 (EQG).

Table 4: 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%		

Figure 4: Residual distributions for the EQG and the NDB based models



4. COST EFFECTIVENESS CASE STUDIES

4.1. METHODS AND CASE STUDY DESCRIPTIONS

The copula models allow the prediction of EQ-5D-5L (responses and utility) from EQ-5D-3L responses (or vice versa). To better understand the likely impact of using 5L in cost effectiveness analysis, either instead of or alongside 3L, we used the copula mapping models in nine case studies.

The case studies were economic evaluations based on individual patient level data using EQ-5D-3L. Ideally, we would use case studies that represent the same balance of technologies and disease areas that reflect the NICE Technology Appraisals work programme. However, NICE TA submissions are almost exclusively model-based analyses with no link to utility data at the individual level. 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. A bespoke Stata command developed by Hernandez and Pudney[8] was provided to all collaborators.

This pragmatic approach ensured that no patient level data were transferred outside the units that conducted the original analyses, thus avoiding any difficulties with research ethics.

4.1.1. 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[9]. The trial enrolled 467 adult patients with active RA of <24-months duration. Patients were followed up for 2 years. EQ-5D-3L was administered to patients at baseline, 6, 12, 18 and 24 months. Detailed resource use data relating to RA were collected in 6-month blocks at months 6, 12, 18 and 24. A within trial economic evaluation was performed[10]. Since there were four treatment options in this trial, we present cost-effectiveness results comparing methotrexate monotherapy to each of the three combination strategies.

4.1.2. 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[11]. CACTUS was a single-blind, parallel-group, stratified, pilot randomized controlled trial in which thirty-four participants with aphasia were randomized, in a UK setting (17 to each arm).

Aphasia occurs in one third of people who survive stroke. The majority of recovery occurs in the first six months after the stroke, but continued treatment for a prolonged period (greater than 6 months) is associated with continued improvement in regaining language skills. Due to the costs associated with treatment, particularly face-to-face treatment with a therapist, continued treatment is often restricted in practice. Self-managed computer therapy in addition to usual stimulation can provide targeted therapy based on individual needs.

A 5-month intervention period was followed by a 3-month period without intervention to explore whether the treatment effect was maintained. Participants were included in the study if they had a diagnosis of stroke and aphasia with word finding difficulties and if they were no longer receiving impairment based speech and language therapy. Patients filled in the 3-level EQ-5D questionnaire on three occasions: at baseline, 5 months and 8 months into the trial. Because of the reading difficulties that aphasia patients may experience, a patient accessible version (based on pictures) of the EQ-5D-3L questionnaire was used instead of the standard EQ-5D questionnaire.

For the cost utility analysis, a model-based approach was taken and costs and health outcomes extrapolated beyond the trial period. Three health states were modelled: the initial aphasia state, a response state (from which patients could also relapse to the aphasia state), and death. An increase of 17% in the percentage of words a patient was able to name correctly was classed as a good response (based on what was achieved on average in the experimental group). None of the patients in the control group achieved a good response.

We applied the 3L to 5L mapping functions to the patient level EQ-5D responses to generate a 5L utility score for each patient. These 5L utility scores were fed back into the model, where they were used to calculate a mean utility score for the control group, and an increase

in utility in the good response state, which is added to the control group utility score for patients in the good response state.

4.1.3. RAIN

Acute traumatic brain injury (TBI) is a major cause of death, disability and cost to society. There is some evidence that management of acute TBI at specialist neuroscience centres is associated with improved outcomes for patients compared with management at non-neuroscience centres, but it is unknown whether adult TBI patients without an acute 'neurosurgical' lesion would benefit from 'early' transfer to a neuroscience centre. Early transfer may pose a risk of death to the patient, but 'late' transfer may result in the patient developing critical lesions, and subsequently be at increased risk of death during the later transfer.

The objectives of the Risk Adjustment in Neurocritical care (RAIN) trial[12] were to compare the effectiveness, costs, and cost-effectiveness of:

1. Management in a dedicated neurocritical care unit versus a combined neuro/general critical care unit, and;
2. '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.

In RAIN, for patients admitted to neuroscience centres, care within a dedicated neurocritical care unit was compared with care within a combined neuro/general critical care unit (n=1,324 vs n=1,341). Secondly, for patients who originally presented at a non-neuroscience centre, an 'early' transfer group was defined as those patients who transferred to a neuroscience centre within 18 hours of initial hospital presentation (n=584). The 'no or late' transfer group were defined as patients who received all their critical care within a non-neuroscience centre, and those who transferred to a neuroscience centre more than 24 hours after initial hospital presentation (n=263). At six months follow up patients completed a 3-level EQ-5D questionnaire. The EQ-5D-3L profiles were combined with health state preference values from the UK general population. QALYs were then reported by combining data on vital status and utility score at six months. Mean differences after adjustment with linear regression were reported for comparison of QALYs.

4.1.4. IMPROVE

The IMPROVE study investigated longer-term outcomes following either endovascular repair or open repair of ruptured abdominal aortic aneurysm (AAA). Ruptured abdominal aortic aneurysm is fatal in over 80% of cases and operative mortality remains high in those who survive repair (42%)[13]. The majority of patients in Europe, the USA, and elsewhere are treated with open surgical repair rather than the less invasive endovascular aneurysm repair (EVAR)[13,14,15,16]. Emergency EVAR is not always available in many centres.

The IMPROVE trial[17] randomised patients at the point of in-hospital clinical diagnosis to an endovascular strategy (EVAR wherever possible, with open repair for those anatomically unsuitable for standard EVAR) vs. open repair. The aim of this trial was to assess the clinical and cost effectiveness of a preferential endovascular strategy for the management of suspected ruptured AAA. The primary outcome was survival at 30 days after randomisation.

The 3-level EQ-5D questionnaire was administered at 3 and 12 months to patients discharged following ruptured AAA repair. The EQ-5D utility index score was calculated by combining the EQ-5D health profile of each patient with health state preference values from the UK general population. The resultant mean QoL utility scores at 3 and 12 months post-randomization were contrasted between the randomized groups, with unpaired t-tests. QALYs up to 1 year were calculated by valuing each patient's survival time by their QoL at 3 and 12 months according to the 'area under the curve' method.

4.1.5. COUGAR-02

Oesophagogastric cancer is the fifth most common type of cancer in the UK and is associated with poor prognosis and survival [18]. The COUGAR-02 randomised, controlled, open-labelled trial (ISRCTN13366390) compared docetaxel chemotherapy plus active symptom control (DXL + ASC) and active symptom control (ASC) only in patients in the UK with advanced adenocarcinoma of the oesophagus, oesophagogastric junction, or stomach[19]. Patients (aged 18 years and over) were included in the trial if their cancer had progressed within 6 months of treatment with a platinum-fluoropyrimidine combination. They were randomised on a 1:1 basis and those in the DXL + ASC arm received a dose of 75 mg/m2 of docetaxel by intravenous infusion every 3 weeks for up to six cycles.

Utility was based on the EQ-5D (three-level) and UK scoring tariff. 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.

4.1.6. 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)[20,21]. It compared fludarabine, cyclophosphamide and rituximab (FCR), which is considered conventional frontline therapy, with fludarabine, cyclophosphamide, mitoxantrone and low dose rituximab (FCM-miniR). The intention was to randomise 206 patients on a 1:1 basis to receive FCR or FCM-miniR. However, interim analysis by the Data Monitoring and Ethics Committee (DMEC) led to early trial closure. Although the response rates in both arms were higher than anticipated, FCM-miniR had a lower CR rate than FCR. 100 participants completed FCR, 79 FCM-miniR and 21 commenced FCM-miniR but switched to FCR following DMEC recommendations.

Participants completed questionnaires, including EQ-5D-3L, 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). Completion rates were highest at baseline, 9 months, 12 months, 18 months and 24 months.

An economic evaluation was conducted to assess the cost-effectiveness of FCM-MiniR compared to FCR from a UK NHS and personal social services (PSS) perspective. The economic evaluation used a within trial analysis, in which cost-effectiveness was assessed within the 24-month trial period using individual patient data collected in the trial; and a decision analytic model analysis, in which cost-effectiveness is assessed over a lifetime horizon using standard modelling techniques applied to the trial data in order to extrapolate the trial results. As the analysis spans more than one year, future costs and health outcomes (beyond one year) were discounted at an annual rate of 3.5% as per the NICE Methods Guide.

4.1.7. SHARPISH

Relapse rates among those who have recently stopped smoking (short-term quitters) are high. The SHARPISH (Self-Help And Relapse Prevention In Smoking for Health) trial sought to estimate the effectiveness and cost-effectiveness of self-help booklets to prevent smoking relapse in people who had stopped smoking for four weeks[22]. The control arm was a single leaflet, quitters were carbon monoxide (CO) verified at 4 weeks after their quit dates in stop smoking clinics. Those who were pregnant, unable to read booklets in English, or younger than 18 years were excluded. Participants were followed up at 3 and 12 months after the quit date (2 months and 11 months post-randomisation). The 3-level EQ-5D questionnaire was administered at baseline, 2 months and 11 months post-randomisation.

Of the 1416 randomised participants, 1049 had complete EQ-5D data at the 3 follow-up points and were included in the complete case analysis. For these, seemingly unrelated regression was used to estimate the incremental cost and QALY gain whilst controlling for baseline EQ-5D score and particular baseline demographic and smoking descriptive variables (see Table 26 the main report[22]).

4.1.8. WRAP

Weight-Reduction Activity Programme (WRAP)[23] was a multi-centre, non-blinded, three-arm parallel groups randomised controlled trial with imbalanced randomisation (5:5:2) of two weight loss programmes, compared to a brief intervention in overweight adults. In the two intervention arms, participants were given free access to commercial weight loss programme (WeightWatchers UK) for a period of either 52 weeks (CP52) or 12 weeks (CP12). In the brief intervention arm (BI) participants were given a printed British Heart Foundation booklet of self-help weight management strategies.

Cost-utility was based on incremental cost-per Quality Adjusted Life Year (QALY) gained using the EQ5D-3L UK tariff. The costing perspective was of UK NHS.

A cost-effectiveness analysis, with 24-month time horizon was the primary analysis based on an assessment of the incremental cost per KG of weight gained. A cost-utility analysis (also with 24-month time horizon) assessing cost per Quality Adjusted Life Year (QALY) was also conducted. Costs were derived from resource use and collected from multiple sources.

QALYs were determined from EQ-5D-3L questionnaires at baseline, 3, 12 and 24 months. Utility scores were measured at each time point and calculated using the Area Under the Curve approach assuming a linear relationship between time points. A bivariate regression analysis of cost and QALYs, with age, sex and baseline utility was conducted. Methods conformed to guidelines currently recommended by NICE.

As this was a three-armed trial, we compared the costs and outcomes for both commercial programmes (CP12 and CP52) to the costs and outcomes of Brief Intervention. 1267 eligible participants were randomised to the brief intervention (211), 12-week programme (528) and 52-week programme (528). In this report we focus on complete case analysis. 571 completed all four EQ5D and health resource use questionnaires to be included in complete case cost effectiveness analysis at two years.

4.1.9. CvLPRIT

Primary percutaneous coronary intervention (P-PCI) is the standard of care for patients presenting with ST-segment elevation Myocardial Infarction (STEMI), with >90,000 such procedures undertaken in the UK each year. Of patients presenting with STEMI, 40-65% are estimated to have bystander stenosis. Cardiologists have long-debated whether these bystander stenosis should be treated ("complete" revascularisation) alongside the heart-attack causing arteries. The CvLPRIT (Complete- compared to Lesion-Only Revascularisation For Myocardial Infarction) trial[24] randomised patients at the point of in-hospital clinical diagnosis 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). The primary outcome was MACE (major adverse cardiac events, a composite of all-cause mortality, repeat revascularisation, recurrent myocardial infarction, or heart failure) at 12 months after randomisation. In addition to MACE, the 3-level EQ-5D questionnaire was administered at two time points: immediately before discharge from index admission, and at 12 months post-discharge.

The CvLPRIT study recruited 296 participants, of which complete EQ5D data was available for 203. In the 12 month follow-up period 14 died and 19 were lost to follow-up. To estimate the incremental cost effectiveness ratio (ICER), we performed bivariate regression of costs and outcomes, adjusting for differences in baseline EQ5D.

4.2. RESULTS

Table 5 and Figure 5 report headline results for all the case studies. Results for individual case studies are expanded upon in subsequent sections.

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, very slightly when using 5L (EQG) compared to 3L. On the face of it, this is not in line with most other results, including those for the 12 –week programme in the same study. Further insight to this is given in Figure 6. For CP52 the incremental QALY gain is made up of the observed difference in utilities prior to 12 months and those between 12 months and 24 months. However, there was a lower utility associated with the intervention compared to control at the final 24 month measure. Using EQ5D-3L there is large reduction in overall QALYs between 12 and 24 months. Using EQ5D-5L, this difference is smaller both because the utility difference at 24 months is smaller and the impact this has on the linear interpolation back to the 12 month observed values. Therefore, the net effect is that QALY gains are marginally larger using 5L (EQG) than with 3L. The incremental QALY gain falls when using the 5L (NDB) in this study. The WRAP study results are qualitatively similar to those observed in other studies. That is, differences in utilities tend to be smaller using 5L than with 3L. The difference with WRAP is the cumulative impact on total QALYs since some time points suggest the intervention is worse than the comparator.

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. This is because COUGAR 2 is the only one of our nine case studies in which 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[19]. 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

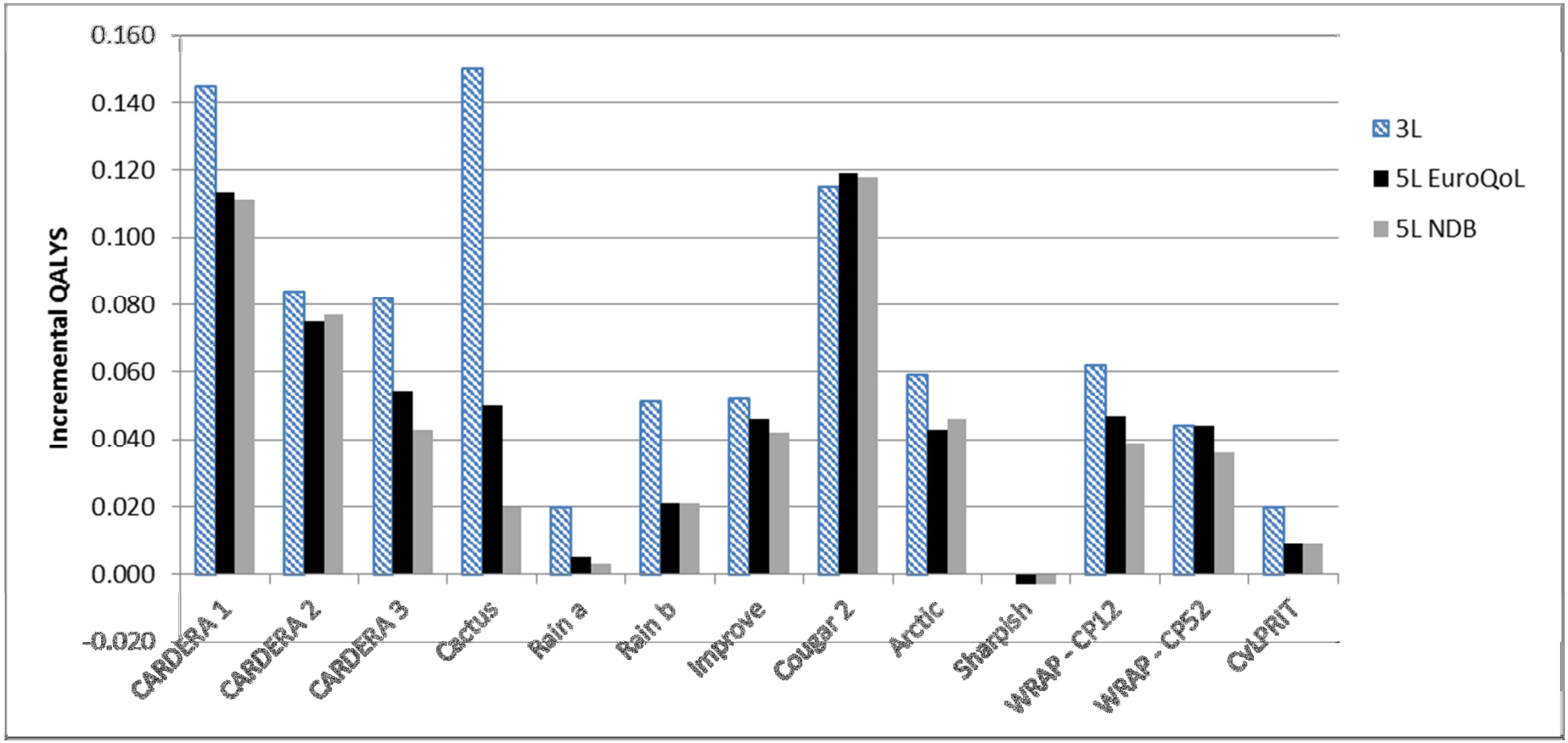
substantial mortality rate (approximately 25% mortality within 6 months) this was substantially lower than in COUGAR-02 (approximate 6-month mortality of 75% in the control group and 60% in the docetaxel arm[19]) and did not outweigh the morbidity effect.

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

	Inc QALYs					ICER				
	3L	**5L EuroQoL**	**% change**	**5L NDB**	**% change**	**3L**	**5L EuroQoL**	***% 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		-0.003						
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	0.009	-55.5%	0.009	-56.9	23208	51614	*122.4%*	53908	*132.3%*

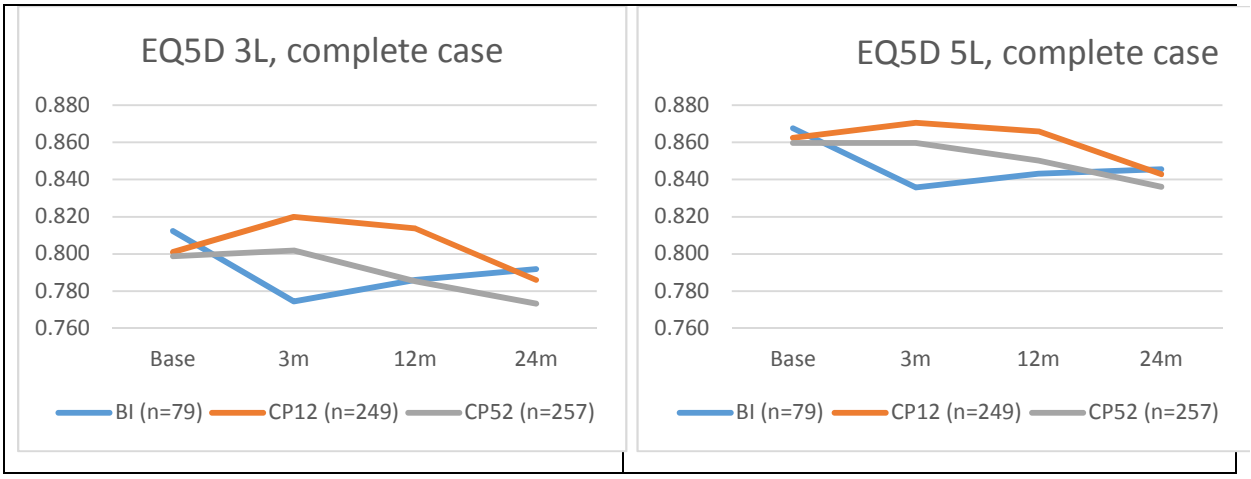
CARDERA 1 = MTX vs MTX + CS, CARDERA 2 = MTX vs MTX + PNS, CARDERA 3 = MTX + CS + PNS vs MTX

Figure 5: 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

Figure 6: Histogram of 3L and 5L EQG in WRAP study



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. Results from several studies illustrate this clearly (see, for example, Figure 9 and Figure 10 for the CACTUS study.

In eight of the thirteen reported comparisons, the incremental QALY gain is greater when measured using EQ5D-5L and the EQG dataset, compared to EQ5D-5L and the NDB dataset. Three of the five remaining comparisons showed no difference.

In those studies where the EQ5D-5L and EQG combination 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 EQ5D-5L (EQG data) and over £30k using EQ5D-5L (NDB data). Similarly, CACTUS changes from a highly cost effective central estimate using EQ5D-3L (£3058) to one that is more borderline (£23022) using EQ5D-5L (NDB data). Other case studies demonstrate changes in cost effectiveness that may

not span boundaries of typically cited cost-effectiveness thresholds but are, nevertheless, very substantial.

4.2.1. Results for CARDERA study

Since CARDERA was a 4-arm study there are 6 pairwise comparisons that can be made. Table 6 below includes all these comparisons, expanding on the three main MTX comparisons in Table 5. It consistently shows that the QALY gain is smaller when measured by either 5L estimate compared to 3L. In all but one comparison, the QALY gain is larger when using the NDB based estimates than the EQG ones. It is worth noting that the NDB is a registry of rheumatology patients and this also allowed us to use a copula model that included the disease specific covariates HAQ and pain[4]. The results with this model varied, with no consistent direction of change across the comparisons.

Table 6: Incremental costs and QALYs from all comparisons in CARDERA

	Inc QALYs						
	3L	**5L EQG**	***% change***	**5L NDB**	***% change***	***5L NDB cov***	***% change***
MTX vs MTX+CS	0.145	0.113	*21.8%*	0.111	*23.2%*	0.097	*33.0%*
MTX vs MTX+PNS	0.084	0.075	*10.4%*	0.077	*8.0%*	0.065	*22.5%*
MTX+CS+PNS vs MTX	0.082	0.054	*33.5%*	0.043	*47.6%*	0.064	*21.5%*
MTX +PNS vs MTX+CS	0.061	0.038	*37.3%*	0.034	*44.2%*	0.032	*47.6%*
MTX+CS+PNS vs MTX+CS	0.226	0.168	*26.0%*	0.154	*32.0%*	0.161	*28.9%*
MTX+PNS vs MTX+CS+PNS	0.165	0.129	*21.8%*	0.120	*27.5%*	0.129	*22.0%*

Notes: MTX – methotrexate, CS – Cyclosporin, PNS – Steroid, Cov - covariates

4.2.2. Results for CACTUS study

Using 3L EQ-5D scores, self-managed computer-assisted therapy was associated with an incremental cost-effectiveness ratio (ICER) of £3,058 per QALY gained (Table 7). With 5L scores derived from the EuroQol dataset, the ICER was significantly larger, at £9,481 per QALY gained. With 5L scores estimated from the RA dataset, the ICER was even larger. These differences were driven by the difference in utility improvement that was attainable with 3L and 5L scores from the two datasets. This utility improvement was much larger with the 3L values than with the 5L values (Table 8).

5L scores were higher than their 3L counterparts on average (see Table 8 and Figure 7). Furthermore, the distribution of 5L scores appears to be slightly narrower than that for 3L scores at different trial points and the different trial arms (Figure 9). An alternative representation of the difference between estimated 3L and 5L scores is given in Figure 10. It

can be seen that whilst the 5L values all lie above their 3L equivalents, the difference is more pronounced at the lower end of disease severity, as measured by EQ-5D. This feature is clear using both mapping models.

Table 7: Comparison of cost-effectiveness results for CACTUS pilot study

Per person treated		*Cost*	*QALYs*	*Incremental Cost*	*Incremental QALY*	*ICER*
3-level	Control	£18,687	3.07			
	Treatment	£19,124	3.22	£ 436.87	0.15	£ 3,058.21
5-level (EuroQol dataset)	Control	£18,687	3.61			
	Treatment	£19,124	3.66	£ 436.87	0.05	£ 9,480.92
5-level (DNB RA dataset)	Control	£18,687	3.71			
	Treatment	£19,124	3.73	£ 436.87	0.02	£23,022.47

Table 8: Comparison of health state utilities for CACTUS pilot study

Health states utilities	*Utility in control group / with no response*	*Utility improvement with good response*	*Utility with good response*
3-level	0.55	0.07	0.62
5-level (EuroQol dataset)	0.65	0.02	0.67
5-level (DNB RA dataset)	0.66	0.01	0.67

Figure 8: Changes in utility over time in CACTUS study

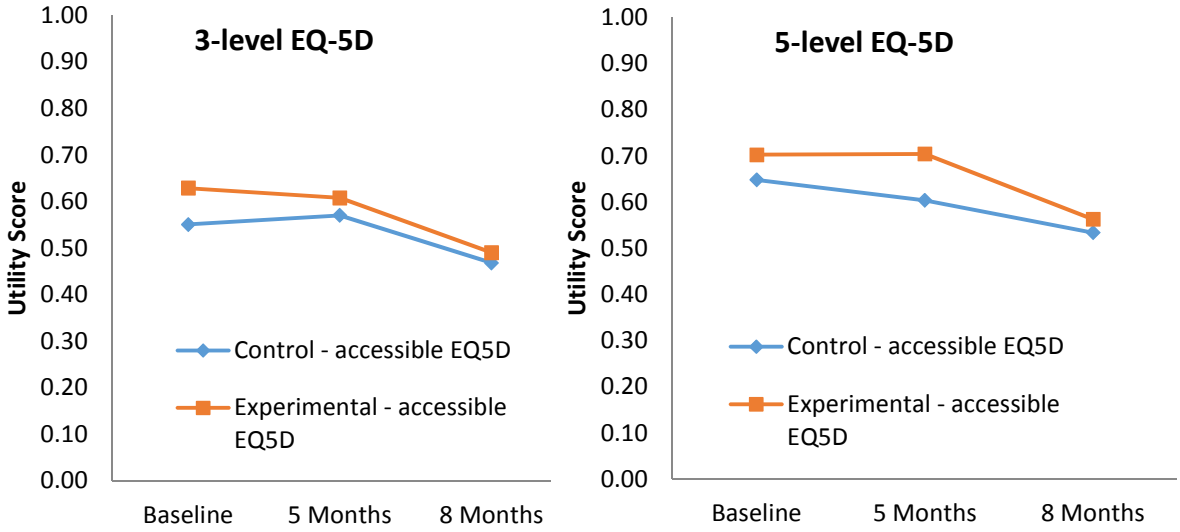


Figure 9: Comparison of utility score distributions (CACTUS pilot study)

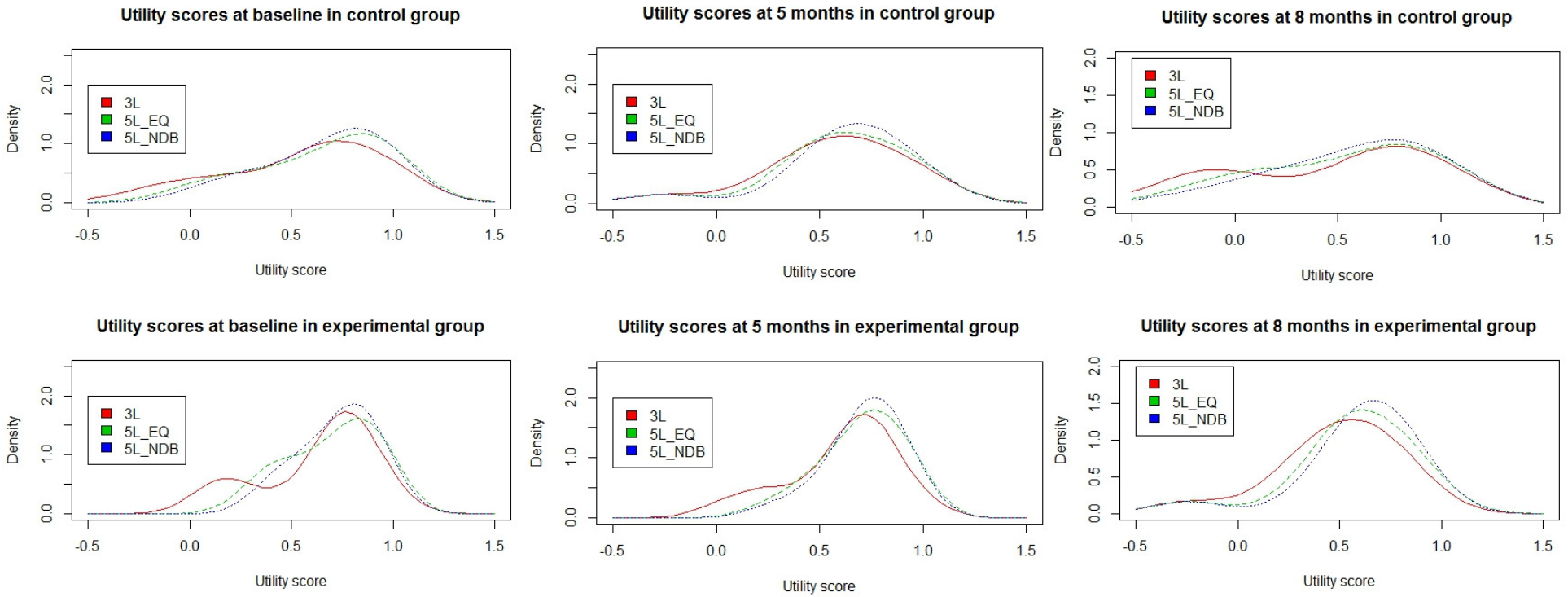
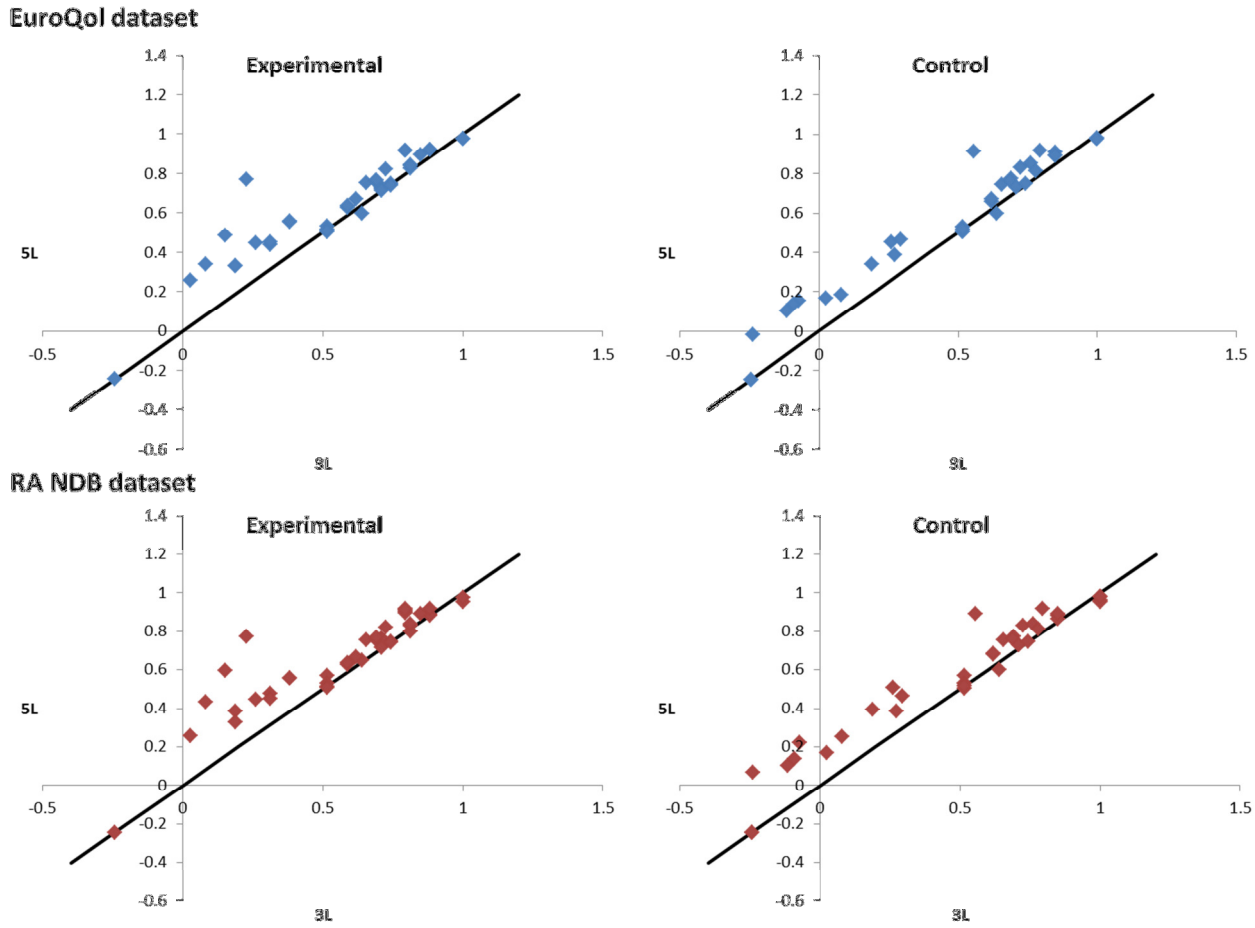


Figure 10: CACTUS 5L vs 3L values



4.2.3. Results of the RAIN study

Two different comparisons were made in the RAIN study: dedicated neurocritical care unit compared with combined neuro/general critical care unit and 'early' compared with 'no or late' transfer to a neuroscience centre. Full results for both comparisons are reported in Table 9.

Table 9: Comparison of cost-effectiveness results for RAIN study at 6 months

	Comparison 1: dedicated neurocritical care unit versus combined neuro and general critical care unit				
Per person treated		**Cost**	**QALYs**	**Incremental Cost***	**Incremental QALY***
3-level	Combined neuro and general critical care unit	£25,466	0.16		
	Dedicated neurocritical care unit	£28,855	0.18	£3,694	0.020
5-level (EuroQol dataset)	Combined neuro and general critical care unit	£25,466	0.15		
	Dedicated neurocritical care unit	£28,855	0.16	£3,694	0.005
5-level (DNB RA dataset)	Combined neuro and general critical care unit	£25,466	0.15		
	Dedicated neurocritical care unit	£28,855	0.16	£3,694	0.003
	Comparison 2: 'Early' versus 'no or late' transfer to a neuroscience centre				
3-level	'No or late' transfer to neuroscience centre	£13,153	0.13		
	'Early' transfer to neuroscience centre	£28,525	0.22	£15,001	0.051
5-level (EuroQol dataset)	'No or late' transfer to neuroscience centre	£13,153	0.13		
	'Early' transfer to neuroscience centre	£28,525	0.19	£15,001	0.021
5-level (DNB RA dataset)	'No or late' transfer to neuroscience centre	£13,153	0.13		
	'Early' transfer to neuroscience centre	£28,525	0.19	£15,001	0.021

* Incremental effects are after case-mix adjustment

Figure 11: Histogram to show distribution of 6 month 3L and 5L utility scores for patients assigned to a) Combined neuro and general critical care unit and b) Dedicated neurocritical care unit

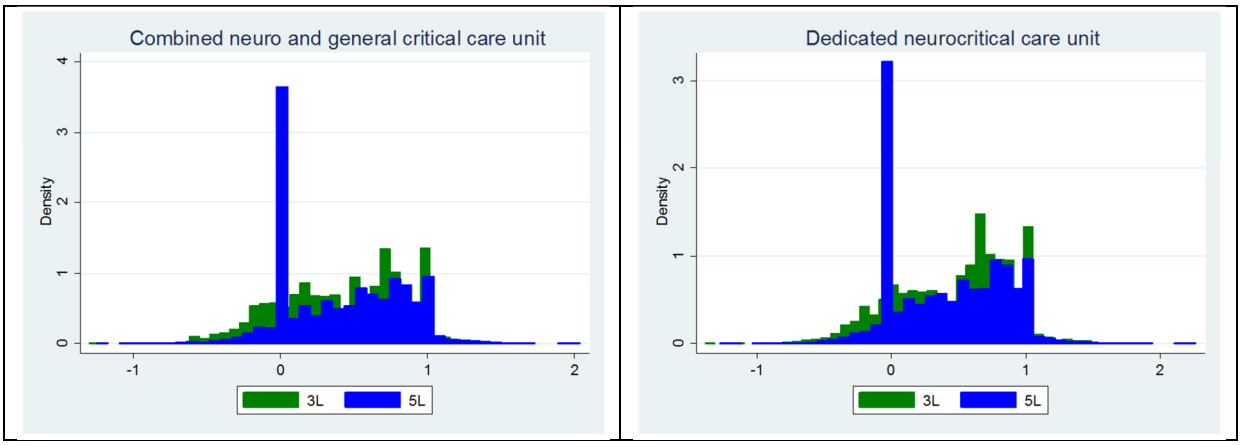
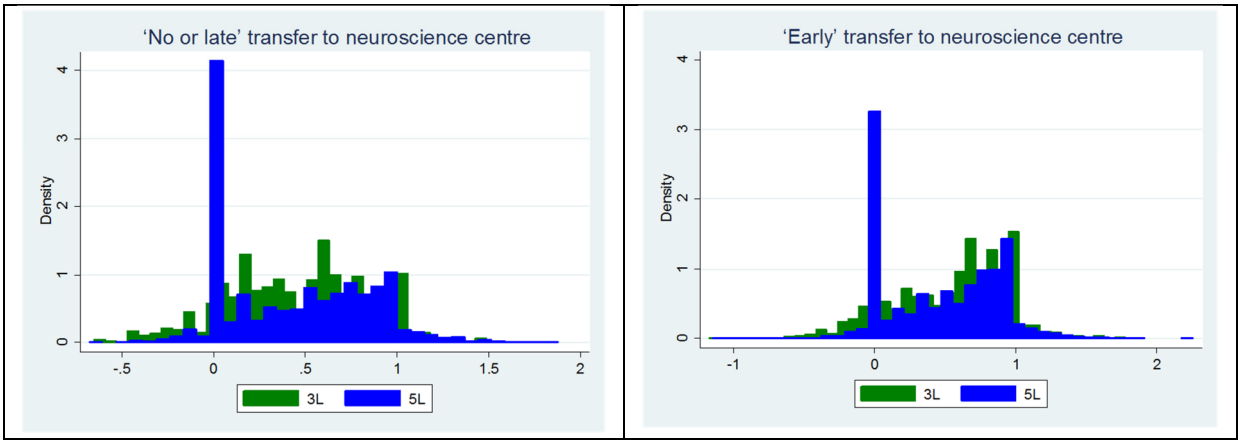


Figure 12: Histogram to show distribution of 6 month 3L and 5L utility scores for patients assigned to a) No or late transfer to neuroscience centre and b) Early transfer to neuroscience centre



4.2.4. Results for IMPROVE study

Full results from the IMPROVE study are presented in Table 10. Histograms of the distribution of 3L compared to 5L are presented in Figure 13, Figure 14, Figure 15 and Figure 16.

Figure 13: Distribution of 3L and 5L (NDB) scores in IMPROVE study, complete cases

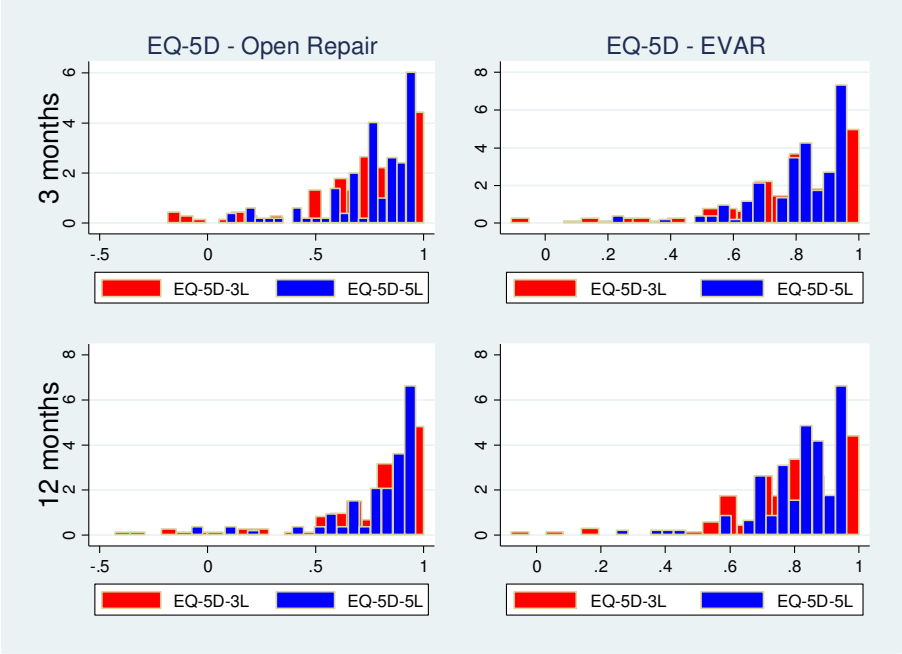


Figure 14: Distribution of 3L and 5L (NDB) scores in IMPROVE study, after imputation

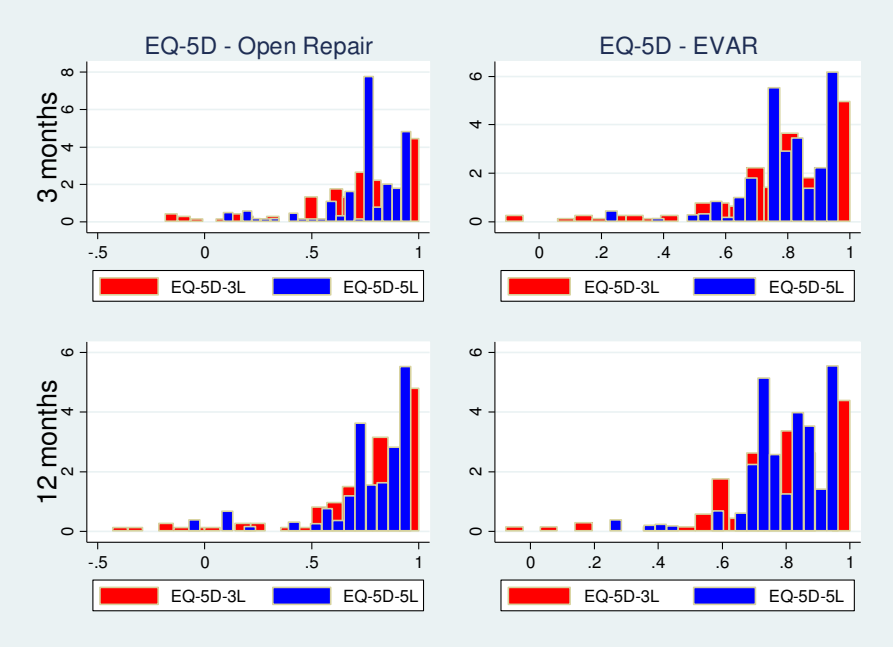


Table 10: Results from IMPROVE trial

	EQ-5D-3L			EQ-5D-5L$_{EuroQol}$			EQ-5D-5L$_{NBD}$		
	Open	EVAR	Difference	Open	EVAR	Difference	Open	EVAR	Difference
EQ-5D$_{3\text{-months}}$*	0.67	0.76	0.087	0.73	0.81	0.074	0.75	0.81	0.061
EQ-5D$_{12\text{-months}}$*	0.71	0.77	0.068	0.77	0.82	0.045	0.77	0.81	0.041
QALY**	0.35	0.40	0.052	0.38	0.42	0.046	0.38	0.42	0.042
Cost (£)	18,723	16,394	-2,329	18,723	16,394	-2,329	18,723	16,394	-2,329
ICER (£/QALY)	-44 617			-48 113			-54 742		
INB [95% CI]***	3877 [253, 7 408]			3617 [92, 7142]			3520 [-51, 7 078]		

* For survivors only; ** For all randomised patients, *** £30 000 per QALY gain

Figure 15: Distribution of 3L and 5L (EQG) scores in IMPROVE study, complete cases

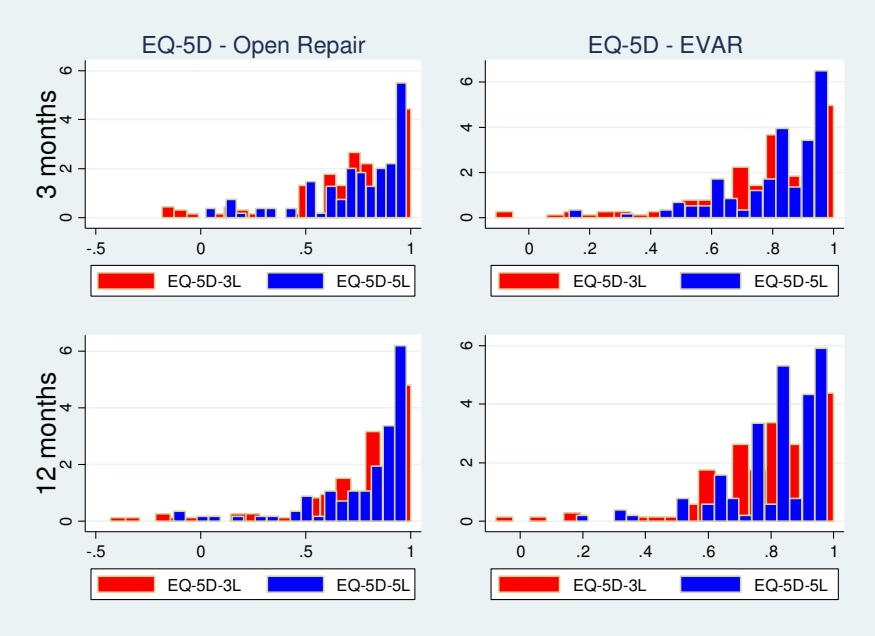
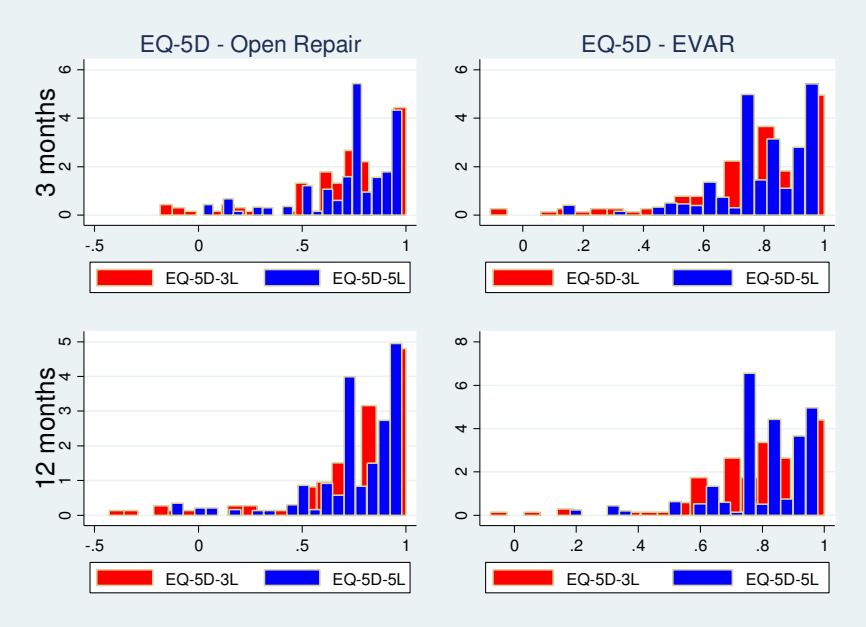


Figure 16: Distribution of 3L and 5L (EQG) scores in IMPROVE study, after imputation

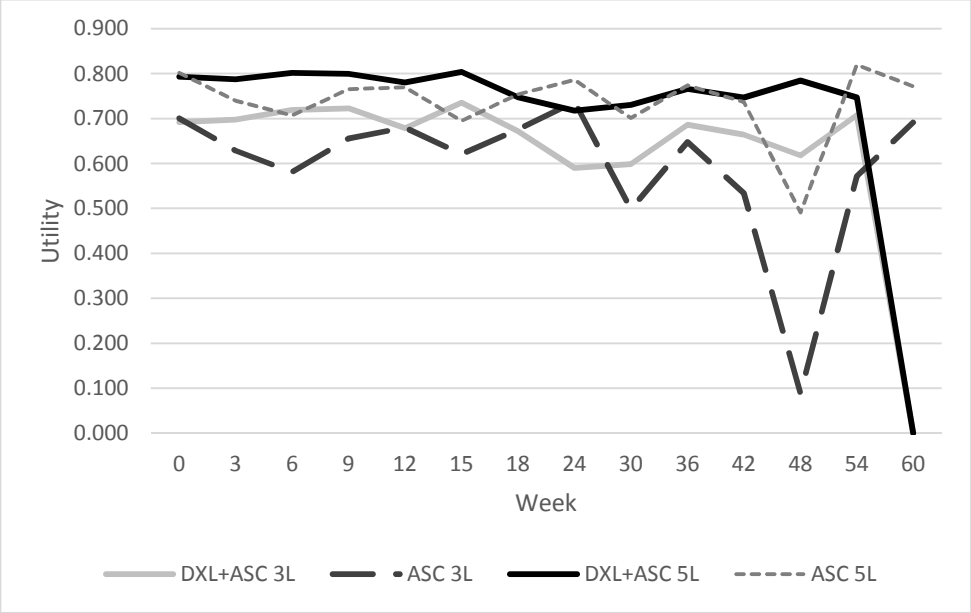


4.2.5. Results of the COUGAR-02 study

Table 11: Results from COUGAR-02 study

Per person treated		*Cost*	*QALYs*	*Incremental Cost*	*Incremental QALY*	*ICER*
3-level	DXL+ASC	£9,352	0.302			
	ASC	£6,218	0.186	£3,135	0.115	£27,180
5-level (EuroQol dataset)	DXL+ASC	£9,352	0.341			
	ASC	£6,218	0.223	£3,135	0.119	£26,434
5-level (DNB RA dataset)	DXL+ASC	£9,352	0.344			
	ASC	£6,218	0.225	£3,135	0.118	£26,484

Figure 17: Mean utility by week in COUGAR-02



4.2.6. Results of the ARCTIC study

Table 12: EQ-5D index scores at the baseline and follow-ups, and QALYs of CLL participants by treatment arm (imputed data).

	Parameter	FCR: EQ-5D-3L	FCR: EQ-5D-5L (NDB)	FCR: EQ-5D-5L (EQG)	FCM-miniR: EQ-5D-3L	FCM-miniR: EQ-5D-5L (NDB)	FCM-miniR: EQ-5D-5L (EQG)
	n	n=92	81	n=88	n=70	61	n=69
Baseline	Mean	0.829	0.86	0.869	0.774	0.843	0.849
	Std Dev	0.200	0.134	0.153	0.275	0.148	0.175
	(min-max)	(-0.016 - 1.000)	(-0.227 -0.961)	(0.167 - 0.980)	(-0.016 - 1.000)	(0.349 -0.961)	(0.286 - 0.980)
3 months after End of Therapy	Mean	0.852	0.871	0.871	0.868	0.876	0.876
	Std Dev	0.141	0.123	0.123	0.194	0.156	0.156
	(min-max)	(0.378 - 1.000)	(0.512 - 0.980)	(0.512 - 0.980)	(-0.239 - 1.000)	(-0.019 -0.980)	(-0.0186 - 0.980)
12 months post randomisation	Mean	0.838	0.86	0.862	0.863	0.884	0.887
	Std Dev	0.177	0.151	0.149	0.218	0.169	0.162
	(min-max)	(0.189 - 1.000)	(0.337 -0.980)	(0.337 - 0.980)	(-0.074 - 1.000)	(0.155 -0.980)	(0.155 - 0.980)
18 months post randomisation	Mean	0.833	0.835	0.833	0.851	0.866	0.869
	Std Dev	0.18	0.149	0.147	0.184	0.139	0.136
	(min-max)	(0.145 - 1.000)	(0.274 -0.945)	(0.274 - 0.980)	(-0.003 - 0.965)	(0.254 -0.945)	(0.254 - 0.945)
24 months post randomisation	Mean	0.852	0.865	0.863	0.871	0.872	0.873
	Std Dev	0.161	0.136	0.137	0.097	0.097	0.099
	(min-max)	(-0.071 - 0.965)	(0.144 -0.945)	(0.144 -0.945)	(0.498 - 0.965)	(0.503 -0.945)	(0.503 -0.945)
Total QALYs	Mean	1.61	1.506	1.509	1.552	1.46	1.466
	Std Dev	0.329	0.43	0.43	0.414	0.52	0.525
	(min-max)	(0.418 -2.000)	(0.481 - 1.925)	(0.481 - 1.934)	(0.049 - 1.974)	(0.324 - 1.925)	(0.324 - 1.933)

4.2.7. Results of the SHARPISH Study

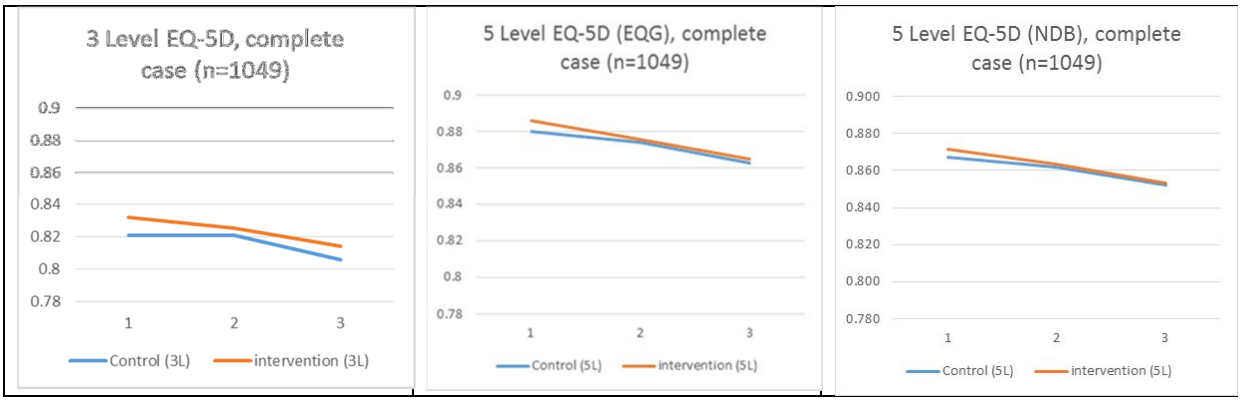
Table 13: Mean utilities and overall QALYs within the SHARPISH study: 3L and 5L (EQG and NDB)

	Control - 3L (n=528)	Intervention - 3L (n=521)	Control - 5L (n=528)	Intervention - 5L (n=521)	Control - 5L (NDB) (n=528)	Intervention - 5L (NDB) (n=521)
Baseline	0.821	0.832	0.880	0.886	0.867	0.872
2 months	0.821	0.825	0.874	0.876	0.862	0.863
11 months	0.806	0.814	0.863	0.865	0.852	0.853
QALY	0.747	0.753	0.798	0.800	0.787	0.788

Table 14: Cost effectiveness results 3L and 5L (EQG and NDB)

Per person treated	*Cost*		*QALYs*	*Incremental Cost*	*Incremental QALY ADJUSTED*	*Incremental net benefit*
3-level	Control	£657.95	0.747			
	Treatment	£553.78	0.753	-£84.49	0.000	£74.80
5-level (EQG)	Control	£657.95	0.798			
	Treatment	£553.78	0.800	-£85.53	-0.003	£34.96
5-level (NDB)	Control	£657.95	0.787			
	Treatment	£553.78	0.788	-£85.92	-0.002	£43.45

Figure 18: Plot of EQ-5D over time



4.2.1. Results of the WRAP Study

Figure 19: Plot of mean 3L and 5L (EQG and NDB) for three arms of the WRAP study

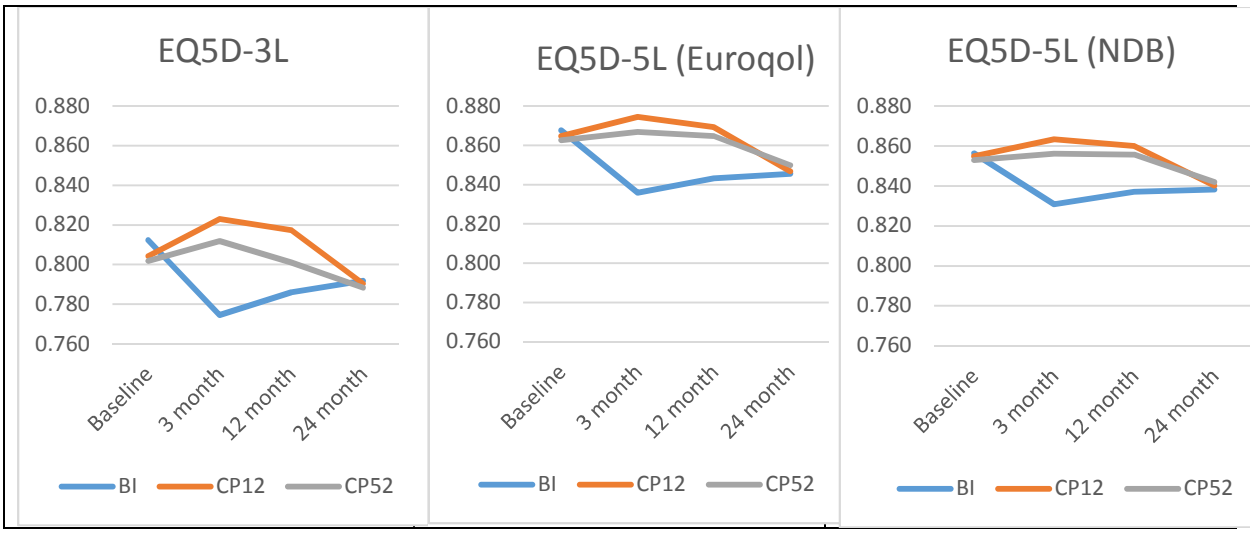


Table 15: EQ-5D index scores at baseline and follow-ups, and total QALYs.

	Arm	BI			CP12			CP52		
	N	79			241			251		
	Tariff	**3 Level**	**5 Level EQG**	**5 Level NDB**	**3 Level**	**5 Level EQG**	**5 Level NDB**	**3 Level**	**5 Level EQG**	**5 Level NDB**
Baseline	**Mean**	**0.812**	**0.868**	**0.856**	**0.804**	**0.865**	**0.855**	**0.802**	**0.863**	**0.853**
	Std. Dev.	0.245	0.161	0.14	0.24	0.163	0.139	0.234	0.162	0.138
	Min	-0.016	0.315	0.378	-0.074	0.146	0.213	-0.181	0.117	0.204
3 month	**Mean**	**0.775**	**0.836**	**0.831**	**0.823**	**0.874**	**0.863**	**0.812**	**0.867**	**0.856**
	Std. Dev.	0.267	0.186	0.157	0.233	0.158	0.134	0.254	0.174	0.15
	Min	-0.016	0.283	0.347	-0.074	0.15	0.213	-0.239	-0.021	0.033
12 month	**Mean**	**0.786**	**0.843**	**0.837**	**0.817**	**0.869**	**0.86**	**0.801**	**0.865**	**0.856**
	Std. Dev.	0.253	0.181	0.154	0.228	0.159	0.135	0.269	0.178	0.15
	Min	-0.074	0.152	0.223	-0.181	0.117	0.213	-0.181	-0.001	0.045
24month	**Mean**	**0.792**	**0.846**	**0.838**	**0.79**	**0.847**	**0.84**	**0.788**	**0.85**	**0.842**
	Std. Dev.	0.266	0.182	0.155	0.26	0.188	0.162	0.244	0.174	0.15
	Min	-0.074	0.152	0.223	-0.239	-0.024	0.036	-0.239	-0.02	0.05
Total QALYs	**Mean**	**1.572**	**1.687**	**1.674**	**1.622**	**1.729**	**1.711**	**1.601**	**1.723**	**1.704**
	Std. Dev.	0.472	0.339	0.288	0.421	0.306	0.261	0.47	0.327	0.278
	Min	-0.112	0.406	0.558	-0.205	0.286	0.423	-0.389	0.144	0.328

4.2.1. Results of the CVLPRIT Study

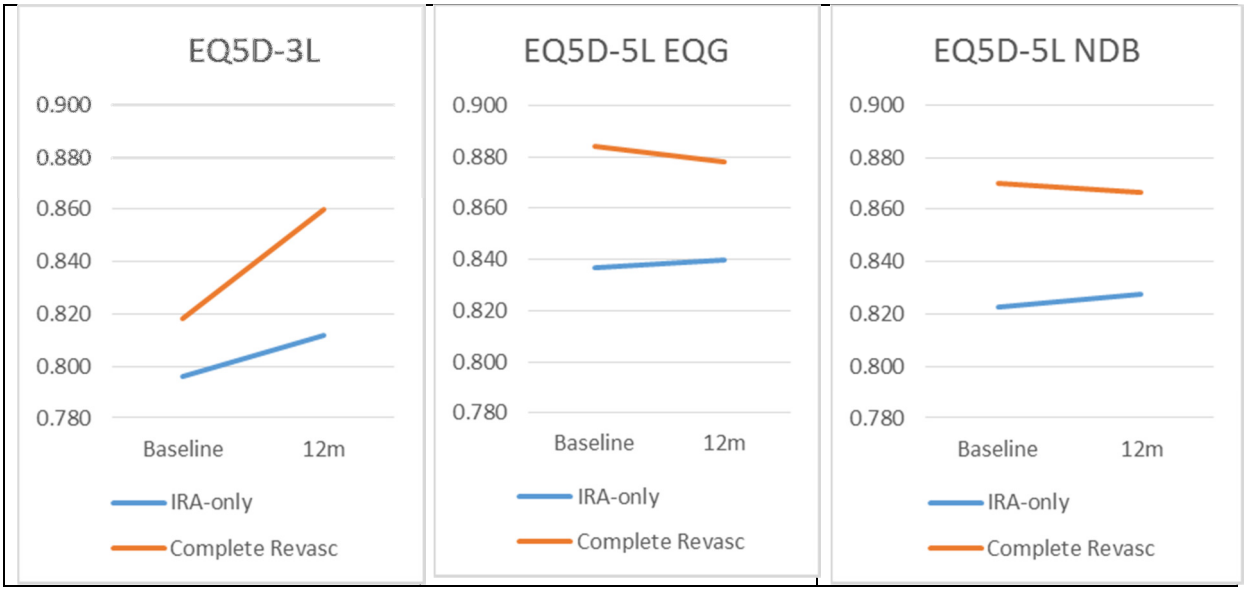
Table 16: Summary results from CVLPRIT Study

Per person treated		*Cost*	*QALYs*	*Incremental Cost*	*Incremental QALY ADJUSTED*	*ICER ADJUSTED*
3-level	Control	£4,918.60	0.801			
	Treatment	£5,551.70	0.833	£446.65	.020	£23,208
5-level EQG	Control	£4,918.60	0.838			
	Treatment	£5,551.70	0.881	£446.65	.0089	£51,614
5-level NDB	Control	£4,918.60	0.825			
	Treatment	£5,551.70	0.869	£446.65	.0086	£53,908

Table 17: Utility scores baseline and 12 months, and QALYs in the CVLPRIT study

	ARM	**IRA-Only**			**Complete Revasc**		
	N	100			103		
	Tariff	3L	5L EQG	NDB	3L	5L EQG	NDB
Baseline	**Mean**	**0.796**	**0.837**	**0.823**	**0.818**	**0.884**	**0.87**
	Std. Dev.	0.269	0.238	0.226	0.221	0.143	0.123
12m	**Mean**	**0.812**	**0.84**	**0.828**	**0.86**	**0.878**	**0.867**
	Std. Dev.	0.291	0.261	0.252	0.228	0.196	0.182
QALY	**Mean**	**0.801**	**0.838**	**0.825**	**0.833**	**0.881**	**0.869**
	Std. Dev.	0.258	0.236	0.228	0.204	0.142	0.126

Table 18: Baseline and 12 month 3L and 5L (EQG and NDB)



5. DISCUSSION

NICE places great emphasis on cost-effectiveness to inform decision-making across many of its guidance-issuing programmes. It is the nature of the Institute's work that decisions need to be consistent across disease areas, patient groups and technologies. This in turn requires a degree of consistency in the methods for undertaking cost-effectiveness analysis in order that results can be compared to each other and be interpreted against some threshold value or range.

It is clear from the work presented here that EQ-5D-3L and 5L versions do produce substantially different estimates of cost effectiveness. In all cases, 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, the incremental cost effectiveness ratio of a clinically effective technology rises (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. The work illustrates that small absolute changes in health utility estimates often have a substantial impact on ICERs. The nature of the utility scale means changes may appear

small in absolute terms but, since QALY gains are rarely large in economic evaluation, these are important differences.

In this sense, 3L and 5L are not consistent with each other. This raises a number of questions and policy challenges for NICE. There are choices to be made about how prescriptive NICE should be about the use of different utility instruments. There are implications for mapping between instruments using different descriptive systems, but with the potential to apply different tariffs via mapping. In addition, there are implications for the threshold – the value of services that are displaced in the NHS as a result of introducing new technologies to the budget constrained health system. 5L is already being used as the descriptive system in many ongoing clinical trials. Therefore, this will increasingly form part of the evidence base of the effectiveness of new technologies. But it is also important to recognise that the relevance of 3L will remain for many years, perhaps decades. This is because comparator therapies will have evidence on their effectiveness conducted either all or in part using 3L and utility values for health states for models drawn from existing published literature will be in 3L terms.

The requirement for consistency in decision-making means that the option of allowing appraisals to be conducted using either the 3L or the 5L instrument, with no adjustment to either, cannot be an appropriate route. It is clear that cost-effectiveness estimates vary by instrument.

It is important to consider the extent to which any change to the distribution of current NICE recommendations is a relevant issue. On purely scientific grounds, with no regard for past NICE decisions, our view is that the 5L is likely superior to the 3L in terms of the richer descriptive system and the valuation methods: inevitably so given the learning that the developers have gained from over 20 years of application. But our results imply that moving wholesale to 5L would likely lead to a much greater level of "not recommended" technologies, *ceteris paribus*, with little change for technologies that are life extending.

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. This is apparent in the results of the statistical modelling and is borne out in the cost-effectiveness case studies, which span a range of disease areas, severity and

health technologies. Simple adjustment of the cost-effectiveness threshold or range will not produce a consistent movement from 3L to 5L. This is apparent in the range of changes in the ICERs for the case studies, which spanned a broad range of disease areas and health technologies.

It is feasible to adjust 3L evidence to its 5L equivalent, as has been done in this report. Whilst the model allows the 5L to 3L translation to be undertaken, the performance is worse. The validity of the model is 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. We therefore recommend more data collection and further investigation of the reasons for differences between mapping functions. Datasets with the degree of separation between 5L and 3L questions, as with the NDB, in patient samples covering different disease types would be valuable. If feasible, the incorporation of both 3L and 5L into clinical studies would allow direct comparisons of clinical and cost-effectiveness results.

We have found that there are also significant differences in utility estimates according to whether we estimate the expected 5L score using data from the EQG or from the NDB. Those differences are even more pronounced when we incorporate disease specific covariates into the mapping model. Each dataset has slightly different characteristics which may influence generalisability and validity for mapping. Importantly, the NDB includes only patients with rheumatoid disease whereas the EQG study comprises patients from a range of different conditions and also includes a healthy student population. The NDB though has a more substantial separation between the questions on 5L and the 3L than the EQG surveys, making it more likely that the responses given are truly independent. The differences observed raise the possibility that future mapping between the instruments may be best performed by disease specific estimates rather than a single generic one. This is important because one option for decision makers like NICE, is to allow evidence to be submitted using either EQ-5D variant and to exploit mapping to achieve consistency between appraisals (both current and historical).

Appendix Table 1: Parameter estimates for EQ5D-3L / 5L mapping models

	EQG dataset		NDB dataset	
	Coefficient	Std. Error	Coefficient	Std. Error

Mobility – 3L				
Male	-0.166	0.077	-0.127	0.080
Age/10	1.178	0.128	0.072	0.025
$(Age/10)^2$	-0.053	0.010	-	-
Common factor	1.664	0.106	1.558	0.058
Threshold 1	4.657	0.374	0.643	0.163
Threshold 2	8.216	0.544	5.122	0.244
Mobility – 5L				
Male	-0.198	0.065	-0.190	0.066
Age/10	1.091	0.113	0.132	0.021
$(Age/10)^2$	-0.050	0.009	-	-
Common factor	1.462	0.086	1.391	0.043
Threshold 1	4.119	0.324	0.391	0.136
Threshold 2	5.107	0.364	1.876	0.141
Threshold 3	6.140	0.406	3.393	0.157
Threshold 4	7.527	0.469	4.981	0.198
Dependency	16.881	1.710	0.598	0.020
Self-care – 3L				
Male	-0.267	0.081	-0.112	0.076
Age/10	0.790	0.123	0.000	0.024
$(Age/10)^2$	-0.028	0.011	-	-
Common factor	1.706	0.108	1.332	0.055
Threshold 1	4.307	0.372	1.242	0.158
Threshold 2	6.559	0.466	4.244	0.224
Self-care – 5L				
Male	-0.328	0.084	-0.050	0.069
Age/10	0.837	0.128	-0.004	0.022
$(Age/10)^2$	-0.026	0.011	-	-
Common factor	1.860	0.116	1.346	0.046
Threshold 1	4.580	0.399	0.756	0.143
Threshold 2	5.523	0.438	2.050	0.150
Threshold 3	6.507	0.481	3.552	0.174
Threshold 4	7.356	0.522	4.440	0.214
Dependency	15.558	1.484	2.475	0.183
Usual activities – 3L				
Male	-0.280	0.067	-0.546	0.106
Age/10	0.371	0.089	0.017	0.033
$(Age/10)^2$	-0.005	0.008	-	-
Common factor	1.432	0.084	2.231	0.104

Threshold 1	1.436	0.215	0.116	0.212
Threshold 2	3.778	0.278	4.598	0.281
Usual activities – 5L				
Male	-0.291	0.068	-0.528	0.086
Age/10	0.309	0.088	-0.003	0.027
$(Age/10)^2$	0.002	0.008	-	-
Common factor	1.582	0.093	1.933	0.069
Threshold 1	1.109	0.211	-0.644	0.174
Threshold 2	2.269	0.234	1.269	0.176
Threshold 3	3.283	0.263	3.270	0.196
Threshold 4	4.317	0.299	4.552	0.222
Dependency	10.653	0.778	0.096	0.044
Pain/discomfort – 3L				
Male	-0.264	0.043	-0.128	0.062
Age/10	0.473	0.059	0.005	0.020
$(Age/10)^2$	-0.029	0.006	-	-
Common factor	0.727	0.042	1.091	0.040
Threshold 1	1.134	0.132	-1.255	0.133
Threshold 2	3.011	0.182	2.067	0.136
Pain/discomfort – 5L				
Male	-0.253	0.042	-0.250	0.061
Age/10	0.473	0.058	-0.075	0.019
$(Age/10)^2$	-0.029	0.005	-	-
Common factor	0.807	0.043	1.263	0.038
Threshold 1	0.968	0.128	-2.600	0.139
Threshold 2	1.873	0.147	-0.325	0.126
Threshold 3	2.730	0.171	1.511	0.129
Threshold 4	3.635	0.204	2.692	0.142
Dependency	11.834	0.572	0.665	0.017
Anxiety/depression – 3L				
Male	-0.171	0.035	-0.149	0.051
Age/10	0.081	0.045	-0.130	0.016
$(Age/10)^2$	-0.010	0.004	-	-
Common factor	0.481	0.028	0.606	0.025
Threshold 1	0.013	0.101	-0.309	0.099
Threshold 2	1.316	0.112	1.519	0.107
Anxiety/depression – 5L				

Male	-0.146	0.032	-0.208	0.048
Age/10	0.023	0.042	-0.140	0.015
$(Age/10)^2$	-0.004	0.004	-	-
Common factor	0.499	0.027	0.630	0.024
Threshold 1	-0.248	0.097	-0.610	0.095
Threshold 2	0.387	0.095	0.344	0.094
Threshold 3	0.970	0.100	1.339	0.100
Threshold 4	1.631	0.116	1.910	0.116
Dependency	14.057	0.582	13.988	0.588
Mixture parameters				
Probability – class1	0.865	0.025	0.967	0.016
Probability – class2	0.135	0.025	0.033	0.016
Mean – class 1	0.151	0.017	0.028	0.011
Mean – class2	-0.967	0.169	-0.820	0.297
Variance – class 1	0.373	0.028	0.820	0.034
Variance – class 2	3.947	0.569	5.665	1.875

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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

[5] Ben van Hout, M.F. Janssen, You-Shan Feng, Thomas Kohlmann, Jan Busschbach, Dominik Golicki, Andrew Lloyd, Luciana Scalone, Paul Kind, A. Simon Pickard, Interim Scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L Value Sets, Value in Health, Volume 15, Issue 5, July–August 2012, Pages 708-715, ISSN 1098-3015, http://dx.doi.org/10.1016/j.jval.2012.02.008. (http://www.sciencedirect.com/science/article/pii/S1098301512000587)

[6] 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

[7] Wolfe, F. & Michaud, K. (2011), `The National Data Bank for rheumatic diseases: a multi-registry rheumatic disease data bank', Rheumatology 50, 16-24.

[8] Hernandez Alava, M. and Pudney S E. eq5dmap: a command for mapping from 3-level to 5-level EQ-5D. University of Sheffield: HEDS Discussion Paper, forthcoming.

[9] 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.

[10] Wailoo, A., Hernandez Alava, M., Scott, IC., Ibrahim, F., and Scott, DL. (2014) "Cost-effectiveness of treatment strategies using combination DMARDs and glucocorticoids in early rheumatoid arthritis", Rheumatology. doi: 10.1093/rheumatology/keu039

[11] 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, 29:4 (2013), 402–409.
[12] Harrison DA, Prabhu G, Grieve R, Harvey SE, Sadique MZ, Gomes M, et al. Risk Adjustment In Neurocritical care (RAIN) – prospective validation of risk prediction models for adult patients with acute traumatic brain injury to use to evaluate the optimum location and comparative costs of neurocritical care: a cohort study. Health Technol Assess 2013;17(23).
[13] Karthikesalingam A, Holt PJ, Vidal-Diez A, Ozdemir BA, Poloniecki JD, Hinchliffe RJ, Thompson MM. Mortality from ruptured abdominal aortic aneurysms: clinical lessons from a comparison of outcomes in England and the USA. Lancet 2014;383: 963–969.
[14] Mani K, Lees T, Beiles B, Jensen LP, Venermo M, Simo G, Palombo D, Troeng T, Wigger T, Bjorck M. Treatment of abdominal aortic aneurysm in nine countries 2005–2009: a Vascunet report. Eur J Vasc Endovasc Surg 2011;42:598–607.
[15] Edwards ST, Schermerhorn ML, O'Malley AJ, Bensley RP, Hurks R, Cotterill P, Landon DE. Comparative effectiveness of endovascular versus open repair of ruptured abdominal aortic aneurysm in the Medicare population. J Vasc Surg 2014;59: 575–582.
[16] Gupta PK, Ramanan B, Engelbert TL, Tefera G, Hoch JR, Kent KC. A comparison of open surgery versus endovascular repair of unstable ruptured abdominal aortic aneurysms. J Vasc Surg 2014;60:1439–1445.
[17] IMPROVE Trial Investigators Endovascular strategy or open repair for ruptured abdominal aortic aneurysm: one-year outcomes from the IMPROVE randomized trial. European Heart Journal (2015) 36, 2061–2069
[18] The Royal College of Surgeons of England. National Oesophago-Gastric Cancer Audit. 2012. http://www.hqip.org.uk/assets/NCAPOP-Library/NCAPOP-2012-13/Oesophago-Gastric-Cancer-National-Audit-INTERACTIVE-pub-2012.pdf.
[19] 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
[20] Peter Hillmen, Donald Milligan, Anna Schuh, Lucy McParland, Anna Chalmers, Tahla Munir, Abraham M Varghese, Andy C Rawstron, David J Allsup, Scott Marshall, Alex Smith, Corinne Collett, Walter Gregory, Andrew Duncombe and Dena Cohen. "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
[21] Howard, DR, Munir, T, McParland, L et al. (9 more authors) (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)
[22] 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.
[23] 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
[24] 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.