An international comparison of EQ-5D-5L and 3L for use in cost effectiveness analysis.

ABSTRACT 246 words (250 words max)

Objectives

To estimate the impact of using EQ5D-5L (5L) compared to EQ5D-3L (3L) in cost effectiveness analyses in 6 countries with 3L and 5L values: Germany, Japan, Korea, Netherlands, China and Spain.

Methods

8 cost effectiveness analyses based on clinical studies with 3L provided 11 pairwise comparisons. We estimated cost-effectiveness by applying the appropriate country values for 3L to observed responses. We re-estimated cost effectiveness for each country by predicting the 5L tariff score for each respondent, for each country, using a previously published mapping method. We compared results in terms of impact on estimated incremental Quality Adjusted Life Year (QALY) gain and cost-effectiveness ratios.

Results

For most countries the impact of moving from 3L to 5L is to lower the incremental QALY gain in the majority of comparisons. The only exception to this was Japan, where 4/11 (37%) of cases saw lower QALYs gained when using 5L. The mean and median reductions in health gain, in those case studies where 5L does lead to lower health gain, are largest in the Netherlands (84% mean reduction, 41% median reduction), Germany (68% and 27%) and Spain (30% and 31%). For most countries, those studies where 5L leads to lower health gain see larger reductions than the gains in studies showing the opposite tendency.

Conclusion

3L and 5L are not interchangeable in these countries. Differences between results are large but the direction of change can be unpredictable. These findings should prompt further investigation into the reasons for differences.

HIGHLIGHTS

i. What is already known about the topic?

EQ5D-3L and the newer 5L version have been shown to value changes in health very differently in the UK.

ii. What does the paper add to existing knowledge?

This paper demonstrates how health benefit and cost effectiveness may be affected from using 5L instead of 3L in 6 countries that have value sets for both.

iii. What insights does the paper provide for informing health care-related decision making?

3L and 5L cannot be treated as interchangeable. For some countries, there may be good reason to seek to improve current 5L value sets before making a decision to use 5L instead of 3L routinely.

INTRODUCTION

Measuring health status through use of the EQ-5D classification systems and valuing health using its associated tariffs or "utility" scores is a central element of assessing health benefits of health technologies in large numbers of cost-effectiveness analyses. EQ-5D is in such widespread use because it offers analysts and decision makers a standardised approach to the valuation of health benefits across many disease areas and intervention types.

We have previously demonstrated how moving from the three level version (EQ-5D-3L) to the newer, five level version (EQ-5D-5L) is likely to have profound impacts on cost-effectiveness estimates for different health care technologies when conducting analyses using the UK/English tariffs[1,2]. The advantage of consistency that EQ-5D offered decision makers, does not extend to comparisons between cost effectiveness analyses using these different versions in the UK/England. Policy makers in the English health system insisted on additional quality assurance of the underlying 5L value set before committing to changes with potentially profound implications for patient care[3]. Findings of that quality assurance raised concerns about data quality and the econometric modelling. The National Institute for Health and Care Excellence (NICE) and the Department of Health and Social Care (DHSC) determined that a new valuation study should be conducted[4]. No other country has yet followed the lead of authorities in the UK in this regard, despite the fact that similar, though not identical methods have been used to construct 5L value sets. The choice of using 3L or 5L, in all countries that use cost effectiveness to inform decisions, is not a simple matter for academic debate but one that has potentially profound impacts on patients, providers of health care services and the incentives provided to manufacturers to bring different types of health technologies to market in the future.

Therefore, in this study, we turn our attention to the likely implications of moving from 3L to 5L for other countries, using the same approach used previously to highlight the likely impact in the UK/England. We take a number of case study economic evaluations that were conducted alongside clinical studies where patients completed the 3L instrument. The analysis asks the question, what would the results of these studies have been if the 5L instrument had been used instead? We attempt to answer this question by using a published statistical model[5] that links 3L and 5L responses together. Analyses presented in our previous paper, and now in this study, are based on methods with proven reliability for mapping between 3L and 5L responses.

In this paper, we examine the impact of moving from 3L to 5L in the six countries, other than England, that have both official 3L and 5L tariffs at the time of analysis (2019), according to the EuroQoL group website: Japan, Korea, Netherlands, China, Spain and Germany. This is a rapidly changing arena with new 5L tariffs constantly being produced for other countries. Eight case study cost-effectiveness analyses, reporting 11 comparisons, originally conducted using the 3L instrument, are replicated using 5L estimated values for each patient observation. All other methods are identical to those reported in our previous estimates of the impact of moving from the 3L to 5L in the UK/England. The paper isolates the predicted impact of switching between 3L and 5L value sets within each country, using a fixed set of cost effectiveness studies and their associated costs.

METHODS

We use the flexible, copula-based model estimates of Hernandez and Pudney[5] to take patient level responses to the 3L instrument and predict the 5L utility score for the six countries of interest using the command eq5dmap in Stata[6]. While two versions of the Hernandez and Pudney model are available, the analyses reported here are all based on the version derived from data collated by the EuroQoL Group in 2009-2010. It comprises 3551 pairs of 3L /5L responses from 6 European countries. 3L values were calculated directly from the respective published tariffs. For both 3L and 5L, we used the versions endorsed by the EuroQoL group and the associated publications and reports for any further details on how they were derived.

Key features of the 3L and 5L tariffs

3L

Table 1 shows the key features of the studies that generated the 3L tariffs for the 6 countries[7,8,9,10,11,12] . All report following similar approaches to the UK valuation study for 3L. They use conventional Time Trade-Off (TTO) methods for elicitation of values but exhibit substantial variation in the proportion of the 243 3L health states that were valued, and therefore the proportion of states whose utility values were extrapolated from the observed data. Rates range from 7% (Japan and the Netherlands) to approximately 40% (Korea and China). Other notable differences between countries include the explicit reporting of exclusions of entire respondent data where 3 or more inconsistencies are identified (Korea and China). All tariffs were based on the inclusion of the so-called "N3" term in valuation regressions, an additional decrement for any dimension score at level 3, with the exception of Japan. The German value set was based on analyses that excluded any coefficients that were not

statistically significant. This resulted in dropping those for both levels 2 and 3 of usual activities (making usual activities an irrelevant domain) and level 2 for anxiety/depression.

5L

Development of 5L value sets has been subject to an approach intended to be more uniform than was the case for 3L. The "EuroQoL Valuation Technology" (EQ-VT) prescribes the numbers of respondents, the form and conduct of the experimental tasks and the health states to be valued. Thus, many aspects of study design are identical across 5L valuation studies. All studies included a set of 10 lead time-trade-off tasks and 7 discrete choice experiments (DCE) as part of the interview process, though not all countries included both sets of data as part of the final value set model. EQ-VT stipulates a sample size of approximately 1,000 respondents, coverage of 86 states in the TTO experiments (2.75% of the possible 3125 5L states) and 196 pairwise comparisons (0.01% of the feasible set).

However, the EQ-VT itself has changed. Three countries here used version 1.0: Netherlands, China, Spain. This is the same version used by the English value set. In response to major concerns about data quality, and following a one year moratorium on 5L studies, v1.0 was superseded with v1.1. In v1.1, interviews failing quality control criteria were flagged as being of suspect quality, allowing feedback to interviewers, retraining and deletion of suspect data. Four criteria were used, three of which relate to the process of the interview (e.g. the time to explain the preliminary wheelchair example) and one relates to extreme inconsistencies in valuations relative to 55555. The Japanese and Korean studies were conducted using v1.1. Version 2.0 built on v1.1 with the addition of a feedback module to respondents as an internal check on their responses. The German valuation study used this version (see Table 1)

The English value set was based on a so-called hybrid statistical model that combines TTO and DCE responses. Three countries followed a similar modelling strategy (Japan, Spain and Germany) whilst the remaining three only included modelling of the TTO data for the preferred value set (Korea, Netherlands, China).

Cost effectiveness case studies

We used the copula mapping models in eight cost-effectiveness case studies. All of these case studies were used in the previous comparison of 3L and 5L using the UK/English tariffs[1]. One study that was previously included

(the Weight-Reduction Activity Programme (WRAP)[13]) could not be reused here due to lapsed permissions in accessing the patient level data.

All included studies were economic evaluations based on individual patient level data using 3L. They were identified and included solely on the basis of pragmatic decisions relating to data access. They do however, span a range of conditions, interventions and patient types. These details are replicated from Hernandez et al[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[14]. 3L was administered to patients at baseline, 6, 12, 18 and 24 months[15]. We report on three pairwise comparisons each using methotrexate monotherapy as the standard comparator.
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[16]. 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. This study therefore contributes two comparisons to the results. 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)[17]. 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[18] . 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)[19,20]. 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[21] 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) CvLPRIT - The CvLPRIT (Complete- compared to Lesion-Only Revascularisation For Myocardial Infarction) trial[22] 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.

RESULTS

For most countries the impact of moving from 3L to 5L is to lower the incremental QALY gain in the majority of comparisons made in the studies (see Figure 1). Appendix Tables 1 to 8 provide the full results for each country. For Spain, as in England, the direction of change as a result of this movement is consistent. Incremental QALYs are lower using 5L than 3L in all but one case study comparison: the COUGAR 2 trial of docetaxel for osophegeal cancer, where both quantity of life gains and quality of life gains are important. All other case studies were of interventions that focussed on quality of life improvement. For the other countries, results are more mixed. For the Netherlands, Korea and Germany 7/11 case studies (64%) have lower QALY gains with 5L. These are largely, though not exactly, the same case studies. The equivalent figure is 6/11 (55%) for China and 4/11 (37%) for Japan. It is notable that the COUGAR2 study results shows lower incremental health gain associated with 5L (the opposite result to that shown for Spain) for every other country.

Figure 2 displays the mean magnitude of change in incremental QALYs as a result of conducting analyses using 5L compared to 3L, categorised by country and by case study comparisons, according to whether the change was to increase or decrease incremental QALYs.

For those cases where 5L leads to lower incremental health gain, the change is often substantial. Using the tariffs for the Netherlands, in the case of the SHARPISH trial, the net health gain using 3L is reversed to a net health loss using 5L, though the incremental QALY gain was particularly small in this example. The mean and median reductions in health gain, in those case studies where 5L does lead to lower health gain, are largest in the Netherlands (84% mean reduction, 41% median reduction), Germany (68% and 27%) and Spain (30% and 31%). For the Netherlands the mean reduction was 28% (median 30%) excluding the SHARPISH result. It should be noted that the country that has the lowest mean reduction in health gain is China with a 10% change.

For most countries, those studies where 5L leads to lower health gain see larger reductions than the gains in studies showing the opposite tendency. This is starkest for the Netherlands where the mean reduction is 84% (median 41%) compared to a mean 11% increase in those studies where 5L leads to a rise (11% median). China and Japan show the opposite tendency – the magnitude of health gain in those studies where 5L leads to a rise is greater than the health loss in those studies where it leads to a loss. For China the mean increase is 29% compared to a 10% decrease. For Japan, these figures are 23% and 15% respectively.

In all countries, the changes in incremental QALYs would result in changes in incremental cost effectiveness that span a £30k (or equivalent) threshold in at least 1 of the 11 case study pairwise comparisons. The CVLPRIT study results were particularly sensitive to the change from 3L to 5L in terms of spanning a notional £30k per QALY threshold. In all countries except China, the movement from 3L to 5L raised the ICER from below £30k to above it. For the Netherlands there are 2/11 comparisons that would span the £30k boundary, and for Japan the figure is 4/11.

DISCUSSION

Changes to estimated health gain, and consequently cost effectiveness, when moving from 3L to 5L are a result of changes between the descriptive systems in 3L and 5L (which in turn differ by language), the ways in which

respondents change their responses from 3L to 5L, the methods used to estimate the tariffs and changes between sampled populations over time. Whilst it may be natural to see the 5L descriptive system as a minor evolution of 3L because of the similarities in domains and wording, our analysis shows that individuals do not provide responses that align with this view. Of course, this could simply reflect genuine changes in preferences over time but even if this were the sole explanatory factor, the analysis illustrates the dangers of treating 3L and 5L as if they were interchangeable. Consistent with this, the degree of difference between 3L and 5L varies by country because many aspects of both describing health and valuing it, both in the 3L and 5L systems, are country specific.

As with England, the Spanish value set exhibits large and predictable changes between 3L and 5L. Of the 6 countries we have analysed, it is only Spain that used the same, early version of the valuation methods (EQ-VT v1.0) and also combined the TTO and DCE data in a hybrid model. All other countries in our analysis that used the now abandoned early version of the EQ-VT discarded the DCE data. Given concerns with both the quality of the data and the hybrid model in the English context[3], coupled with widely expressed concerns about EQ-VT1.0 more generally[23], such as interviewer effects, confusion or fatigue amongst respondents, this is concerning for health care decision-makers in Spain.

Results for China and the Netherlands do not follow quite such a clear pattern. These two countries both used version 1.0 of the valuation methods but based the final tariff only on the TTO data. The design of the valuation tasks in the 5L valuation is such that very few checks on data quality can be made of the DCE data. It is notable that in those comparisons using the tariff values for the Netherlands, and in those cases where incremental QALYs decreased following use of the EQ5D-5L, the mean magnitude of change was greater than for either England or Spain.

Our case studies were generally selected on the basis that we were able to access the patient level data, not to represent any specific disease or intervention types. By virtue of the number of case studies they do inevitably span quite a range but we would not claim they are representative of any set of issues any individual decision maker is likely to face. These are all UK based studies. We do not believe this impacts conclusions about the likely impact of switching from 3L to 5L in different countries and note that it is common for assessments of

many technologies to draw on clinical studies conducted in many different countries, using many different language variants of quality of life instruments. We do not seek to make any claims about the cost effectiveness of the intervention case studies in other countries. All discussions of ICERs are based on UK costs and a UK threshold in order to isolate the impact of the different value sets for different countries. We focus on the changes to incremental health gains because costs are unrelated to the use of 3L/5L. Many of our case studies are far from commonly agreed cost effectiveness boundaries and, thus, even the large changes in incremental health gain that the shift to 5L implies rarely changes this. In those settings where cost effectiveness is often at the boundary of decision maker thresholds, for example in the evaluation of new pharmaceuticals, the shift from 3L to 5L and the resultant changes in estimated incremental health gain will be likely to change reimbursement decisions and/or require manufacturers to change the prices of their products substantially. The policy choice of using 3L or 5L, in all countries that use cost effectiveness to inform decisions, is not a simple matter for academic debate but one that has profound impacts on patient care, those that work in health services to provide that care and the incentives provided to manufacturers to bring future health technologies to market.

It is often claimed that mapping is an approach that is inherently biased and that studies like this, which rely on mapping of the relationship between responses observed in large datasets where respondents were asked to complete both 3L and 5L, are of limited value, see for example Yang et al[24]. It is certainly true that analysts have tended to use mapping methods that are not theoretically well suited to the task and these have, predictably, led to results that are not reliable. However, poor practice should not be confused with the underlying principles of mapping. The application of appropriate methods in the general mapping literature[25,26,27] has repeatedly been demonstrated to eliminate such bias. In the context of 3L/5L, the mapping that the analysis presented here is based on uses novel, flexible methods, demonstrated to suffer none of the biases of naïve methods[5,28]. Concerns that do remain stem from the features of the dataset used for the mapping study.

It has been suggested that a better approach for understanding the true relationship between 3L and 5L is via direct application of both for patients recruited to clinical studies. Notwithstanding the ethical and practical difficulties associated with increasing patient burden in such studies, to do so relies on the assumption that the inclusion of both instruments yields patient responses that are identical to those that would have been observed had they completed only one. This is also assumed in mapping studies. It is an assumption that has not yet been

tested, the validity of which may well differ according to study design (e.g. the degree of separation between the two instruments in a patient survey, the mode of administration), and may not be consistent with evidence from the survey literature in other scenarios (see for example McFarland[29]). This is a weakness of our study, but it is a weakness that may be more pronounced in approaches that use direct observation of both instruments in a clinical study setting.

CONCLUSION

Moves to adopt EQ5D-5L lead to substantial changes in estimates of cost-effectiveness compared to 3L in all countries. These differences are predictable in Spain which shows a similar pattern to previous analyses for England/UK: incremental QALY gains are smaller with 5L where these gains are generated mostly from quality of life improvements. Other countries show less predictable changes.

These findings should prompt further investigation into the reasons for differences that this study cannot answer. This should not be restricted to the 5L valuation process, where some clear concerns have previously been raised, but should also consider the potential impact of wording in the both the 3L and 5L systems (where language specific anomalies may be relevant) and the relevant 3L valuations.

Irrespective of the reasons for observed differences, our findings suggest that 3L and 5L cannot be considered interchangeable.

---

[1] Hernandez Alava, M., Wailoo A., Grimm S., Pudney S., Gomes M., Sadique Z., Meads D., O'Dwyer J., Barton G., Irvine L. (2018) "EQ-5D-5L versus 3L: the impact on cost-effectiveness in the UK", Value in Health, Vol.21(1):49-56.

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

[3] Hernandez Alava M, Pudney S, Wailoo A. (2018) "Quality review of a proposed EQ5D-5L value set for England", EEPRU Report, available at http://www.eepru.org.uk/wp-content/uploads/2017/11/eepru-report-eq-5d-5l-27-11-18-final.pdf (last accessed 25th Feb 2019)

[4] https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/technology-appraisal-guidance/eq-5d-5l (last accessed 31st October 2019)

[5] Hernandez Alava M, Pudney S. (2017) "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, Vol.55:139-152.

[6] Hernandez Alava M, Pudney S. (2018) "Eq5Dmap: A Command for Mapping between EQ-5D-3L and EQ-5D-5L", The Stata Journal, Vol.18:395-415.

[7] Tsuchiya A, Ikedab S, Ikegamib N, Nishimurac S, Sakaid I, Fukudae T, et al (2002) "Estimating an EQ-5D population value set: the case of Japan", Health Economics, Vol. 11: 341–353

[8] Lee Y, Nam H, Chuang L, Kim K, Yang H, Kwon I, et al. (2009) "South Korean Time Trade-Off Values for EQ-5D Health States: Modeling with Observed Values for 101 Health States", Value in Health, Vol12: 1187-1193.

[9] Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ. 2006;15(10):1121-32.

[10] Liu GG, Wu H, Li M, Gao C, Luo N. Chinese time trade-off values for EQ-5D health states. Value Health. 2014;17(5):597-604.
[11] Badia X, Roset R, Herdman, M, Kind P.A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making 2001; 21(1): 7-16
[12] Claes C, Greiner W, Uber A, Graf von der Schulenburg JM. An interview-based comparison of the TTO and VAS values given to EuroQol states of health by the general German population. n: Greiner W, J-M. Graf v.d. Schulenburg, Piercy J, editors. EuroQol Plenary Meeting, 1-2 October 1998. Discussion papers. Centre for Health Economics and Health Systems Research, University of Hannover, Germany. Uni-Verlag Witte, 1999; 13-39.
[13] 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
[14] 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.
[15] 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
[16] 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.
[17] IMPROVE Trial Investigators. Endovascular or open repair strategy for ruptured abdominal aortic aneurysm: 30 day outcomes from IMPROVE randomised trial. BMJ 2014;348:f7661
[18] 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
[19] 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
[20] 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)
[21] 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.
[22] 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.
[23] Shah, K., Rand-Hendriksen, K., Ramos, J.M., Prause, A.J. and Stolk, E. (2014). Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. 31st Scientific Plenary Meeting of the EuroQol Group, Stockholm September 2014; Proceedings: 1-18. Available at: http://eq-5dpublications.euroqol.org/download?id=0_53918&fileId=54332
[24] Yang F, Devlin N, Luo N. (2019) “Cost-Utility Analysis Using EQ-5D-5L Data: Does How the Utilities Are Derived Matter?” Value in Health, Vol.22:45-49.
[25] Hernandez Alava, M., Wailoo, AJ, and Ara, R. (2012) “Tails from the Peak District: Adjusted Limited Dependent Variable Mixture Models of EQ-5D Health State Utility Values”, Value in Health, doi:10.1016/j.jval.2011.12.014.
[26] Hernandez Alava, M., Wailoo, A., Wolfe, F., Michaud, K. (2014) “A comparison of direct and indirect methods for the estimation of health utilities from clinical outcomes.”, Medical Decision Making, Vol: 34:919–930.
[27] Gray L., Hernandez Alava M., Wailoo A. (2018) “Development of methods for the mapping of utilities using mixture models: Mapping the AQLQ-S to EQ-5D-5L and HUI3 in patients with Asthma”, Value in Health, Vol.21:748-57

[28] Hernandez Alava M, Pudney S, Wailoo A. (2017) "Methods for mapping between the EQ-5D-5L and the 3L for Technology Appraisal", Report by the NICE Decision Support Unit, available at http://nicedsu.org.uk/wp-content/uploads/2017/05/Mapping-5L-to-3L-DSU-report.pdf (last accessed 25th Feb 2019)

[29] McFarland S. (1981) "Effects of Question Order on Survey Responses", Public Opinion Quarterly, Vol.45: 208–215,

Tables

Table 1: Characteristics of included value sets

	N States valued	(%)	Method	N3 term	n interviews	n respondents in analysis	inconsistency rate (weak)	5L VT version	Method
Japan	17	(7.00)	TTO	No	543			1.1	TTO/DCE
Korea	101	(41.56)	TTO	Yes	1307	1264	17.80%	1.1	TTO
Netherlands	17	(7.00)	TTO	Yes	300	298		1.0	TTO
China	97	(39.92)	TTO	Yes	1222	1147	33%	1.0	TTO
Spain	43	(17.70)	TTO	Yes	1000	975		1.0	TTO/DCE
Germany	35	(14.40)	TTO	Yes	344	339		2.0	TTO/DCE

Figures

Figure 1: Percentage of case studies in which the move from EQ5D-3L to 5L led to a reduction in incremental QALYs gained.

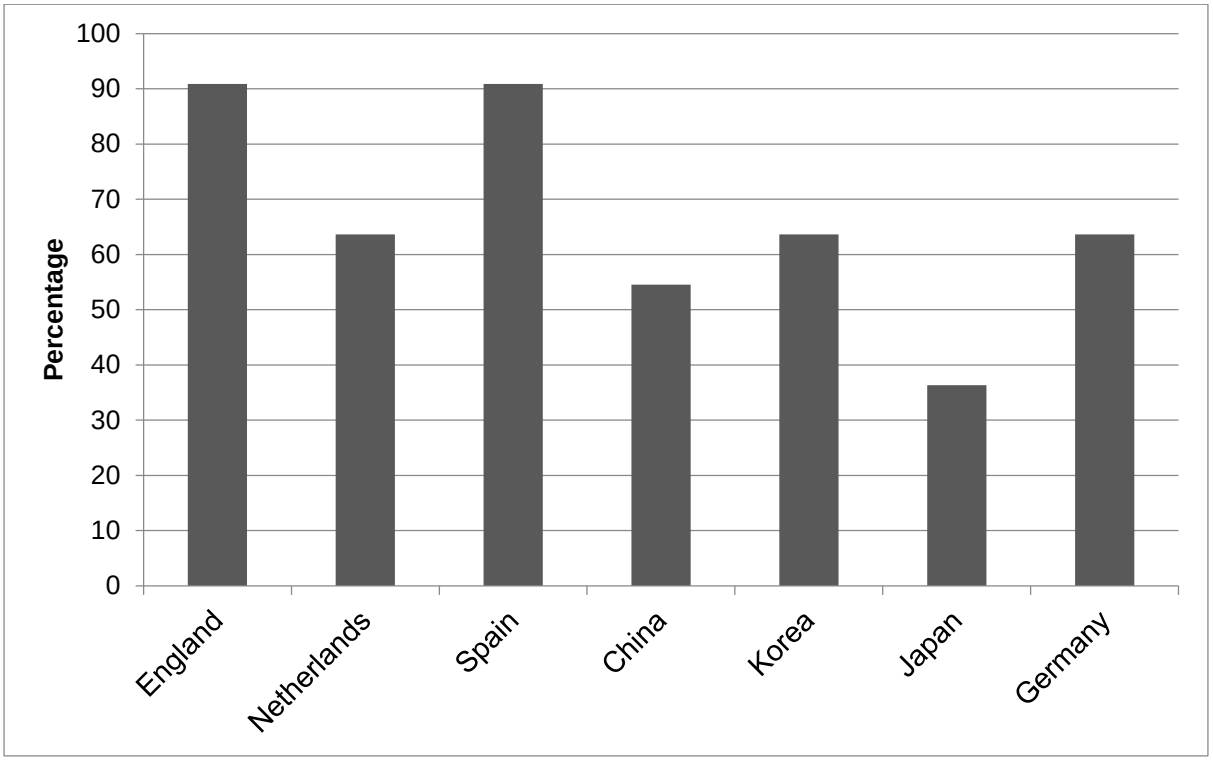

Figure 2: Mean percentage change in incremental QALYs as a result of moving from EQ5D-3L to 5L by direction of change.