Monograph:
Rowen, D, Young, T, Brazier, J et al. (1 more author) (2011) Comparison of generic, condition-specific and mapped health state utility values. Discussion Paper. HEDS Discussion Paper (11/06). (Unpublished)

HEDS Discussion Paper 11/06


ScHARR
School of Health and
Related Research

Comparison of generic, condition-specific and mapped health state utility values

Donna Rowen[a] (PhD, MSc, BA), Tracey Young[a, b] (PhD, CStat, MSc, BSc), John Brazier[a] (PhD, MSc, BA), Sabine Gaugris[c] (MSc, MSc)

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

[b] NIHR Research Design Service for Yorkshire and the Humber

[c] Janssen-Cilag Ltd., High Wycombe, UK

Corresponding author: Donna Rowen, Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA. d.rowen@sheffield.ac.uk
Tel: +44(0)114 222 0728. Fax: +44(0)114 272 4095.

Financial support: This study was partly funded by MRC-NIHR Methodology Research Programme, project number 06/97/04**.**

Keywords: Preference-based measures, QALYs, utility, mapping, condition-specific measures

Running title: Comparison of generic, condition-specific and mapped utilities

Acknowledgements

We would like to thank Janssen-Cilag Ltd for use of the data. The article has not been published elsewhere and there are no conflicts of interest.

Abstract

Objective: Resource allocation informed by cost-utility analysis requires that the benefits are comparable across patient groups and interventions. One option is to recommend the use of one generic utility measure, but this raises the issue of comparability when the preferred measure is inappropriate or unavailable. Many cancer trials do not include generic measures such as EQ-5D and instead include condition-specific measures and use these to generate utility estimates. We analyse the comparability of generic, condition-specific and mapped utility values for a Multiple Myeloma cancer patient dataset.

Methods: Generic EQ-5D, condition-specific EORTC-8D and mapped EQ-5D utility values are compared using psychometric and statistical analysis to determine discrimination across severity groups, responsiveness and agreement.

Results: Generic, condition-specific and mapped utility estimates were responsive and show discriminative validity. EQ-5D had higher responsiveness and detected a greater change across severity groups and treatment periods than EORTC-8D, but has a higher proportion of responses at full health (12.2%). Differences in EQ-5D and EORTC-8D were due to both differences in classification system and preference weights.

Conclusion: Our findings suggest that condition-specific EORTC-8D or mapped EQ-5D utility estimates are comparable to directly obtained EQ-5D utilities. EORTC-8D estimates captured problems in quality of life at the upper end of the utility scale that were not captured by EQ-5D, but estimated lower utility gains than the use of EQ-5D directly.

Introduction

Resource allocation informed by economic evaluation using cost-utility analysis has become increasingly popular in recent years. This analysis requires that the measures of benefit and cost for each evaluation are comparable both across different patient groups and different interventions. Benefit is measured using quality-adjusted life years (QALYs) which are a measure of both quantity and quality of life. Often generic preference-based measures such as the EQ-5D (1), HUI3 (2) or SF-6D (3;4) are used to calculate the 'Q' component of the QALY. However it is well documented that different generic measures produce different results when applied to the same patient group at the same point in time (5). This raises issues for comparability, and one solution is to recommend the use of a single measure for all evaluations. This is the approach taken by the National Institute of Health and Clinical Excellence (NICE) (6) where the most commonly used generic measure, EQ-5D, is recommended for use in all technology appraisals. This raises the question of how utility values should be generated if EQ-5D is either unavailable or inappropriate, and the comparability of evaluations undertaken in these circumstances.

Cancer is one condition where it remains unclear whether the generic EQ-5D is appropriate, but is further complicated by the fact that EQ-5D is often unavailable as many cancer trials do not include it. NICE state that if a measure is thought inappropriate empirical evidence should be provided demonstrating why it is inappropriate, covering properties such as content validity, construct validity, responsiveness and reliability. A recent report argues that the EQ-5D may not be sufficiently sensitive to capture changes in health status of cancer patients, as, for example, there is no EQ-5D dimension to specifically capture changes in vitality or energy (7). However there is little guidance provided by NICE or similar agencies of when a measure can be deemed inappropriate for a patient group or intervention, and this is an area requiring further research and guidance. If EQ-5D is inappropriate, NICE state that other measures can be used (6).

Clinicians and researchers often choose to include condition-specific profile measures in trials rather than generic preference-based measures such as EQ-5D. Condition-specific profile measures, such as EORTC QLQ-C30 are often included as these capture the effects of interventions across a wide range of relevant symptoms, side effects and aspects of functioning and quality of life and their validity is well established. These profile measures have great clinical utility and are recommended by the US FDA (8), whereas the EQ-5D is recommended only for economic evaluation and can be viewed as being an additional burden for completion for patients who are very unwell. However these condition-specific profile measures typically provide a description rather than a valuation of health and cannot be used to populate cost-effectiveness models. In recent years there has been a growth in preference-based measures derived from existing condition-specific measures that enable these measures to be used directly to generate utilities. The EORTC-8D is a recently developed

condition-specific preference-based measure derived from EORTC QLQ-C30 for use in cancer patients (9). This measure allows a utility estimate to be generated for every individual each time the EORTC QLQ-C30 is used and enables the direct estimation of utility without placing any burden on patients to complete an extra measure or additional questions. Mapping is an alternative method that can be used to obtain utility values when only a condition-specific non-preference-based measure was included in the trial. Mapping applies the statistical relationship between, for example, QLQ-C30 and EQ-5D to obtain predicted EQ-5D values from QLQ-C30 data. This relationship is typically obtained by estimating regressions on a separate dataset which has similar patient characteristics to the trial. Published mapping algorithms are available that map the condition-specific QLQ-C30 onto EQ-5D, and these algorithms can be applied to the trial dataset to produce EQ-5D estimates. If EQ-5D is unavailable in a trial, NICE (6) recommend that either mapping or other validated measures are used to produce utility values. NICE stipulate that the mapping must be based on empirical data and the other measures should have valuation methods that are comparable to those used for EQ-5D (MVH tariff) (10).

A small number of studies have examined the impact of using mapped EQ-5D estimates rather than directly generated EQ-5D utilities, finding different results across studies (11-13). A large number of studies compare the performance of EQ-5D to the other main generic preference-based measures such as SF-6D and HUI2 (see (5) for an overview) but there are few comparisons of condition-specific and generic preference-based measures (see (14)). Furthermore as far as the authors are aware no study has examined the comparability of all preferred options for use in technology appraisals to agencies such as NICE; as although EQ-5D is the preferred option, under certain circumstances other generic, condition-specific or mapped EQ-5D utility estimates can be used.

This paper compares utility values generated using the EQ-5D to utilities generated using a condition-specific preference-based measure and mapping for a cancer patient dataset. We compare utility values obtained using: generic preference-based EQ-5D; condition-specific preference-based EORTC-8D derived from EORTC QLQ-C30; and two published algorithms mapping QLQ-C30 onto EQ-5D. We further compare the performance of EORTC-8D and EORTC QLQ-C30 summary scores to determine whether the EORTC-8D maintains the desirable properties of the original measure. This paper seeks to inform researchers and policy makers in their choice of source of utility values and interpretation of these values regarding discrimination across severity groups, responsiveness and agreement.

Summary of measures

EQ-5D

EQ-5D has 5 dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) each with 3 levels of severity from no problems to severe problems (1).

The health state classification system describes 243 unique health states and utility values range from 1 to -0.594 for the UK value set collected in the Measuring and Valuing Health (MVH) study (10).

EORTC QLQ-C30 and EORTC-8D

The QLQ-C30 is widely used in cancer clinical trials in Europe and Canada (15) and has been found valid for many cancer conditions. The QLQ-C30 has 30 questions that cover functioning (physical, role, social, emotional and cognitive functioning) and common cancer symptoms (pain, fatigue, nausea, vomiting, dyspnea, appetite loss, sleep disturbance, constipation and diarrhea) plus financial impact of the disease and treatment (excluded from analyses here as this is inappropriate for inclusion in health-related quality of life measurement to generate QALYs). The QLQ-C30 has fourteen summary scales ranging from 0 to 100, each representing an aspect of functioning (5 summary scales, higher scores represent higher functioning) or a particular symptom (9 summary scales, higher scores represent greater symptoms), with one additional global quality of life scale.

The EORTC-8D has 8 dimensions (physical functioning, role functioning, pain, emotional functioning, social functioning, fatigue and sleep disturbance, nausea, constipation/diarrhoea) each with 4 or 5 levels of severity. The health state classification system was derived from 10 QLQ-C30 items and describes 81,920 unique health states with a range of utility values from 1 to 0.291 (9).

Methods

Utility values were generated using the available preference weights for EQ-5D and EORTC-8D for each patient at each time point in the dataset. Mapped utility values were also estimated for each patient at each time point using published algorithms described below.

Estimating EQ-5D utilities by mapping QLQ-C30 onto EQ-5D

The easiest way to produce mapped estimates is to use published algorithms. Six published algorithms use mapping to produce utilities using EORTC QLQ-C30 data, two of which were used here (11;16). The other four algorithms are not used here as one paper requires FACT data not available in our dataset (17), one paper maps to patient TTO values rather than EQ-5D (18), one paper uses only females as their patient group (19), and one paper does not publish the mapping function (20). Patient valuations of own health using preference elicitation techniques such as time trade-off or visual analogue scales are not preferred by agencies such as NICE or Washington Panel of Cost Effectiveness (21) as public preferences are preferred given that public funding is often used to provide healthcare. Patient values can also be affected by vested interests and can be difficult to obtain on patients who are very unwell as to generate QALYs the utility values must be anchored against death on a full health-death 1-0 scale. For mapping to provide accurate and appropriate estimates the

patient group used to estimate the mapping algorithm should be representative of the patient group it is applied to. Although there is no evidence stating that sex affects the accuracy of mapped estimates, given that mapping algorithms estimated using datasets containing both sexes are available, these were chosen here in preference.

The first algorithm used here is an algorithm by McKenzie and van der Pol (11), who estimated OLS regressions on 877 patients with esophageal cancer. The regression equations used here include all 15 functioning scales and symptom scales from the QLQ-C30 to map onto EQ-5D. McKenzie and van der Pol also report regressions that predict EQ-5D responses to each dimension rather than to a utility score, yet found that these performed worse and so they are not used here.

The second algorithm by Kontodimopoulos et al. (16) estimated OLS regressions on 48 patients with gastric cancer. Explanatory variables included in the model were selected using a stepwise inclusion procedure and the remaining variables were the physical functioning, emotional functioning and global health status scales.

Cancer patient dataset

The analysis was undertaken for a sample of patients newly diagnosed with multiple myeloma cancer. The data was collected in VISTA, a phase III randomized open-label trial (ClinicalTrials.gov number, NCT00111319) completed in June 2007. Patients were requested to complete both the EQ-5D and EORTC QLQ-C30 at their screening visit, day 1 of each of the 9 cycles of treatment, end of treatment visit and during the post treatment phase (every 6 or 8 weeks) until disease progression. To remove any differences in analyses due to missing values, observations are included in the analysis only where both EQ-5D, EORTC-8D and QLQ-C30 items used in each of the mapping algorithms are available.The dataset used here contains 5650 observations in total across 674 individuals and 16 time periods (all periods in the trial where n>70). Mean age of the sample is 71.58 (standard deviation of 5.25) and 50.8% of the sample is female.

Analysis

Psychometric and statistical analyses were used to compare the utility estimates produced using different methods and EORTC QLQ-C30 summary scores.

Validity: Discrimination across different severity groups

Construct validity is examined by assessing ability to discriminate between patients with different levels of severity. It is important that a utility measure or method of producing utilities can discriminate correctly amongst groups of different severity as this determines whether the utility values measure an improvement in quality of life due to a health improvement in the condition of interest. The Karnofsky Performance Scale is reported by the doctor and

classifies patients according to functional impairment typically using 10 point markers, where a score of 100 indicates that the patient is normal with no signs of disease and a score of 0 is equivalent to death (22). As clinical severity is conceptually different to quality of life discrimination was also captured across different groups according to self-reported quality of life using item 30 from the EORTC QLQ-C30 ('How would you rate your overall quality of life during the past week? Please circle the number between 1 and 7 that best applies to you, 1=very poor and 7=excellent'). Discrimination was examined using the statistical significance of differences using an overall F-test from an ANOVA and the sensitivity of differences using standardised effect size (ES). ES is estimated using the difference in mean scores between two adjacent sub-groups of study participants with different levels of severity divided by the standard deviation of scores for the mildest of the two sub-groups. Utilities were also plotted for severity groups categorised according to the Karnofsky Performance Scale and self-reported quality of life.

Responsiveness to change over time

Responsiveness is the sensitivity of a measure to known changes in health over time. Here this is examined in terms of sensitivity to change in trial data before and after treatment across all study arms. Responsiveness was examined using floor and ceiling effects, standardised response mean (SRM), ES and t tests. Floor and ceiling effects report the percentage of patients in either full health or the most severe health state 'PITS', where a high percentage indicates that the measure is unable to capture either an improvement or deterioration in health respectively. Relative floor and ceiling effects are important as they indicate that one measure cannot distinguish whereas another can. SRM is the mean change score of a measure between two different time points divided by the standard deviation of the change score (23). ES in this case is the mean change score of a measure between two different time points divided by the standard deviation of the score at baseline. Standardised response mean and effect size are generated to assess the responsiveness of the different methods between screening and cycle 9 of the trial and between screening and end of treatment. These points were chosen as screening represents the only period before treatment (n=604), cycle 9 represents the end of treatment for patients completing all 9 treatment cycles (n=283) whereas end of treatment includes respondents at the end of their treatment (n=406). Statistical significance of any difference is examined using t-tests. Utilities generated using each method were plotted by period to determine whether they show comparable movements in quality of life throughout the trial. All statistics were reported using all responses where observations were available for every measure of interest.

Correlation and agreement

The estimates produced using the different methods are compared using Pearson correlation coefficients and the intraclass correlation coefficient (ICC). ICC assesses the consistency of

the methods given that they are all generating utility values on the same 1-0 full health-dead scale.

Differences across classification systems and preference weights

Further analysis was undertaken to determine why any differences in values were observed between EQ-5D and EORTC-8D. This analysis can also be used to highlight whether the measures have content validity for this patient group. If EQ-5D and EORTC-8D produce different values possible explanations include: classification system; preference weights for the classification system; recall period. The recall period is likely to explain some difference as EORTC-8D measures health during the past week whereas EQ-5D measures health today. This is an issue for research using qualitative analysis and is beyond the scope of this paper. Further analysis was here undertaken to explore differences due to the classification system and preference weights.

Differences in classification system were examined using Spearman rank correlations of each dimension of each measure. Observed frequencies of each dimension are also reported when each measure is at full health to determine differences across the measures in ability to detect a health improvement at the ceiling of the measure.

Differences in preference weights were analysed using different weightings for EQ-5D. Two alternative sets of EQ-5D utility weights were used here (24;25). The first alternative set of preference weights by Craig and Busschbach (25) were derived by remodelling the time-trade-off (TTO) values collected in the MVH study that were used to produce the standard UK value set (here referred to as the MVH value set) (10). The model was developed to deal with a criticism of how worse than dead responses were modelled to produce the UK MVH value set (25). The second alternative set of preference weights by Yang el al. were estimated on TTO data collected in a separate study using a small UK general population sample (n=81) (24).

Results

Discrimination

Severity groups were generated using the Karnofsky performance scale and overall quality of life (using a QLQ-C30 item). EORTC-8D had generally higher effect sizes than EORTC QLQ-C30 and similar effect sizes to EQ-5D (see Table 1). Mapped estimates using both methods had higher effect sizes than EQ-5D and the highest effect sizes when severity groups were divided by overall quality of life, most likely due to the inclusion of QLQ-C30 global health status in both mapping algorithms. The difference in scores across adjacent severity groups was statistically significant at the 1% level for all measures. The EORTC-8D had a narrower range of mean values across severity groups than EQ-5D (0.597 to 0.852 compared to 0.259 to 0.810 for severity groups defined using the Karnofsky performance scale). Despite these

differences, the smaller standard deviation of EORTC-8D resulted in similar effect sizes to EQ-5D. Figure 1 parts a) and c) indicate that EORTC-8D values have a much shallower gradient than the other methods, showing smaller differences across different severity groups.

Responsiveness

The range of utility values covers the full severity range for EQ-5D and EORTC-8D, yet the mapped EQ-5D estimates do not reflect the full range of severity at the lower end (Table 2). EQ-5D and mapped EQ-5D estimates have a much larger utility range than EORTC-8D due to the differences in the utility ranges of the measures. EQ-5D and the Kontodimopoulos et al mapped estimates suffer from ceiling effects (12.2% and 11.4% respectively). EORTC QLQ-C30 summary scores also suffer from ceiling effects (up to 80.3% for one symptom summary score), yet the EORTC-8D does not.

The mapped estimates have values above 1 as the mapping algorithms used here were estimated using OLS, meaning that predictions are not constrained to 1. As EQ-5D values greater than 1 are impossible to obtain this raises the issue of whether we should censor mapped EQ-5D estimates above 1. Censoring the values above 1 changes the mean and standard deviation of the mapped estimates, and although the change is minimal for the McKenzie and van der Pol estimates this would produce a large change for the Kontodimopoulos et al estimates from 0.703 (0.250) to 0.695 (0.239). As the mapping literature provides no guidance on this, the uncensored mapped estimates are used in all analyses. The McKenzie and van der Pol and Kontodimopoulos algorithms correctly predicted (1 or above) 19.5% and 51.8% respectively of observed EQ-5D values at 1.

All methods used to produce utility values show significant differences in utilities between screening and cycle 9 and between screening and end of treatment (Table 2). The size of the change in utilities varies across method with EQ-5D showing the largest mean change with the largest standard deviation (0.189 (0.337) between screening and cycle 9) and the EORTC-8D showing the smallest mean change and smallest standard deviation (0.049 (0.143) between screening and cycle 9). For the methods used to produce utilities the effect size and standardised response means are largest for the Kontodimopoulos et al estimates and smallest for EORTC-8D. Effect sizes and standardised response means for EORTC-8D are towards the upper range produced for the EORTC QLQ-C30 summary scores, which may be expected given that EORTC-8D is made up of 10 EORTC QLQ-C30 items. Overall the EORTC QLQ-C30 global QOL summary score has the highest effect size and standardised response mean. Figure 1e) indicates that there is a noticeable gap between the utilities for EQ-5D and EORTC-8D, where EORTC-8D values are always higher. The McKenzie and van der Pol mapped EQ-5D estimates follow a similar pattern to the EQ-5D, whereas the Kontodimopoulos estimates follow a similar pattern to EORTC-8D.

The mapped EQ-5D estimates have a lower standard deviation than observed EQ-5D estimates and this has been observed elsewhere in the mapping literature. Mapped values contain error, measured as the difference between predicted and observed values. Mean absolute error is 0.144 for the McKenzie and van der Pol estimates and 0.156 for the Kontodimpopoulos et al estimates. Error in predictions increases for more severe health states (analysis not reported, available on request), where MAE is more than doubled when observed EQ-5D is less than 0.5 compared to EQ-5D greater than or equal to.0.5, and a similar pattern has been previously observed in the mapping literature (26).

Correlation and agreement

Utility values generated using each of the values have high correlation coefficients (Table 3) yet the mapped estimates are more highly correlated with EORTC-8D than EQ-5D. This is perhaps unsurprising given that the mapped estimates and EORTC-8D are all generated using QLQ-C30 responses, but has not been explored in the mapping literature previously.

Differences across classification system and preference weights

The above analysis demonstrates differences in EQ-5D and EORTC-8D utilities. EQ-5D and EORTC-8D dimensions are most highly correlated where expected, such as EQ-5D pain/discomfort and EORTC-8D pain (Table 4). The correlations are below 0.5 between all EQ-5D dimensions and EORTC-8D fatigue, nausea and constipation/diarrhoea dimensions and between EQ-5D self-care and all EORTC-8D dimensions. This suggests differences in the quality of life captured by the two classification systems for these dimensions.

Table 5 summarises EORTC-8D responses when EQ-5D=1 meaning that the respondent is in full health, demonstrating that the EORTC-8D captures an impact on quality of life according to dimensions such as fatigue and physical functioning that are not captured in the EQ-5D for these patients. This is most noticeable for fatigue where 52.77% of observations at EQ-5D full health have fatigue in the EORTC-8D, similarly 45.71% of observations have problems in the physical functioning dimension. Table 6 summarises EQ-5D responses when EORTC-8D=1 meaning that the respondent is in full health, finding that the EQ-5D captures some differences in pain/discomfort and anxiety/depression not captured by EORTC-8D, but the proportion of these responses is small (14.08% and 9.23% respectively).

The use of alternative EQ-5D preference weights reduces the range and standard deviation of EQ-5D utility values and standard deviation and using the Craig and Bussbach weights raises the mean value (Table 7). Mean change is smaller using these alternative preference weights (0.059 to 0.098, Table 7) than using the standard UK EQ-5D weights (MVH value set (0.100 to 0.189, Table 2). Figure 1 parts b), d) and f) demonstrate the pattern of discrimination across different severity groups and utility values by period using different preference weights for EQ-5D. These figures indicate that the Craig and Busschbach estimates are much closer

to EORTC-8D estimates than the standard UK EQ-5D MVH value set. Despite a striking resemblance in these graphs summarising data at the mean level, this masks differences at the individual level. For example, the correlation between EORTC-8D and Craig and Busschbach EQ-5D values is 0.658 (other plots available from authors on request). The Yang et al EQ-5D estimates have virtually the same gradient in all of the figures as EORTC-8D, but with a systematic difference in mean utilities of around 0.13.

Discussion

Generic EQ-5D, condition-specific EORTC-8D and two published mapping algorithms were used to generate utility estimates for Multiple Myeloma patients in a clinical trial dataset. We observed differences in mean utilities and in mean change across time periods using the different methods, with the EQ-5D consistently showing the largest mean utility gain. However all methods were able to discriminate between severity groups measured using the Karnofsky Performance Scale and an overall quality of life item and were responsive. We further compared the performance of EORTC-8D to the EORTC QLQ-C30 measure it was derived from, finding that discriminative validity and responsiveness of EORTC-8D was comparable to the QLQ-C30 functioning and symptom summary scores, but inferior to the QLQ-C30 global quality of life summary score. The QLQ-C30 global quality of life score was excluded from the health state classification system of EORTC-8D as it is inappropriate for inclusion in a multi-attribute preference-based measure, yet was included in both the McKenzie and van der Pol and Kontodimopoulos et al mapping algorithms and their discriminative validity and responsiveness may in part be attributed to this. Analyses have been conducted using one patient dataset containing patients with one type of cancer and this is a limitation of this research. Replicating these analyses using data for other cancer types is recommended.

The analysis has been performed using observations with no missing data for any of the approaches used to generate utilities or EORTC QLQ-C30 summary scores. However, high levels of missing data mean that the utilities are not representative of the entire trial sample, and this is important as the data may be missing for systematic reasons where patients in the poorest health are unable to complete the appropriate questionnaire. Missing values for the overall dataset do vary by method, where overall the Kontodimopoulos et al estimates has the smallest proportion of missing values (2.2%), followed by EQ-5D (2.8%), with the EORTC-8D (4.4%) and the McKenzie and van der Pol estimates (5.1%) having the largest proportions.

Mapping is advantageous as it enables EQ-5D utilities to be estimated when EQ-5D was not included in the trial. However these mapped values contain error and should be considered only as a second best alternative to including EQ-5D directly in the trial. Here the McKenzie and van der Pol algorithm (11) produced more accurate EQ-5D estimates than the Kondimopoulos et al algorithm (16) yet both have high mean absolute error. We suggest that EORTC-8D utilities are more accurate than these mapped estimates as they do not contain

error. However NICE suggests that the mapped estimates should be presented in the main economic evaluation analyses and EORTC-8D utilities should be included in a separate analysis (6). NICE further recommend that when using condition-specific measures such as EORTC-8D researchers should indicate the extent to which their choice of instrument has impacted on the valuations. Our findings suggest that the use of EORTC-8D would generate lower mean change in utilities with smaller standard deviation than the use of EQ-5D to generate utilities for the same cancer patient group.

NICE recommend using the same valuation methodology as the UK valuation of EQ-5D to ensure comparability to EQ-5D when using condition-specific measures to produce utilities for use in economic evaluation (6). Yet differences between EQ-5D and EORTC-8D were observed despite the EORTC-8D using the same valuation methodology as the UK MVH valuation of the EQ-5D (10). Further analysis suggested that some of the differences were due to the classification system yet some of the differences were due to the preference weights. The alternative EQ-5D preference weights produced closer utility estimates to the EORTC-8D utilities than the use of the standard UK EQ-5D weights (MVH value set).

Here we have not analysed the impact these differences have on QALY estimates. It is possible that the differences in utility values will have an impact on QALY estimates and change in QALY estimates, particularly where survival differs across interventions. Two studies have compared mapped estimates to direct EQ-5D estimates for use to generate QALYs in economic evaluation (11;12). Although both studies found no significant difference between QALY estimates generated using mapping to EQ-5D and directly observed EQ-5D values, one of these studies found that incremental cost per QALY estimates differed across four interventions depending on whether mapped or directly observed EQ-5D values were used (12). Research analysing the impact on QALY estimates from using generic or condition-specific preference-based measures is recommended.

EQ-5D is not always included in cancer trials, sometimes because it is thought to be inappropriate or unresponsive. In contrast, condition-specific measures are often included, as they are thought appropriate and responsive. Here EQ-5D has higher responsiveness than the condition-specific EORTC-8D and was also able to discriminate between severity groups using the Karnofsky Performance Scale and an overall quality of life item from the QLQ-C30. This raises the issue of why it is thought the EQ-5D is inappropriate for capturing change in cancer patients. Our findings suggest it may be due to content validity, as here 12.2% of EQ-5D responses are at full health whereas for a large proportion of these observations the EORTC-8D captures problems on dimensions such as fatigue and physical functioning. This indicates problems with content validity, yet this can only be appropriately determined using qualitative analysis which is beyond the scope of this paper. This raises the issue of how to determine whether EQ-5D is inappropriate. Guidance suggests that it is not simply a question

of ability to detect a change; content validity has an important role in the new US FDA guidance on patient reported outcomes for use in labelling claims (8), and a recent report on economic evaluations in cancer found that EQ-5D did not contain all domains thought important for sensitivity in an outcome measure for cancer patients (7).

The recommendation of one generic measure such as the EQ-5D for use in all economic evaluations is advantageous for comparability, but raises issues of best practice when this measure is unavailable or inappropriate. Recommended alternatives are to use mapped estimates or other preference-based measures. Our analysis suggests that these methods are able to discriminate across severity groups and are responsive, but that the mean change and standard deviation across time periods or severity groups is affected by the alternative method used, and all methods produced lower mean change and standard deviation than the use of EQ-5D directly. Mapped estimates contain error and this will affect the accuracy of the utility estimates. In contrast, EORTC-8D estimates captured problems in quality of life at the upper end of the utility scale that were not captured by EQ-5D, but overall produced higher utility estimates and smaller mean change. The preference-based EORTC-8D performed comparably to the non-preference-based EORTC QLQ-C30 measure it was derived from.

Table 1: Discrimination across severity groups

Measure	Range of mean (s.d.) across groups	Range of ES	ANOVA
Karnofsky performance scale (6 severity groups, N=36 to 1410 per group)			
EQ-5D	0.259 (0.358) to 0.810 (0.179)	0.179 to 0.547	<0.001
EORTC-8D	0.597 (0.117) to 0.852 (0.122)	0.354 to 0.497	<0.001
McKenzie and van der Pol mapped EQ-5D estimates	0.324 (0.225) to 0.791 (0.192)	0.293 to 0.508	<0.001
Kontodimopoulos et al mapped EQ-5D estimate	0.381 (0.206) to 0.879 (0.195)	0.287 to 0.570	<0.001
EORTC QLQ-C30 functioning summary scores	16.0 (24.3) to 88.5 (15.5)	0.021 to 0.657	<0.001
EORTC QLQ-C30 symptom summary scores	70.2 (25.0) to 4.4 (12.4)	-0.012 to -0.476	<0.001 to 0.004
EORTC QLQ-C30 global QOL summary score	41.3 (16.0) to 69.05 (17.78)	0.045 to 0.499	<0.001
Overall quality of life (EORTC QLQ-C30 item: 7 severity groups, N=122 to 1688 per group)			
EQ-5D	-0.025 (0.309) to 0.885 (0.184)	0.388 to 0.816	<0.001
EORTC-8D	0.483 (0.123) to 0.918 (0.086)	0.550 to 0.811	<0.001
McKenzie and van der Pol mapped EQ-5D estimates	0.061 (0.191) to 0.930 (0.135)	0.659 to 1.041	<0.001
Kontodimopoulos et al mapped EQ-5D estimate	0.079 (0.162) to 1.080 (0.126)	0.946 to 1.461	<0.001
EORTC QLQ-C30 functioning summary scores	12.3 (23.2) to 95.0 (12.2)	0.045 to 0.859	<0.001
EORTC QLQ-C30 symptom summary scores	84.3 (20.9) to 1.7 (7.0)	-0.013 to 0.671	<0.001

Note: ES=effect size, ANOVA=analysis of variance.

Table 2: Responsiveness

	All (n=5650)				Screening to cycle 9 (n=255)				Screening to end of treatment (n=370)			
	Min	Max	% at worst state	% at (or above) full health	Mean change (s.d.)	SRM	ES	Paired t-test (P value)	Mean change (s.d.)	SRM	ES	Paired t-test (P value)
EQ-5D	-0.594	1.000	0.02%	12.2%	0.189 (0.337)	0.559	0.539	<0.001	0.100 (0.371)	0.270	0.289	<0.001
EORTC-8D	0.291	1.000	0.01%	3.7%	0.049 (0.145)	0.338	0.329	<0.001	0.021 (0.154)	0.137	0.142	0.009
McKenzie and van der Pol mapped EQ-5D estimates	-0.268	1.058	0%	2.5%	0.137 (0.276)	0.495	0.486	<0.001	0.078 (0.289)	0.271	0.283	<0.001
Kontodimopoulos et al mapped EQ-5D estimate	-0.181	1.186	0%	11.4%	0.154 (0.263)	0.586	0.581	<0.001	0.084 (0.302)	0.278	0.310	<0.001
EORTC QLQ-C30 functioning summary scores	0	100	0.5% to 8.4%	7.4% to 35.0%	0.523 (20.909) to 11.035 (24.871)	0.025 to 0.444	0.023 to 0.431	<0.001 to 0.690	1.216 (22.341) to 5.563 (26.418)	0.041 to 0.211	0.042 to 0.228	<0.001 to 0.100
EORTC QLQ-C30 symptom summary scores	100	0	0.7% to 5.4%	8.1% to 80.3%	1.569 (22.665) to -18.039 (33.321)	0.069 to -0.541	0.105 to -0.540	<0.001 to 0.270	-1.441 (19.269) to 14.054 (34.301)	-0.075 to -0.410	-0.088 to -0.422	<0.001 to 0.426
EORTC QLQ-C30 global QOL summary score	0	100	1.6%	2.6%	14.705 (25.108)	0.586	0.664	<0.001	9.437 (26.532)	0.356	0.413	<0.001

Note: SRM=standardised response mean, ES=effect size.

Table 3: Correlation and agreement

Measures	Pearson correlation coefficient	ICC (mean (95% confidence interval))	ICC p value
EQ-5D and EORTC-8D	0.713	0.481 (0.263 to 0.624)	<0.001
EQ-5D and McKenzie and van der Pol mapped EQ-5D estimates	0.757	0.749 (0.737 to 0.760)	<0.001
EQ-5D and Kontodimopoulos et al mapped EQ-5D estimate	0.749	0.713 (0.626 to 0.775)	<0.001
EORTC-8D and McKenzie and van der Pol mapped EQ-5D estimates	0.933	0.638 (0.072 to 0.832)	<0.001
EORTC-8D and Kontodimopoulos et al mapped EQ-5D estimate	0.875	0.721 (0.646 to 0.775)	<0.001
McKenzie and van der Pol mapped EQ-5D estimates and Kontodimopoulos et al mapped EQ-5D estimate	0.921	0.868 (0.517 to 0.943)	<0.001

Note: ICC=Intraclass correlation coefficient.

Table 4: Correlation by dimension (n=5650)

Spearman rank correlation	EQ-5D dimensions				
EORTC-8D dimensions	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression
Physical functioning	0.599	0.444	0.610	0.428	0.287
Role functioning	0.536	0.482	0.654	0.431	0.314
Pain	0.502	0.437	0.560	0.613	0.332
Emotional functioning	0.260	0.267	0.343	0.303	0.626
Social functioning	0.452	0.422	0.588	0.386	0.334
Fatigue	0.424	0.323	0.496	0.383	0.342
Nausea	0.204	0.194	0.244	0.207	0.216
Constipation and diarrhoea	0.219	0.233	0.268	0.255	0.228

Table 5: EORTC-8D responses when EQ-5D=1 (n=722)

EORTC-8D dimensions	Level 1 %	Level 2 %	Level 3 %	Level 4 %	Level 5 %
Physical functioning	54.29	37.12	7.76	0.83	0
Role functioning	77.01	20.91	1.94	0.14	N/A
Pain	88.09	11.50	0.42	0	N/A
Emotional functioning	85.04	14.27	0.28	0.42	N/A
Social functioning	82.83	15.65	1.39	0.14	N/A
Fatigue	47.23	48.06	4.57	0.14	N/A
Nausea	92.94	6.51	0.14	0.42	N/A
Constipation and diarrhoea	69.39	25.48	3.74	1.39	N/A

Table 6: EQ-5D responses when EORTC-8D=1 (n=206)

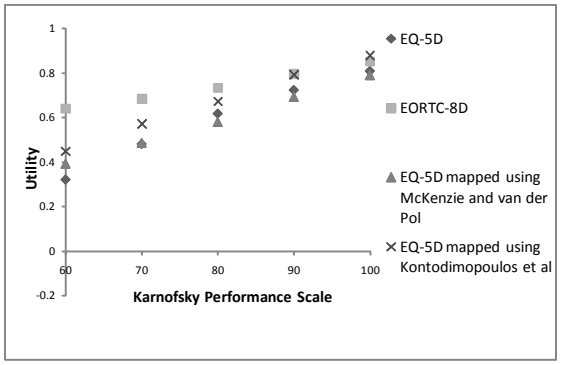
EQ-5D dimensions	Level 1 %	Level 2 %	Level 3 %
Mobility	96.60	3.40	0
Self-care	99.03	0.97	0
Usual activities	99.51	0.49	0
Pain/discomfort	85.92	14.08	0
Anxiety/depression	90.78	8.74	0.49

Table 7: EQ-5D descriptive statistics using alternative preference weights (n=5650)

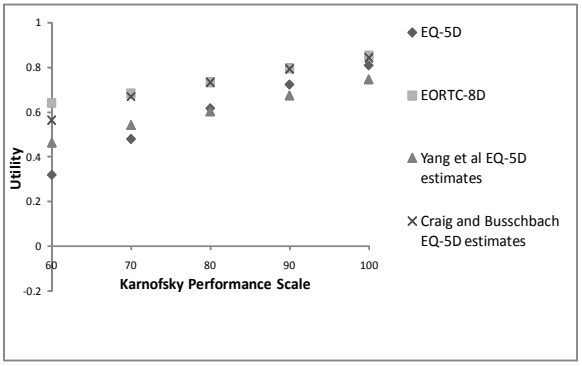
	Craig and Busschbach	Yang et al
Mean (s.d.)	0.741 (0.192)	0.623 (0.185)
Min	-0.298	-0.236
Max	1	1
Screening to cycle 9 (n=255)		
Mean change (s.d.)	0.112 (0.232)	0.098 (0.212)
SRM	0.496	0.489
Effect size	0.509	0.550
Paired t-test (P value)	p<0.001	p<0.001
Screening to end of treatment (n=370)		
Mean change (s.d.)	0.061 (0.268)	0.059 (0.237)
SRM	0.229	0.243
Effect size	0.264	0.299
Paired t-test (P value)	p<0.001	p<0.001

Note: SRM=standardised response mean, ES=effect size, ANOVA=analysis of variance

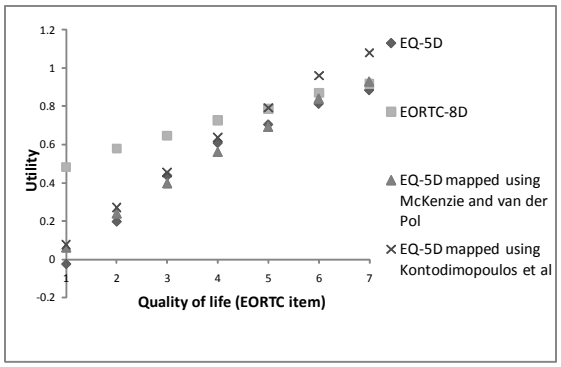
Figure 1: Discrimination across severity groups and mean utility values by period



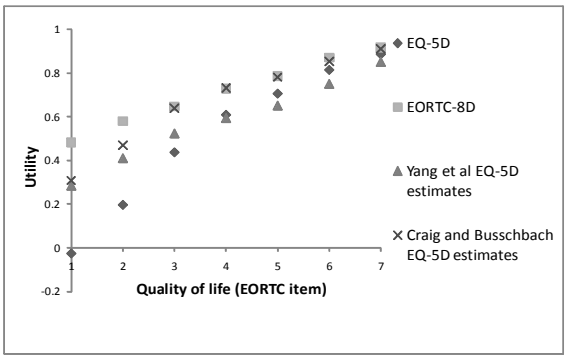
a) Discrimination across severity groups by Karnofsky Performance Scale



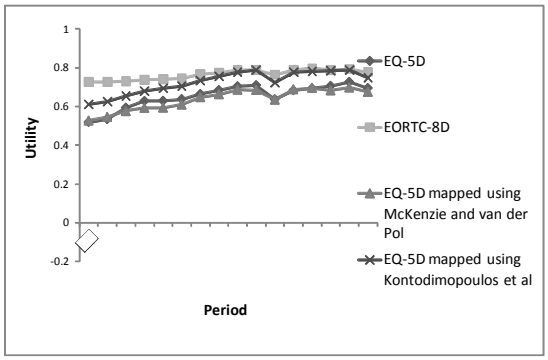
b) Discrimination across severity groups by Karnofsky Performance Scale using alternative EQ-5D weights



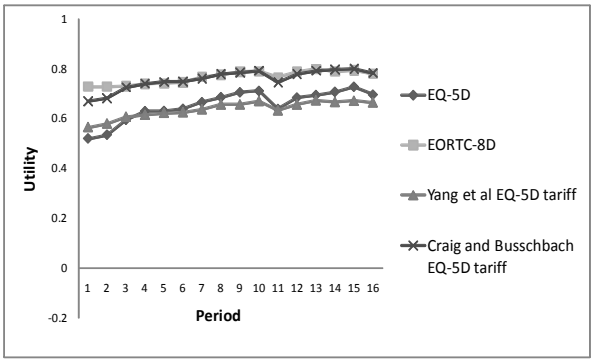
c) Discrimination across severity groups by overall quality of life



d) Discrimination across severity groups by overall quality of life using alternative EQ-5D weights



e) Utility values by period



f) Utility values by period using alternative EQ-5D weights

Note: n<40 for Karnofsky Performance Scale<60 so is not reported for 1a) and 1d).

References

(1) Brooks R. EuroQol: the current state of play. Health Policy 1996 Jul;37(1):53-72.

(2) Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002 Feb;40(2):113-28.

(3) Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 2002 Mar;21(2):271-92.

(4) Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Medical care 2004 Sep;42(9):851-9.

(5) Brazier JE, Ratcliffe J, Solomon JA, Tsuchiya A. Measuring and valuing health for economic evaluation. Oxford: Oxford University Press, 2007.

(6) National Institute of Health and Clinical Excellence (NICE). Guide to the methods of technology appraisal. London: NICE; 2008.

(7) Garau M, Shah K, Towse A, et al. Assessment and appraisal of oncology medicines: does NICE's approach include all relevant elements? What can be learnt from international HTA experiences? Report for the Pharmaceutical Oncology Initiative (POI) 2009.

(8) U.S.Department of Health and Human Services Food and Drug Administration (FDA). Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Maryland: FDA; 2009.

(9) Rowen D, Brazier J, Young T, et al. Deriving a preference-based measure for cancer using the EORTC QLQ-C30. Value in Health 2011;14(5):721-31.

(10) Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997 Nov;35(11):1095-108.

(11) McKenzie L, van der Pol M. Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data. Value in Health 2009 Jan;12(1):167-71.

(12) Barton GR, Sach TH, Jenkinson C, et al. Do estimates of cost-utility based on the EQ-5D differ from those based on the mapping of utility scores? Health and quality of life outcomes 2008;6:51.

(13) Chuang LH, Kind P. Converting the SF-12 into the EQ-5D: an empirical comparison of methodologies. PharmacoEconomics 2009;27(6):491-505.

(14) Brazier JE, Rowen D, Mavranezouli I, et al. Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). Health Technology Assessment 2012;Forthcoming.

(15) Aaronson NK, Ahmedzai S, Bregman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. Journal of the National Cancer Institute 1993;85(5):365-76.

(16) Kontodimopoulos N, Aletras VH, Paliouras D, Niakas D. Mapping the cancer-specific EORTC QLQ-C30 to the preference-based EQ-5D, SF-6D, and 15D instruments. Value in Health 2009;12(8):November-December.

(17) Wu EQM. Mapping FACT-P and EORTC QLQ-C30 to patient health status measured by EQ-5D in metastatic hormone-refractory prostate cancer patients. Value in Health 2007;10(5):408-14.

(18) Pickard AS, Shaw JW, Hsiang-Wen L, et al. A Patient-Based Utility Measure of Health for Clinical Trials of Cancer Therapy Based on the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire. Value in Health 2009;12(6):977-88.

(19) Crott R, Briggs A. Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. Eur J Health Econ 2010 Aug;11(4):427-34.

(20) Versteegh MM, Rowen D, Brazier J, Stolk EA. Mapping onto EQ-5D for patients in poor health. Health & Quality of Life Outcomes 2010;8(141):1-13.

(21) Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-effectiveness in health and medicine. Oxford: Oxford University Press, 1996.

(22) Karnofsky D, Burchenal J. The Clinical Evaluation of Chemotherapeutic Agents in Cancer. In: MacLeod C, ed., Evaluation of Chemotherapeutic Agents. Columbia: Columbia University Press, 1949.

(23) Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press, 1977.

(24) Yang Y, Brazier J, Tsuchiya A. The effect of adding a sleep dimension to the EQ-5D. Health Economics Study Group Meeting, University of East Anglia 2008.

(25) Craig BM, Busschbach J. The episodic random utility model unifies time trade-off and discrete choice approaches in health state valuation. Population Health Metrics 2009;7(3).

(26) Brazier JE, Yang Y, Tsuchiya A, et al. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010 Apr;11(2):215-25.