APPROPRIATE MAPPING FUNCTIONS IN HEALTH RELATED QUALITY OF LIFE: MAPPING FROM FACT-G OR QLQ-C30 TO EQ-5D

Tracey Young[1], Clara Mukuria[1], Donna Rowen[1], John Brazier[1], Louise Longworth[2]

[1] ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA

[2] HERG, Brunel University, Uxbridge, Middlesex UB8 3PH

Contact details: t.a.young@sheffield.ac.uk

Acknowledgements

We would like to thank Simon Pickard, Stewart Peacock and Galina Velikova for providing us with the NCCN, Vancouver Cancer Clinic and VISTA trial datasets (ClinicalTrials.gov number NCT—111319) that were used in this paper. The usual disclaimer applies.

INTRODUCTION

In the UK, the National Institute for Health and Clinical Excellence (NICE) recommend using the EQ-5D to measure preference-based health related quality of life (HRQL) to estimate quality adjusted life years (QALYs) for use in economic evaluations [NICE,2008]. However, it is acknowledged that the EQ-5D is not always an available or appropriate measure and in these situations NICE guidelines recommend using mapping to estimate EQ-5D scores when they are not available. Mapping allows health state utility values to be predicted when no preference-based measure has been included in the study. This approach involves estimating the relationship between a non-preference-based measure and a generic preference-based measure using statistical association and requires a degree of overlap between the descriptive systems of the two measures. It also assumes that the two measures have been administered on the same population.

A review of mapping functions by Brazier et al (2010) showed researchers used a number of models including ordinary least squares regression (OLS), generalised linear models, tobit, censored absolute deviance (CLAD), two part models, and response mapping to predict quality of life. Studies also report a variety of methods to assess model and predictive performance including: predicted mean and standard deviation, median, range of predictions, Akaike Information Criteria (AIC), Bayes Information Criteria (BIC), R^2, pseudo R^2, mean estimates across severity groups, root mean squared error (RMSE), mean square error (MSE) and mean absolute error (MAE).

It has been demonstrated that the EQ-5D is sensitive to changes in HRQL in patients with cancer [For example: Cella *et al,* 2010; Janson *et al,* 2007; Sandblom *et al,* 2004]. Many cancer studies do not include the EQ-5D and are more likely to include one of two cancer specific HRQL questionnaires the European Organization for Research and Treatment Quality Of Life Questionnaire Core 30 (EORTC QLQ-C30) or the Functional Assessment of Cancer Therapy Scale (FACT-G). Only one mapping function has been published mapping from FACT-G to EQ-5D, it fitted OLS and CLAD models at the domain level and showed that scores were poorly predicted away from the mean [Cheung*et al*, 2009]. There are several mapping functions for the EORTC QLQ-C30 in the literature. Potentially the most useful function is that of McKenzie and van der Pol (2009) who used OLS to predict EQ-5D index and ordered probit to predict dimension levels, the ordered probit model did not produce reliable predictions however the OLS gave reasonable EQ-5D estimates. It is possible that other mapping functions such as the tobit model would produce more accurate estimates though these were not explored in this paper. Of the remaining mapping functions Kontodimopoulous*et al* (2009)

use OLS to predict EQ-5D, the model is based on a small sample of patients and they state that their function is unreliable; Wu et al (2007) requires data on FACT-G as well as EORTC to produce mapping estimates, which studies may not routinely collect together; Pickard*et al* (2009) map to patient TTO values rather than EQ-5D index. Crott and Briggs (2010) uses only females as their patient group and therefore the results are not generalisable as evidenced by results from assessment of this mapping function in different datasets [Crott, Versteegh and Uy-de-Groot, 2012] and Versteegh*et al* (2010) do not provide the mapping functions for other users to apply.

Given the lack of robust mapping studies in this area the aim of this paper is to estimate mapping functions from two cancer-specific HRQL measures, EORTC QLQ-C30 and FACT G, to the EQ-5D for use in future studies. Additionally we also wanted to test the applicability of different mapping approaches that have been used in the literature in order to provide recommendations for mapping studies. We tested different modelling techniques that have been recommended in the literature and used recommended criteria to identify the most appropriate mapping functions.

METHODS

Data

Four datasets were used in this analysis, one contained the FACT-G and EQ-5D the remaining three the EORTC QLQ-C30 and EQ-5D, the latter three datasets were combined to produce mapping functions.

FACT-G

The FACT-G is a 27 item cancer specific HRQL measure that has been widely validated [Cella*et* al, 1993]. Each item has 5 options ranging from not at all (score 0) to very much (score 4) and these are summed to obtain a global score as well as four subscale scores: physical well-being, social/family well-being, emotional well-being, functional well-being.

The dataset used to produce mapping functions consisted of 530 US respondents with 13 different types of stage III and IV cancers who completed the EQ-5D and FACT-G [Pickard*et al*, 2007]. 52% (N = 273) respondents were male and the average age of the sample was 59 years.

QLQ-C30

The EORTC-QLQC30 is a 30 item cancer specific HRQL measure that has also been widely validated [Aaronson *et al,* 1993], two items ask about overall quality of life and overall health, the remainder cover 5 functioning scales: physical, role, social, emotional and cognitive and 9 symptoms scales: fatigue, nausea and vomiting, pain, dyspnoea, sleep disturbance, appetite loss, constipation, diarrhoea and financial impact.

The three datasets that were combined in this analysis were: a randomised controlled trial of 572 patients with multiple myeloma (VISTA study) [Griepp *et al*, 2005], 100 patients with breast cancer and 99 patients with lung cancer seen at the Vancouver cancer clinic in Canada (VCC data). This gave a total of 771 cases for the mapping study, the mean age of patients was 68 years and 44% of responders were male.

Models that were considered

Five alternative types of model were fitted to the data: OLS, tobit, two-part models, splining and response mapping, the first four model types are used to map to EQ-5D index and the latter to individual EQ-5D domain scores.

The most commonly used mapping model is OLS. These models are typically able to predict the mean scores but are poor at predicting those in poor health and full health and do not allow for the fact that the EQ-5D is bounded. Therefore, tobit models were fitted to allow for the bounded nature of EQ-5D, thus limiting predictions to within a credible range. An alternative model that can be fitted in an attempt to predict responders in perfect health is the two-part model, this model uses a combination of two different model types to predict different parts of the distribution of the data. Logistic regression was fitted to predict the probabilty of whether responders were in full health (FH) or not and a truncated OLS was applied to predict EQ-5D scores for those not in full health. The results from the two parts of the model were combined to obtain an overall score.

Expected(EQ-5D) = Probability(FH) + (EQ-5Dscore if not FH*(1 – Probility(FH)))

where FH is full health

The final model that was fitted to the EQ-5D index was splining, also known as fractional polynomials. Splining can be used to identify changes (cut points) in the distribution of the data and to model these changes using different mathematical functions. The function mfp in Stata was used to identify possible cut off values (Royston & Sauerbrei, 2007), this function fits all possible polynomial functions to the data and identifies the best fitting model. We applied splining functions to the best fitting OLS/tobit dimension based models to test whether splines offered an improvement over using squared terms.

OLS, tobit, two-part models and splining models are usually reliable at predicting the group EQ-5D mean and median scores, and are able to distinguish between severity levels but are poor at predicting the overall range of EQ-5D scores. An alternative to modelling the EQ-5D index is to use response mapping to fit multinomial logistic regression models to each of the five dimensions of the EQ-5D (Gray *et al,* 2006; Rivero-Arias *et al,* 2010). The estimates from multinomial regressions were used to categorise respondents into levels 1, 2 or 3 of the EQ-5D dimensions. The expected value of each dimension is then calculated to estimate the expected response level and the EQ-5D index was then calculated using the York MV tariff [Dolan*et al*, 1997].

Models were fitted to global HRQL score (FACT-G only), domain scores, squared, square root and interaction terms, item levels and included significant respondent characteristics and severity characteristics. A summary of the models fitted are listed below:

Model 1 EQ-5D index/dimensions = Global FACT-G score

Model 2 EQ-5D index = All FACT-G/QLQ-C30 domain scores

Model 3 EQ-5D index/dimensions = Significant FACT-G/QLQ-C30 domain scores

Model 4 EQ-5D index/dimensions = Significant FACT-G/QLQ-C30 domain scores, squared and square root terms

Model 5 EQ-5D index/dimensions = Significant FACT-G/QLQ-C30 domain scores, squared and square root and interaction terms

Model 6 EQ-5D index/dimensions = Significant FACT-G/QLQ-C30 items

Model 7 EQ-5D index/dimensions = Significant FACT-G/QLQ-C30 items with collapsed item levels if items were unordered

Model 8 EQ-5D index/dimensions = Best of models 1 to 7 with significant respondent/disease characteristics

To avoid over fitting models we use the rule of 10 participants per variable for continuous models and 10 events for the smallest category for response mapping models. Models were fitted using backwards regression and variables were removed from the model if non-significant at $p < 0.1$. When variables are highly correlated (correlation > 0.7) the variable that was most significant and cognitively appears more likely to map to the EQ-5D was selected. Standard errors of regression coefficents were calculated from bootstrap estimates, 5000 bootstrap samples are run for each model.

Model goodness of fit was measured using AIC and BIC and MAE, where the smaller the value the better the model fit. Model performance was also assessed visually by plotting observed and predicted EQ-5D values by EQ-5D health state. For OLS we examined the R^2 and adjusted R^2 to show the models explanatory power, the higher the value the better the model and use the Ramsey Regression Equation Specification Error Test (RESET) to test wether non-linear conbinations of varaibles in the model help explain the variability. For tobit models, logistic regression models and multinomial regression models we use the pseudo R^2 which uses the ratio of model predicted probabilites or the likelihood statistics from the null and full models to obtain a "Pseudo" R^2 value, where the higher the value the better the model. Sigma is reported for tobit and truncated regression models and is the equivalent to root mean squared error in linear regression models. The link test was used to check model specification. The Hosmer-Lemeshow test was used to assess goodness of fit for logistic regression models (First part of two –part models) the test assess whether predicted probabilities agree with observed probabilities and should be non-significant for a model that accuratly predicts observed values.

Model performance and discrimination

Summary statistics including mean and range were examined to asses overall model predictions. We used a severity measure to assess the discriminative performance of the predicted EQ-5D score. For FACT-G, respondents were asked a variation of the Eastern Cooperative Oncology group (ECOG) performance status [Okan et al, 1982] The ECOG has 5 response categories: normal activity without symptoms, some symptoms but do not require bed rest during the waking day, require bed rest for less than 50% of the waking day, require bed rest for over 50% of the waking day and unable to get out of bed. No patients were in the most severe level (unable to get out of bed) and few patients (n = 21 (4%)) required bed rest for more than 50% of the waking day therefore these two categories are merged with do not require bed rest less than 50% of the waking day. The ECOG responses are

included in mapping models as a measure of disease severity and to test the predictive ability of the mapping models across different severity groups. There was no common severity measure in the EORTC QLQ-C30 datasets and item 29 – reported health status – was used instead. Response options ranged from poor to excellent (1 to 7). Discriminative ability across severity groups using these measures was tested using ANOVA. MAE were reported for each subgroup.

Model validation

Models were validated internally using the bootstrapping techniques reported by Steyerberg *et al* (2000) to assess all models and shrinkage coefficients are reported in order to counter over optimism of estimates. A shrinkage coefficient of less than 1 (typical value expected for a shrinkage coefficient) reflects an "over fitting" of the data.

Model selection

When producing a mapping model the factors that are important in selecting a model are accuracy of the predicted mean and standard error, the range of predictions, MAE, shrinkage and the reproducibility of the model across different severity states. Mapping and model fitting literature does not suggest a single criteria for use in selecting the best fitting model and the criteria that we might focus on when selecting a model may depend on what we want the mapping function to acheive. Therefore when selecting models all criteria were given equal wieghting and models were ranked based on these factors and the mean rank per model was estimated. The model with the best ranking was then selected and these were then compared across the different estimation methods (OLS, tobit, two-part, splining, response mapping).

RESULTS

Descriptive statistics

The FACT-G dataset did not include responders with very poor HRQL, the mean EQ-5D index score was 0.721 (SD = 0.22) with median of 0.735, score ranged from -0.135 to 1 and 18% of responders were in full health and 0.9% scored worse than death. No responder had extreme problems for mobility and few responders indicated extreme problems for self-care (0.4%), usual activities (6%), pain/discomfort (3%) or anxiety depression (2%). The distribution of the EQ-5D index for the FACT-G dataset is shown in Figure 1, there is a large gap (0.117) between those in full health and the next allowable value according to the EQ-5D tariff score and there are at least two more components in

the distribution with peaks around 0.7 and 0.2. Average FACT-G scores were 20, 23, 18 and 18 for physical, social, emotional and functional dimensions respectively. The average global score was 78 and ranges from 33 to 108, thus like the EQ-5D, it did not cover the worse end of the FACT-G scale. The relationship between global FACT-G score and EQ-5D was moderate (Spearman's correlation = 0.575). The EQ-5D also correlated moderately with the physical and functional domains of the FACT-G.

The characteristics and a summary of EQ-5D scores and EORTC QLQ-C30 responses are presented in Table 1 for the combined sample and for each dataset. Mean age and proportion of males varied by dataset as did mean EQ-5D scores which were lowest for the multiple myeloma dataset and higher for the breast and lung cancer datasets. Only the multiple myeloma dataset covered the entire range of the EQ-5D, and had lower ceiling effects than the other datasets, with 8% of responses at full health on EQ-5D in comparison to 24% and 17% for the breast and lung cancer datasets respectively. Figure 1 presents the histograms for each dataset and the combined dataset showing that the distributions differ by dataset but without further information we cannot conclude whether this is differences in the severity of the patients in each dataset or differences in the pattern of EQ-5D by condition. Like the FACT-G, there is evidence of a large gap between full health and the next allowable EQ-5D score as well as evidence of peaks. The scores for the EORTC QLQ-C30 scales most noticeably varied across the three datasets for physical functioning, role functioning, pain, dyspnoea, constipation and global quality of life. Assessment of the correlations between the independent variables indicated that the highest correlations were between role functioning, physical functioning and fatigue variables.

Mapping results

Selecting models

To illustrate the model selection process we present summary results for the eight models fitted to FACT-G using OLS (Table 2). Model 2 to 5 present OLS models using FACTG domain scores. Physical, Emotional and Functional are all significant predictors of EQ-5D and an increase in these FACT-G domains results in an increase in EQ-5D scores, and this was consistant with findings of Cheung (2009). Model 5 included an interation term between physical and functional scores, the only high correlation found between domains. Model 6 and 7 were item level models including only significant items, these were lack of energy, trouble meeting the need of family and pain from the physical domain, sad and loosing hope from the emotional domain and able to work from the functional

domain. Level 0 (very much) of I feel sad had less than 20 observations, these responses were merged with level 1 and model 6 was then refitted, Collapsing item levels did not improve the overall model fit. Model 8 considered adding patient and disease characteristics, however no characteristic was found to be a significant predictor of EQ-5D score, therefore the results for this model are not shown.

Table 2 presents ranking across each of the performance statistics for the above OLS models as well as the mean rank for each model. Item level models consistantly performed better than the domain and global score models though were most likely to overfit and were poor at predicting scores away from the overall mean for the dataset. All models predicted the overall mean EQ-5D score for the dataset, underestimated those in near full or full health and overestimated those in poorer health states. No model predicted a score below 0.155 (observed scores ranged from -0.135 to 1). All models were able to discriminate between different levels of health, as measured by ECOG. MAE was large for those in poor health, which is unexpected given the range of model predictions. Giving all performance statistics equal weighting suggested that model 6 was the best performing OLS model for estimating EQ-5D scores. This process was then repeated for tobit, two-part, splining and response mapping (Results not shown but available from the authors on request) for the FACT-G and the EORTC QLQ-C30. The best ranked functions for each model (OLS, tobit etc.) were then compared using the same approach for the two measures.

Best fitting models – FACT-G

Table 4 and Figure 2 summarise the best models fitted OLS, tobit, two part modelling, splining and response mapping models for the FACT-G. OLS gave the best estimates of the overall mean and mean by severity group and had one of the two best ranges of predicted scores (the two-part model covered the widest range). OLS was the poorest at predicting the median and had the lowest shrinkage factor suggesting it would be the most likely to over-predict results in other studies applying the mapping algorithm. The response mapping model gave reasonable estimate of the mean and median but the poorest MAE across severity groups. A mean ranking of models across the different model performance statistics showed OLS to give the best predictions (mean rank = 2.17), followed by tobit (mean = 2.5), with two-part models (mean = 3.75) and response mapping (mean = 3.58) giving the poorest predictions. All models failed to predict anyone in perfect health, under predicting at the top of the EQ-5D scale and under predicting at the bottom end of the scale.

However, the under prediction at the lower end of the scale is perhaps unsurprising given that few responders in the FACT-G dataset reported severe problems with quality of life.

Best fitting models – EORTC QLQ-C30

Table 5 and Figure 2 presents the predicted EQ-5D scores for the best fitting models for each estimation technique alongside model performance statistics for the EORTC QLQ-C30. As with FACT-G the item-level models gave the best model predictions for OLS and Tobit models (Model 8 items and respondent characteristics), these models were best at predicting the overall mean EQ-5D score. Unlike FACT-G item-level models with respondent characteristics gave the best model performance for two-part models, these models were better at predicting the median. The splining model had the least deviation from the shrinkage coefficient of 1 (Model 3).The best performing response mapping model included all domains with age and gender for some of the dimensions, this model had the lowest mean absolute errors on average. None of the models predicted the full range of observed EQ-5D scores, with no predictions at the best or worst EQ-5D scores. The mean ranking indicated that the response mapping was the best performing model (mean rank = 2.5), with OLS also performing well (mean = 2.6) and splining giving the poorest overall performance (mean = 3.6).

DISCUSSION

The best performing model for the EORTC QLQ-C30 and the FACT-G differed with response mapping and OLS performing best for EORTC QLQC30 and OLS and Tobit performing best for FACT-G with response mapping producing poor predictions for the FACT-G. These differences are unlikely to represent general findings and may be due to the nature of the FACT-G dataset which had a limited number of responders in poor health. This was surprising as the responders all had stage III or IV cancer and covered a range of different cancers but might be due to the FACT-G study asking respondents to fill in a large number of questionnaires, in addition to the ones reported here responders completed a EQ-5D -5L, disease specific FACT-G modules, ECOG performance measure, cancer and treatment distress scale (CTXD) the renal cell carcinoma (RCC) symptom index and the symptom index for anxiety and depression (SCLDA), making the task quite lengthy and thus potentially biasing the sample to more healthy respondents. This means that the FACT-G mapping results are not necessarily generalizable to other studies, unless they also consist of a fairly healthy population.

Furthermore, the FACT-G sample size may have added to the poor performance of response mapping. With a larger sample e.g. 2000 respondents you would obtain more accurate predictions of those in level 3, as although the percentage of observations for this level might remain the same (e.g. 3%) the number of observations from which estimates could be made would increase giving more reliable estimates. Further work is needed on sample size recommendations for the more complex models such as response mapping.

Other studies have fitted CLAD and GLM models as mapping functions. Like the tobit model the CLAD model also deals with the censored nature of the data and produces consistent estimates in the presence of heteroscedasticity and non-normality, but is a median-based model rather than a mean-based model and so is not suitable for economic evaluation (Sullivan and Ghushchyan, 2006; Powell, 1984). Therefore this model was not fitted here. Generalised linear models were not fitted either as initial GLM model fitting gave similar results to OLS.

It is evident from mapping studies in the literature that studies report different model fit and model selection criteria, some focusing on model goodness of fit, others on the predictive ability of the model. Models should be selected on their predictive ability, however within this there are still a number of criteria from which a model can be selected and different choices can result in alternative models being selected. In this paper we have given equal weighting to all model fitting criteria. The statistical literature suggests looking at model performance criteria such as AIC and BIC (Zucchini, 2000), whereas in mapping criteria such as distribution of predictions across severity and MAE are felt to be more important (Longworth & Rowen, 2011). We would be interested in HESG audiences' thoughts on this.

References

Aaronson, N.K., Ahmedzai, S., Bregman, B., Bullinger, M., Cull, A., Duez, N., Filiberti, A., Flechtner, H., Fleishman, S., de Haes, J.C.J.M., Kaasa, S., Klee, M., Osoba, D., Razavi, D., Rofe, P.B., Schraub, S., Sneeuw, K., Sullivan, M., & Takeda.F 1993. 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, 85, (5) 365-376

Brazier JE, Yang Y, Tsuchiya A, Rowen DL (2010) A review of studies mapping (or cross walking) non-preference-based measures of health to generic preference-based measures. Eur J Health Econ; 11: 215-225

Cella D, Michaelson MD, Bushmakin AG, Cappelleri JC, Charbonneau C, Kim STm Li JZ, Motzer RJ (2010) Health-related quality of life in patients with metastatic renal cell carcinoma treated with sunitinib *vs* interferon-α in a phase III trial: final results and geographical analysis. British Journal of Cancer; 102(4): 658-664

Cella DF, Tulsky DS, Gray G (1993) The functional assessment of cancer therapy scale: development and validation of the general measure. Journal of Clinical Oncology; 11: 570-579

Cheung YB, Thumboo J, Gao F, Ng GY, Pang G, Koo WH, *et al (*2009) Mapping the English and Chinese Versions of the Functional Assessment of Cancer Therapy–General to the EQ-5D Utility Index. Value in Health; 12 (2) 371-376

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.

Crott, R., Versteegh, M., & Uyl-de-Groot, C. (2012). An assessment of the external validity of mapping QLQ-C30 to EQ-5D preferences. *Quality of Life Research*, 1-10.

Dolan P (1997) Modelling valuations for EuroQol health states. Medical Care; 35 (11): 1095-1108

Gray AM, Rivero-Arias O, Clarke PM (2006) Estimating the association between SF-12 response and EQ-5D utility values by response mapping. Medical Decision Making; 26 (18): 18-29

Greipp, P.R., Miguel, J.S., Durie, B.G.M., Crowley, J.J., Barlogie, B., Blade, J., Boccadoro, M., Child, J.A., Avet-Loiseau, H., Kyle, R.A., Lahuerta, J.J., Ludwig, H., Morgan, G., Powels, R., Shimuzu, K., Shustic, C., Sonneveld, P., Tosi, P., Turesson, I., & Westin, J. 2005. International Staging System for Multiple Myeloma. Journal of Clinical Oncology, 23, (15) 3412-3420

Janson M, Lindholm E, Anderberg B, HaglindE (2007) Randomised trial of health related quality of life after open and laparoscopic surgery for colon cancer. Surg Endosc; 21: 747-753

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.

Longworth, L. & Rowen, D. 2011. The use of mapping methods to estimate health state utility values. NICE DSU Technical Support Document 10

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.

National Institute for Health and Clinical Excelence (2009) Guide to the methods of technology appraisal.

Oken, M.M., Creech, R.H., Tormey, D.C., Horton, J., Davis, T.E., McFadden, E.T., Carbone, P.P.: Toxicity And Response Criteria Of The Eastern Cooperative Oncology Group. Am J Clin Oncol 5:649-655, 1982.

Pickard AS, De Leon MC, Kohlmann T, Cella D, Rosenbloom S (2007) Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. Medical Care; 45(3): 259-263

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.

Powell, J.L. 1984. Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics,* 25: 303-325.

Rivero-Arias O, Ouellet M, Gray AM, Wolstenholme J, Rothwell PM, Luengo-Fernandez R (2010) Mapping the modified rankin scale (mRS) measurement into the generic EuroQol (EQ-5D) health outcome. Medical Decision Making; 341-354

Royston P & Sauerbrei W (2007) Multivariable modelling with cubic regression splines: A principled approach. The Stata Journal; 7(1): 45-70

Sandblom G, Carlson P, Sennfalt K, Varenhorst E (2004) A population-based study of pain and quality of life during the year before death in men with prostate cancer. Br J Cancer;90(6):1163-8.

Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small datasets. Statistics in Medicine; 19: 1059-1079

Sullivan PW, Ghushchyan V (2006). Preference-Based EQ-5D index scores for chronic conditions in the United States. Med Decis Making. 2006 Jul-Aug;26(4):410-20.

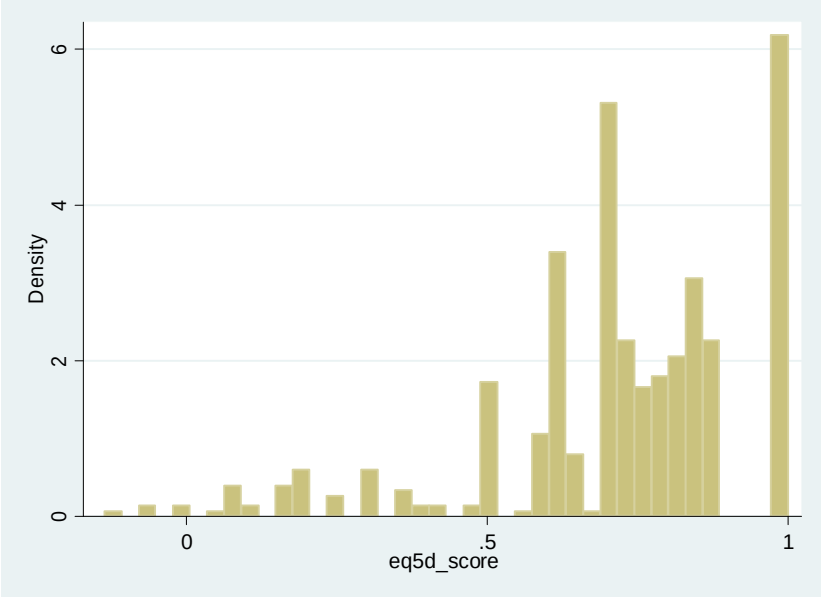
Versteegh, M.M., Rowen, D., Brazier, J., & Stolk, E.A. 2010. Mapping onto EQ-5D for patients in poor health. Health & Quality of Life Outcomes, 8, (141) 1-13

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.

Zucchini W (2000) An introduction to model selection. Journal of mathematical psycology 44: 41-6

Figure 1: EQ-5D distribution for a) FACT-G dataset and b) QLQ-C30 datasets

a) FACT-G dataset



b) QLQ-C30 datasets

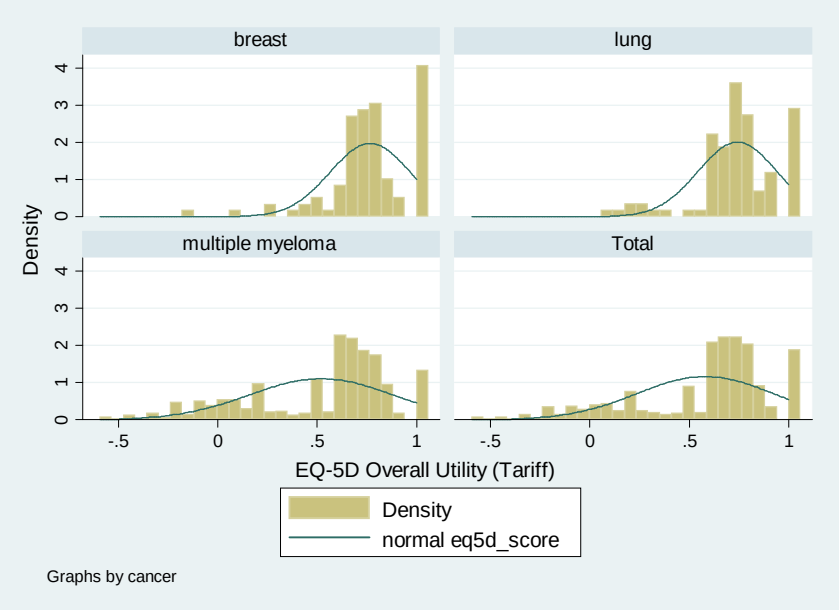
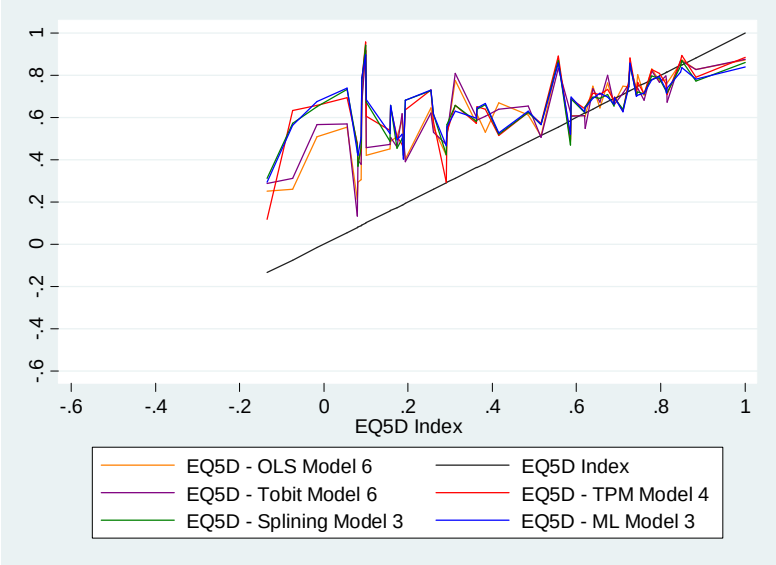


Figure 2: Observed and predicted EQ-5D scores for best performing models for a) FACT-G dataset and b) QLQ-C30 datasets

a) FACT-G dataset



b) QLQ-C30 datasets

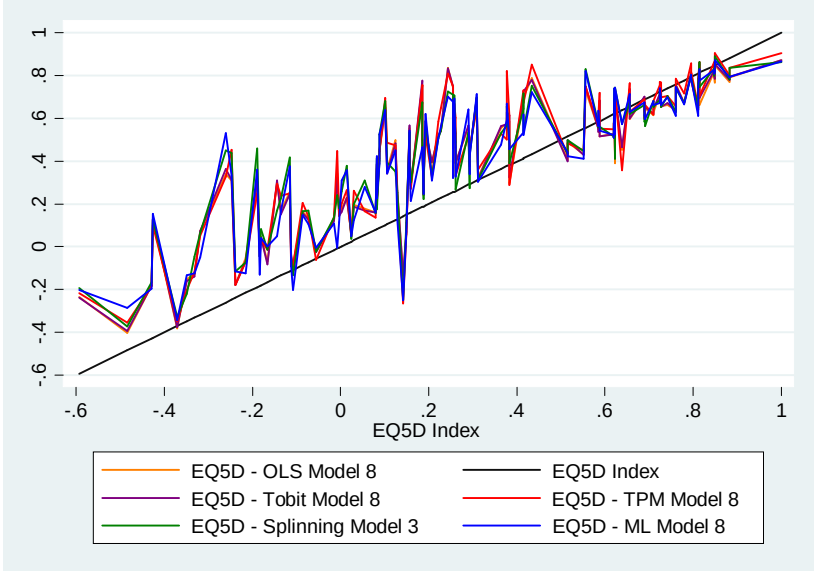


Table 1 Characteristics of the EORTC QLQC30 patient samples

	All (*n*=771)		*Breast* (*n=100*)		*Lung* (*n=99*)		*Multiple Myeloma* (*n=572*)	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Mean age (s.d.)	68.31	9.56	53.9	10.94	62.73	10.5	71.79	5.449
Male (%)	44.1		0		48.0		50.0	
EQ-5D								
EQ-5D utility score	0.579	0.342	0.765	0.202	0.742	0.199	0.519	0.360
EQ-5D = 1 (%)	10.7		24.0		17.1		7.9	
Range of EQ-5D	-0.594 to 1		-0.144 to 1		0.088 to 1		-0.594 to 1	
EORTC QLQ-C30								
Physical functioning	64.81	25.59	78.27	19.87	69.76	19.62	61.6	26.5
Role functioning	59.14	33.18	72.67	27.68	67.51	26.98	55.33	34.17
Emotional functioning	69.71	24.91	73	22.69	76.26	21.47	68.01	25.61
Cognitive functioning	76.05	22.74	76.83	22.83	77.27	20.54	75.7	23.11
Social functioning	69.13	29.82	72	26.15	73.74	23.82	67.83	31.25
Fatigue‡	45.42	26.16	39.11	20.86	42.87	23.11	46.97	27.3
Nausea‡	9.014	17.87	11	19.85	10.27	16.79	8.45	17.69
Pain‡	40.4	32.99	23	24.25	22.56	23.49	46.53	33.54
Dyspnea‡	24.73	28.97	16.67	22.47	36.7	30.67	24.07	29.09
Sleep disturbance‡	32.68	32.6	34	31.06	30.98	28.27	32.75	33.59
Appetite loss‡	27.37	32.53	19.67	28.46	28.62	32.3	28.5	33.1
Constipation‡	23.13	30.71	11.67	23.39	22.9	29.98	25.17	31.56
Diarrhoea‡	9.511	19.86	15.67	26.99	11.45	20.29	8.1	18.04
Financial impact‡	19.76	28.78	23.67	30.45	22.9	28.83	18.53	28.42
Global quality of life	52.76	23.18	67.92	18.17	62.12	21.04	48.48	22.75

EORTC QLQ-C30 dimension score range 0 -100, higher scores indicate better functioning and quality of life: ‡ higher scores for symptom scales indicate worse symptoms

Table 2: Summary of observed and predicted values per model – FACT-G dataset

	Observed values	OLS 1	OLS 2	OLS 3	OLS 4	OLS 5	OLS 6	OLS 7
		Total score	Domain scores	OLS significant domains	OLS significant domains and squared terms	OLS significant domains, squared and interaction terms	OLS item levels – significant levels only	OLS item levels – significant levels only, collapse unordered items
Mean (SD)	0.721 (0.223)	0.721 (0.128)	0.721 (0.138)	0.721 (0.138)	0.721 (0.144)	0.721 (0.146)	0.721 (0.163)	0.721 (0.161)
Median	0.735	0.730	0.735	0.735	0.738	0.744	0.755	0.750
Range	-0.135 to 1	0.319 to 0.975	0.357 to 0.971	0.357 to 0.972	0.198 to 0.981	0.161 to 0.946	0.115 to 0.962	0.169 to 0.961
R^2		0.331	0.383	0.383	0.417	0.432	0.535	0.524
Adjusted R^2		0.330	0.378	0.379	0.413	0.425	0.513	0.507
AIC		-298.40	-335.20	-337.12	-365.34	-374.98	-445.43	-443.38
BIC		-289.86	-313.84	-320.11	-343.97	-345.07	-338.60	-357.92
Ramsey RESET		$F_{3, 525} = 3.19$ ($p = 0.024$)	$F_{3,522} = 0.83$ ($p = 0.477$)	$F_{3,525}=0.84$ ($p = 0.471$)	$F_{3,524} = 2.96$ ($p = 0.032$)	$F_{3,521} = 2.06$ ($p = 0.104$)	$F_{3,502} = 0.72$, $p =0.539$	$F_{3,507} = 1.17$, $p = 0.320$
MAE		0.129	0.126	0.126	0.124	0.122	0.111	0.112
Shrinkage		1.005	0.992	0.996	0.995	0.991	0.850	0.909

		Observed values	OLS 1		OLS 2		OLS 3		OLS 4		OLS 5		OLS 6		OLS 7	
	n	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG																
Normal, no sympt	122	0.8645	0.8156	0.1113	.8339	0.0958	0.8339	0.0958	0.8429	0.0966	0.8404	0.0973	0.8464	0.0868	0.8464	0.0870
Some symptoms do	256	0.7219	0.7280	0.1220	.7325	0.1237	0.7325	0.1236	0.7263	0.1227	0.7281	0.1214	0.7318	0.1080	0.7319	0.1087
Require some bed	152	0.6055	0.6344	0.1568	.6121	0.1540	0.6121	0.1540	0.6154	0.1465	0.6143	0.1451	0.6033	0.1353	0.6032	0.1343
ANOVA		$F_{2,527} = 55$, $p < 0.001$	$F_{2,527} = 92$, $p < 0.001$		$F_{2,527} = 134$, $p < 0.001$		$F_{2,527} = 135$, $p < 0.001$		$F_{2,527} = 125$, $p < 0.001$		$F_{2,527} = 117$, $p < 0.001$		$F_{2,527} = 107$, $p < 0.001$		$F_{2,527} = 108$, $p < 0.001$	

Table 3: EORTC QLQ C30 Mean observed and predicted EQ-5D values per model and summary model performance – OLS models

	Observed values	OLS 1	OLS 2	OLS 3	OLS 4	OLS 5	OLS 6	OLS 7
		Total score	Domain scores	OLS significant domains	OLS significant domains and squared terms	OLS significant domains, squared and interaction terms	OLS item levels – significant levels only	OLS item levels – significant levels only, collapse unordered items
Mean (SD)	0.721 (0.223)	1 (7)	1 (5)	1 (5)	1 (4)	1 (3)	1 (1)	1 (2)
Median	0.735	4	1	1	3	5	7	6
Range	-0.135 to 1	5	7	6	4	3	1	2
R^2		7	5	5	4	3	1	2
Adjusted R^2		7	6	5	4	3	1	2
AIC		7	6	5	4	3	1	2
BIC		7	6	5	3	2	4	1
Ramsey RESET		6	2	3	7	5	1	4
MAE		7	5	5	4	3	1	2
Shrinkage		2	4	1	2	5	7	6

		Observed values	OLS 1		OLS 2		OLS 3		OLS 4		OLS 5		OLS 6		OLS 7	
	n	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG																
Normal, no sympt	122	0.8645	7	7	5	3	5	3	3	5	4	6	1	1	1	2
Some symptoms do	256	0.7219	2	4	6	7	6	6	1	5	3	3	4	1	5	2
Require some bed	152	0.6055	7	7	3	5	3	5	6	4	5	3	1	2	2	1
Mean ranking			5.53		4.53		4.12		3.76		3.53		2.12		2.53	

Table 4: FACT-G Mean observed and predicted EQ-5D values per model and summary model performance – best fitting model across modelling techniques

		Observed values	OLS Model 6		Tobit Model 6		TPM Model 4		Splining Model 3		ML Model 3	
			Significant item levels		Significant item levels		Significant domain scores, squared and square root terms		Significant domain scores		Significant domain scores	
Mean (SD)		0.721 (0.223)	0.721 (0.163)		0.723 (0.161)		0.739 (0.154)		0.723 (0.144)		0.720 (0.133)	
Median		0.735	0.755		0.738		0.753		0.736		0.737	
Range		-0.135 to 1	0.115 to 0.962		0.132 to 0.957		0.119 to 0.993		0.312 to 0.974		0.268 to 0.934	
MAE			0.111		0.181		0.120		0.198		0.125	
Shrinkage			0.850		0.962		0.944		0.982		1.019	
	N	Mean	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
ECOG												
Normal, no symptoms	122	0.8645	0.8464	0.0878	0.8498	0.0878	0.8302	0.0896	0.8460	0.097	0.7933	0.1009
Some symptoms	256	0.7219	0.7318	0.1080	0.7320	0.1108	0.7359	0.1211	0.7277	0.121	0.7201	0.1219
Require some bed rest	152	0.6055	0.6033	0.1353	0.6074	0.1365	0.6713	0.1410	0.6152	0.148	0.6601	0.1485
ANOVA		$F_{2,527} = 55$, $p<0.001$	$F_{2,527} = 107$, $p < 0.001$		$F_{2,527} = 109$, $p < 0.001$		$F_{2,527}=122$, $p<0.001$		$F_{6,527} = 130$, $p < 0.001$		$F_{2,527} = 120$, $p<0.001$	

Table 5 EORTC QLQ C30 Mean observed and predicted EQ-5D values per model and summary model performance – best fitting model across modelling techniques

	N	Observed values	OLS 8 Significant item levels + age		TBT 8 Significant item levels + age		TPM 8 Significant item levels + age (P1)		SPL 3 Significant domains		MLM 8 all domains +age/gender	
Mean (SD)	771	0.5793 (0.3423)	0.5793 (0.2866)		0.5792 (0.2891)		0.6066 (0.2997)		0.5793 (0.2833)		0.5726 (0.2914)	
Median		0.6910	0.6502		0.6517		0.6892		0.6457		0.6569	
Range		-0.594 1	-0.4046 to 0.9714		-0.3937 to 0.9463		-0.3936 to 0.9898		-0.3718 to 0.9438		-0.3376 to 0.9416	
MAE			0.139		0.139		0.140		0.143		0.134	
Shrinkage			1.042		1.020		0.940		0.997		1.179	
Health status (Eortc 29)			Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE	Mean	MAE
(very poor)1	42	-0.0057	0.0642	0.205	0.0638	0.203	0.0649	0.195	0.0660	0.245	0.0473	0.181
2	53	0.1763	0.2470	0.179	0.2433	0.177	0.2670	0.185	0.3345	0.236	0.2262	0.159
3	144	0.4286	0.4629	0.182	0.4602	0.183	0.4808	0.184	0.5166	0.142	0.4515	0.182
4	226	0.6220	0.5823	0.138	0.5816	0.139	0.6091	0.141	0.5694	0.143	0.5827	0.139
5	186	0.7180	0.7176	0.109	0.7205	0.109	0.7566	0.107	0.7353	0.084	0.7094	0.097
6	94	0.8321	0.8181	0.098	0.8195	0.099	0.8511	0.104	0.8151	0.072	0.8137	0.100
(excellent)7	26	0.9029	0.8546	0.080	0.8546	0.081	0.8925	0.060	0.8660	0.134	0.8596	0.075
ANOVA		F_6=97, p=0.000	F_6=114, p=0.000		F_6=113, p=0.000		F_6=114, p=0.000		F_6=117, p=0.000		F_6=116, p=0.000	