

Estimating the association between SF-36 responses and EQ-5D utility values by direct mapping.

Alastair Gray¹, Philip Clarke^{1,2}, Oliver Rivero.¹

1. Health Economics Research Centre , University of Oxford, UK.
2. Diabetes Trials Unit, University of Oxford, UK.

CES - HESG Workshop 2004

1st Franco-British Meeting in Health Economics
Centre de Conférences Internationales
Paris - 14-16 January 2004

Introduction

The ability to make reliable translations or mappings from generic or disease-specific health status measures into health state utilities is of interest to health economists, given the increasing importance of the quality adjusted life year as the standard metric of outcome in economic evaluation. As a result, a number of attempts have been made to estimate the utility values associated with responses to generic instruments such as the SF-36.

To date, five studies have been published that report algorithms designed to generate utility values from the SF-36 and/or the SF-12. Lundberg and colleagues sent a postal questionnaire to 8,000 adults asking them to complete the SF-12, a rating scale question, and a time-trade-off question, and then used age, gender and the individual items of the SF-12 as explanatory variables in a linear regression analysis of health-state utilities indicated by the time-trade-off question.(Lundberg et al. 1999) Fryback adopted a different approach, in which the SF-36 and Quality of Well-being index (QWB) instruments were both administered by interview to 1,430 people in the Beaver Dam Health Outcomes Study, and domain scales of the SF-36, their squares, and all pairwise cross-products were then used as candidate variables in stepwise and best-subsets regressions to predict QWB scores.(Fryback et al. 1997) Shmueli used face to face interviews with a sample of 2,030 adults who rated their own health using the SF-36 and a visual analogue scale, and then used linear and nonlinear regression to estimate the association between domains of the SF-36 and the VAS score.(Shmueli 1999)

Another approach was adopted by Brazier and colleagues (1998), who restructured the SF-36 into the SF-6D health state classification, drew a sample of 59 multidimensional health states from the 9,000 possible states defined by this classification, valued them using visual analogue scale ratings and standard gamble questions on a convenience sample of 165 people, estimated the association between health states and valuations using OLS regression, and finally converted the model results into an algorithm potentially capable of providing utility values from SF-36 responses.(Brazier et al. 1998) More recently, Brazier and colleagues reported further results using a larger sample: in this study, the SF-36 was again reclassified into a six-dimensional health state classification called the SF-6D, a sample of 249 states was drawn from the 18,000 potential states defined by the reclassification, a representative population of 611 people was used to derive valuations of these health states using standard gamble, and the association between health states and valuations was then modelled, generating results that offered a methods for obtaining utility estimates from responses to the SF-36.(Brazier, Roberts, & Deverill 2002)

The studies listed above differ widely in their methods and in their objectives. It is not surprising, therefore, that direct comparisons between them suggest they may also provide different estimates of quality of life or utility for the same set of items responses to the SF-36. For example, Hollingworth and colleagues applied algorithms from five of the above studies to a cohort of patients with low back pain, and found differences in the mean utility value generated by each algorithm, in the ability to discriminate between groups of patients with known differences in disease severity, and in the effect size they are capable of detecting.(Hollingworth et al. 2002) Similar findings were obtained from a comparison of these algorithms in a sample of patients with asthma.(Lee, Hollingworth, & Sullivan 2003). Lorgelly has also found substantial variation in the utility scores of different algorithms when applied to patients with knee pain.(Lorgelly PK 2001)

Of particular interest is the likely association between the utility values generated by these algorithms, and the utility value that would have been obtained from the EQ-5D. The EQ-5D is a multi-attribute instrument for measuring preferences associated with an individual's health state.(EuroQol Group 1990) Using a descriptive system covering five dimensions (mobility, self-care, usual activity, pain/discomfort, anxiety and depression) each of which has three levels (no problem, some problem, extreme problems), a utility value or "tariff" can be assigned to each of the 243 possible health states indicated by a respondent, in the UK by using the results of a survey of the general British population that involved valuing key states using the time-trade method and then using regression methods to impute values for all states.(Dolan et al. 1996) While gaining an understanding of the degree of agreement between utility values obtained from algorithms applied to the SF-36 and/or SF-12 and utility values obtained from the EQ-5D is important, the EQ-5D utility values should not be regarded a "gold standard". Indeed the developers of an algorithm for the SF-36 note that while the instrument can be used as an alternative to the EQ-5D, it may produce different utilities for similar health states due to the larger number of health states it accommodates, and other differences such as the use of standard gamble rather than time-trade off used to value predefined health states that form the basis of the estimates.(Brazier, Roberts, & Deverill 2002)) However, the availability of alternative quality of life instruments producing different estimates of utility is potentially

troubling for applied health economists, as the acceptance of a new therapy or technology may rely on the choice of instrument used. In these circumstances it would seem desirable to be able to obtain a consistent set of utility values from both instruments in order to test the sensitivity of any results to the choice of measure. Further, a translation or mapping algorithm would facilitate the pooling of utility values when different instruments have been used to assess quality of life of different patients (e.g. it would allow the outcomes of trials that have only used either the EQ-5D, or the SF-36 to be combined).

Here we address this issue in two ways. First, using a large population survey in which the SF-36 and the EQ-5D were administered, we calculate utility values using the Brazier (Brazier, Roberts, & Deverill 2002) and Lundberg (Lundberg, Johannesson, Isacson, & Borgquist 1999) algorithms, and compare the results with the actual EQ-5D results. Second, we report the results of an analysis in which we estimate directly using two methods the association between SF-36 responses and EQ-5D responses from the same individuals; we then derive an algorithm that uses SF-36 data to predict responses to EQ-5D questions as well as the utility values that would have been obtained had the EQ-5D been administered.

Methods

Individuals who had completed the SF-36 and EQ-5D in the Health Survey for England 1996, (Health Survey 1998) were included in the study. The SF-36 individual items and domains were scored conventionally: for each of 8 domains the item scores were coded, summed and transformed onto a scale from 0 (worst possible health state measured by the questionnaire) to 100 (best possible health state). Table 1 reports the 8 domains into which individual items are grouped, and the interpretation of high and low scores in each domain. (The guide to the interpretation of very high or very low scores on the SF-36 is adapted from the definitions provided by Ware and Gandek. (Ware & Gandek 1998)) EQ-5D tariff values were calculated using the results of the British MVA survey. (Dolan, Gudex, Kind, & Williams 1996) The Brazier algorithm was based on the most recent published results, (Brazier, Roberts, & Deverill 2002) and the Lundberg algorithm used their reduced form model with time trade-off results as the dependent variable. (Lundberg, Johannesson, Isacson, & Borgquist 1999)

Cases were included if the respondent was aged 18 or over and there were no missing items in their SF-36 and EQ-5D responses or in other variables potentially of interest (age, sex, smoking status, socio-economic group, presence of long-standing illness). The analysis was undertaken in two stages: mapping to the EQ-5D ordinal responses, and mapping to the EQ-5D tariff.

Mapping to ordinal EQ-5D responses

In the first part of the analysis, regression analysis is used to explore the association between responses to the SF-36 (again using either individual items or domain summary scores) and responses to each EQ-5D question: that is, whether the respondent indicated level 1, 2 or 3 on the EQ-5D mobility, self-care, usual activities, pain/discomfort, and anxiety/depression questions. It is then possible to compare

actual and predicted responses to each EQ-5D question, and to construct a new tariff from the 5 predicted responses, again using the MVA survey results. In this case, the dependent variables are categorical variables with discrete outcomes. One option would be the use of multinomial logistic regression to predict probability of each response level for the five questions.(Nerlove & Press 1973) While the multinomial logit model provides unbiased parameter estimates, it is inefficient because it ignores the fact that the categorical responses to each EQ-5D question are ordered. Instead we use ordered or ordinal logistic regression.(Zavoina & McElvey 1975) Here we follow expositions by Greene(Greene 1997) and Long.(Long 1997) We can consider y as a measure that provides partial information about the underlying or latent variable y^* as follows:

$$y_i = 1 \text{ if } y_i^* \leq \tau_1 \quad (1.1)$$

$$y_i = 2 \text{ if } \tau_1 < y_i^* \leq \tau_2 \quad (1.2)$$

$$y_i = 3 \text{ if } y_i^* \geq \tau_2 \quad (1.3)$$

Each τ is a parameter representing a cutpoint or threshold separating the categories in the observed variable, and can be estimated according to a structural model where

$$y_i^* = \sum_{k=1}^K \beta_k x_{ki} + \varepsilon_i = Z_i + \varepsilon_i \quad (2.)$$

We therefore estimate the unknown τ parameters with β , where K represents the set of measured independent predictors, in this case SF-36 individual responses or domain scores. We assume that the random disturbance term ε has a logistic distribution. This gives us

$$Z_i = \sum_{k=1}^K \beta_k X_{ki} \quad (3.)$$

It is then possible to compute Z_i for each case and calculate the probability that the case falls into each category J , using the estimated ε parameters as threshold limits:

$$P(y = 1) = \frac{1}{1 + \exp(Z_i - \tau_1)} \quad (4.1)$$

$$P(y=2) = \frac{1}{1 + \exp(Z_i - \tau_2)} - \frac{1}{1 + \exp(Z_i - \tau_1)} \quad (4.2)$$

$$P(y=3) = 1 - \frac{1}{1 + \exp(Z_i - \tau_2)} \quad (4.3)$$

Here, ordered logit was performed in Stata v.7 using the ordered logit procedure OLOGIT, and replicated in SPSS with the PLUM procedure: these gave identical results.

Mapping to the EQ-5D tariff

In the second part of the analysis, regression analysis was employed to model the relationship between actual EQ-5D tariff values and responses to the SF-36 using either domain summary scores or individual items. The time-trade off valuations that form the basis of the tariff values for the EQ-5D states mean that tariff scores are bounded by 1.0 (the score for full health) when a respondent indicates no problems in any dimension (i.e., 11111 on the EQ-5D survey). (Dolan, Gudex, Kind, & Williams 1996) Population health surveys that have included the EQ-5D indicate that a significant fraction of respondents rate themselves in full health (e.g. 52% of respondents for the Health Survey for England 1996). Another feature of the EQ-5D tariffs is that they can be highly skewed, with some patients recording health states that yield negative tariff values to a minimum of -0.594. Both these features mean that it is inappropriate to employ conventional Ordinary Least Squares analysis. In the following, we focus on deviation of tariff from full health and denote this as M_i where $M_i = 1 - U_i$.

We employ a two part model. In the first part, Probit regression is used to model the probability of the respondent being at full health. More formally probability of a patient having full health is calculated by:

$$\Pr(M_i = 0 | X_i^j) = \Phi(\alpha^j X_i^j) \quad (5.)$$

Where M_i is the deviation in tariff values from full-health; X_i a vector of j independent variables the $i = (1..N)$ individuals responding to the survey; α^j is a vector of coefficients; and Φ is the cumulative distribution function of the standard normal distribution. As the maximum likelihood estimates for the standard Probit model have shown to be inconsistent in the presence of heteroscedasticity, (Yatchew & Griliches 1985) we employ a more general Probit model that allows a vector of variables ($\gamma^k H_i^k$) that influence the variance of the error term. (Harvey A 1976) Under this more general model (1) is replaced by:

$$\Pr(M_i = 0 | X_i^j) = \Phi(\alpha^j X_i^j / \exp(\gamma^k H_i^k)) \quad (6.)$$

In the second stage only the deviations from full health are modelled. Two part models typically transform the dependent variable when it is skewed. Adopting this approach we use a Box-Cox transformation to explore alternative models. The transformation is:

$$MT_i = \frac{M_i^\lambda - 1}{\lambda}$$

Linear model are then estimated using maximum likelihood methods and standard statistical test applied to determine an optimal value for λ (i.e. one that produces the highest value of the log likelihood function).

The predicted values for these deviations can then be obtained by:

$$E[MT_i | M_i > 0] = \beta^j X_i^j \quad (7.)$$

We test the predictive performance of the non-linear model against a linear alternative (i.e. $E[M_i | M_i > 0] = \beta^j X_i^j$).

Comparisons

Having conducted the analyses above, we calculated for each case in the sample the most probable response category for each dimension of the EQ-5D using ordered logit, a predicted tariff by applying the MVA valuations to the set of predicted response categories, and a predicted EQ-5D tariff using the two-part model. We then examined the correspondence between the predicted and actual EQ-5D question responses, and the correspondence between predicted and actual EQ-5D tariffs, by plotting differences, calculating the proportion of predicted tariff values within 5 points of the actual tariffs, and calculating correlation. We also examined the performance of the actual tariff, the predicted tariffs and the utility values derived from the Brazier and Lundberg algorithms in discriminating between individuals reporting the presence or absence of long-standing illness.

Data

We used information from the Health Survey for England 1996, one of a series of annual surveys commissioned by the Department of Health in which a representative sample of adults and children living in private households in England were interviewed to provide information about health and risk factors; in the 1996 survey, the interview schedule included the SF-36 and the EQ-5D. (Health Survey 1998) A total of 15,978 persons over the age of 18 were included in the survey, of whom 3,171 (19.9%) were dropped because of one or more missing data items, giving a final total of 12,753 cases for analysis. Table 2 provides descriptive data for the dataset.

Results

Mapping to ordinal EQ-5D responses

Table 3 reports the ordinal logit results for each of the EQ-5D questions, using the SF-36 domain scores and change in health score as the predictor variables. Model fit statistics indicate that the null hypothesis that all effects of independent variables are zero can be rejected, and the Nagelkerke pseudo R^2 measures indicates that the model performs well. The coefficients for individual variables in this type of model are not straightforward to interpret, (Greene 1997) but the results indicate that the domain scores most likely to be related to EQ-5D questions are generally highly significant, while those least directly related are non-significant: for example, the SF-36 bodily pain domain score is very significantly related to the EQ-5D pain question but not to the EQ-5D anxiety question; the SF-36 role emotional domain score is not significantly related to the EQ-5D mobility, usual activity or pain questions, but is significantly related to the EQ-5D anxiety question, and so on.

For any individual case in the sample, the coefficients can be used to predict Z_i , as set out in equation 2 above, and this can then be combined with the cutoff parameters to calculate the probability that the case is in each category of the EQ-5D question.

For example, a patient scoring 85, 100, 0, 55.56, 56, 25, 66.67, 62, and 25 respectively on the 8 domains and change-in-health items of the SF-36 would have a Z_i of -5.61, and using the equations given above (4.1 to 4.3) it can be estimated that their probability of answering 1, 2 or 3 to the EQ-5D anxiety question is 0.18, 0.77 and 0.04 respectively; they would therefore be classified in category 2 – moderately anxious or depressed.

A similar approach was adopted using responses to all 36 individual questions in the SF-36, but with each predictor variable entered as a categorical variable converted into a set of dummy variables for each level. The full sets of coefficients are not reported here for reasons of space, but Table 4 provides summary data on the actual frequency distribution across each EQ-5D question in the survey, and the predicted distribution based on all SF-36 questions and on the domain scores. Looking at the domain score results first, it is evident that the ordered logit procedure correctly places between 80% and 97% of all cases into the category respondents actually selected (that is, the proportion of cases on the cross-tab diagonal). For the mobility and self-care questions the overall numbers correctly classified are particularly high. However, it is also evident that the proportion of people in category 3 is consistently under-predicted, although the actual numbers involved is small. This is true of all 5 EQ-5D questions, with no category 3 responses predicted at all in the mobility and self-care questions. Finally, Table 4 shows that the improvement obtained by using all SF-36 questions rather than the domain scores as predictor variables is very slight.

Having predicted for each case in the survey a response category for each EQ-5D question, it was possible to calculate a predicted tariff score. Table 5 reports the mean and dispersion statistics for the actual EQ-5D tariffs, the predicted tariffs based on the 8 SF-36 domain scores and on all 36 individual questions, and the estimated utility from the Brazier algorithm and the Lundberg algorithm (reduced form TTO model).

The table shows that the two ordered logit predicted tariffs have a mean value approximately 4 percentage points above the actual tariff, which may reflect the underestimation of category 3 responses. The Brazier utility estimate is approximately 10 percentage points lower than the actual EQ-5D tariff, and the Lundberg algorithm approximately 3 percentage points above the actual tariff. Also of note, the ranges of the two ordered logit predictions, and in particular the estimate based on SF-36 domain scores, are substantially wider than the Brazier or Lundberg estimates, which give minimum utility values of 0.30 and 0.38 respectively in this sample.

An alternative way of looking at agreement and divergence between these different estimates of utility is given in Figure 1. The actual EQ-5D tariff value is subtracted from the estimated utility level from the two ordered logit models, the Brazier algorithm and the Lundberg algorithm, and the differences are plotted on barcharts. It is immediately evident that the two ordered logit models consistently predict utility values that are close to the actual value: in the case of the ordered logit model using all SF-36 questions as predictors (p1dist), 65% of predictions are within 5 percentage points of the actual value, and for the ordered logit model based on SF-36 domain scores the corresponding figure is 64%. In comparison, the Brazier algorithm produces estimates within 5 percentage points of the actual EQ-5D tariff in only 16% of cases, and the Lundberg algorithm in only 37% of cases.

Finally, we compare the 5 measures of utility in terms of their ability to discriminate between individuals who reported the presence or absence of long-standing illness in a separate standard question administered during the Health Survey for England. In the sample examined here 5,316 answered “yes” and 7,437 answered “no”. Figure 3 shows the mean utility values for each of these groups according to the 7 measures being considered (including 2 based two part models - see below), and the mean difference in utility, with corresponding confidence intervals. In the Survey, those with long-standing illness scored a mean tariff of 0.75 on their EQ-5D responses, while those without long-standing illness had a mean tariff of 0.93, a difference of 0.18. the ordered logit approach using SF-36 domain scores predicted slightly higher mean scores in both groups (0.81, 0.96) and a mean difference of 0.15, similar to ordered logit using all SF-36 questions. The Brazier algorithm estimated significantly lower utility scores in both groups and a significantly smaller mean difference of 0.10, while the Lundberg algorithm estimated significantly higher utility values than the actual EQ-5D tariff in the group with long-standing illness, and the smallest mean difference of just 0.9.

Mapping to the EQ-5D tariff

Table 6 reports the two-part model for the EQ-5D tariff, using the SF-36 domain scores and change in health score as the predictor variables. The first part reports a Probit model that allows for a heteroscedasticity in the error term. While all independent variables were included as auxiliary variables explaining the variance, an initial analysis indicated that heteroscedasticity was closely associated with the pain domain of the SF-36 and so we report an equation with this more parsimonious specification. Table 6 also report two regression models for the second part. The first model uses ordinary least squares on the untransformed tariff values. In the second, a grid search indicated that optimal value of λ was -0.631 and the regression after applying this transformation is also reported in Table 6. A similar approach was used to model tariff values using individual SF-36 item responses, but again for the sake of brevity these are not reported.

These two-part models were then used to predict a tariff value for each individual in the data set. This involved using (6) to predict probability of an individual being at full health and then we assume if $\Pr(M_i = 0 | X_i^j) > 0.5$ then $\hat{M}_i = 0$. The second part of the model uses (7) to estimate \hat{M}_i for any individual whose $\Pr(M_i = 0 | X_i^j) < 0.5$. Summary statistics of the predictions are reported in bottom half of Table 5, in the same format as the ordered logit results. The predictions from the linear model are closer to the actual tariff values while the non-linear model predicts tariffs values over a greater range. The predictions of the regression model based on SF-36 individual item responses are slightly closer to the actual mean tariff value and so we focus on these models in the remaining analysis.

The difference between predicted & actual tariffs for the linear model is illustrated in Figure 2a and 2b, again using the same format as used to report ordered logit results. In the case of the linear model that is based on domains 56% of predictions are within 5% of the actual values (see figure 2a), while in the model based on items 59% fell in this range (see figure 2b). Finally the ability of these algorithms to discriminate

between the utilities of patients with longstanding illness is illustrated in Figure 3. Both two part models that use linear regression produce a difference that is not significantly different from tariff values.

Discussion

In this paper we have shown that previously published algorithms to obtain utility values from the SF-36 or SF-12 do not necessarily provide results that correspond either with each other or with those obtained from the EQ-5D, and we have set out alternative methods for reliable prediction of EQ-5D responses and utility levels.

We have produced similar results using ordered logit to predict EQ-5D response categories from SF-36 domain scores and thence calculate utilities from these predicted responses, and by using two-part regression models to predict tariffs directly from these domain scores. The ordered probit approach produced a slightly higher proportion of estimated tariffs within 5 points of the actual tariff values, but the two-part models produced estimates of mean tariff and mean difference with respect to long-standing illness that were closer to the actual mean. Both methods performed better in these respects than the Brazier or Lundberg algorithms. Analysts may find both methods useful: the ordered logit approach will indicate the EQ-5D dimensions in which most health gains or losses are concentrated, and the two-part model will provide mean tariff estimates closely approximating those the EQ-5D would have provided.

We found that the ordered logit approach slightly under-estimated the number of respondents likely to indicate level 3 on EQ-5D responses, and this in turn resulted in a slight (4%) over-prediction of utility levels. This may be a manifestation of recognised floor effects in the SF-36, especially in the role physical and role emotional domains.

We noted earlier and wish to restate that we do not consider the EQ-5D to be a gold standard, and we also recognise that previously published algorithms did not necessarily have the objective of producing results consistent with the EQ-5D. However, given the widespread use of the EQ-5D and its high degree of support amongst technology appraisal and reimbursement agencies, an algorithm that can produce from the SF-36 utility estimates consistent with the EQ-5D should be a useful addition to the armamentarium of health economists.

There are a number of areas in which we are currently extending the work reported here. First, we are examining the effect of adding a small number of other readily available independent variables to our models, to reduce unexplained variance and further improve prediction. In particular, we are looking at the impact of including age and sex. Second, we are examining the model specification through the use of additional explanatory variables constructed from interactions between SF-36 domain scores, in recognition that the values taken by these domain scores are unlikely to be independent. Third, we are applying the approach to other datasets, in particular the 2000 Medical Expenditure Panel Study, in which SF-12 and EQ-5D responses were obtained simultaneously from a large US population sample. Fourth, we are testing the validity of the approach in trial-based datasets, to make blinded predictions of

actual EQ-5D responses, mean utility levels and mean differences by allocation. Fifth, we are looking at ways of dealing with the apparent floor effects in domains of the SF-36, to further improve the mean tariff values we obtain with the ordered logit approach. Finally, we are preparing easily accessible versions of our algorithms in Microsoft Excel, SPSS and STATA, which will be downloadable from the HERC website at the University of Oxford at some stage in the future. We will also provide reference health profiles and results so that users can ensure correct implementation of these algorithms.

In conclusion, for analysts who might wish to derive from SF-36 or SF-12 responses estimates of health states and utility values that will approximate to those the EQ-5D would have given had it been administered, we have demonstrated that our direct mapping method provides reliable and accurate results: the predicted responses and utility values will correspond quite closely to the values that would have been obtained had the EQ-5D been used, and will almost certainly correspond more closely to EQ-5D utilities than other published algorithms.

Table 1: Description of SF-36 domains and what they measure

Dimension title and Description	No. of items	Low Scores	High Scores
Physical Functioning (PF)	10	Limited a lot in performing types of physical activities including bathing and dressing	Performs all activities without limitations due to health
Role Physical (RP)	4	Problems with work or other daily activities as a result of physical health	No problems with work or other daily activities due to physical health
Bodily Pain (BP)	2	Severe and limiting bodily pain	No pain or limitations due to pain
Energy / Vitality (VT)	4	Feels tired and worn down all the time	Feels full of energy all the time
Social Functioning (SF)	2	Extreme and frequent interference with normal social activities due to physical or emotional problems	Performs normal activities without interference due to physical or emotional problems
Role Limits Emotional (RE)	3	Problems with work or other daily activities as a result of emotional problems	No problems with work as a result of emotional problems
Mental Health (MH)	5	Feelings of nervousness and depression all of the time	Feels peaceful, happy and calm all of the time
General Health (GH)	5	Believes personal health is poor and likely to get worse	Believes personal health is excellent
Physical component summary	35	Limitations in self-care, physical, social, and role activities, severe bodily pain, frequent tiredness	No physical limitations, disabilities, or decrements in well-being and high energy level
Mental component summary	35	Frequent psychological distress, social and role disability due to emotional problems	Frequent positive affect, absence of psychological distress and emotional problems

Table 2: Descriptive data from Health Survey for England 1996 (n=12,753)

	Mean	S. Dev.	Minimum	Maximum
Age last birthday	46.6	17.2	18	102
SF-36 domain scores:				
sf36pfs = physical functioning	83	25	0	100
sf36rps = role physical	81	35	0	100
sf36res = role emotional	85	31	0	100
sf36scs = social functioning	76	20	0	100
sf36mhs = mental health	76	17	0	100
sf36evs = energy/vitality	63	20	0	100
sf36ps = bodily pain	80	25	0	100
sf36hps = general health perception	70	21	0	100
Sf36chs Change in health	51	17	0	100
EQ-5D tariff	0.85	0.22	-0.36	1.00
Sex:	47.3% men, 52.3% women			
Smoking status:	28.4% current, 26.5% ex, 45.2% never			
Long-standing illness:	41.7% yes, 58.3% no			
Social class:	Professional 4.5%, Managerial technical 27.2%, skilled nonmanual 25.2%, skilled manual 19.7%, semi-skilled manual 16.7%, unskilled manual 6.3%, other 0.3%			

Table 3: Ordered logit SF-36 domain score coefficients and significance for each EQ-5D question, using Health Survey for England data

	Mobility		Self-care		Usual activity		Pain		Anxiety	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Level 1 cutoff	-4.612	0.000	-2.44	0.000	-5.727	0.000	-7.124	0.000	-7.071	0.000
Level 2 cutoff	3.147	0.000	1.082	0.000	-1.386	0.000	-1.678	0.000	-2.398	0.000
sf36pfs – physical functioning	-0.052	0.000	-0.053	0.000	-0.029	0.000	-0.018	0.000	0.001	0.520
sf36rps - role physical	-0.007	0.000	-0.009	0.000	-0.017	0.000	0.002	0.017	0.005	0.000
sf36res - role emotional	0.002	0.152	0.006	0.000	-0.001	0.572	-0.001	0.395	-0.012	0.000
sf36scs – social functioning	-0.001	0.571	-0.014	0.000	-0.012	0.000	0.01	0.000	-0.012	0.000
sf36mhs – mental health	0.01	0.000	-0.011	0.001	0.002	0.396	0.002	0.224	-0.078	0.000
sf36evs - energy/vitality	-0.003	0.220	0.001	0.827	-0.011	0.000	-0.003	0.069	-0.006	0.001
sf36ps - bodily pain	-0.024	0.000	-0.01	0.000	-0.019	0.000	-0.069	0.000	-0.001	0.527
sf36hps – gen. health percep.	-0.014	0.000	-0.013	0.000	-0.016	0.000	-0.022	0.000	-0.014	0.000
sf36chs - change in health	-0.001	0.525	0.006	0.022	-0.003	0.132	-0.003	0.062	-0.003	0.089
Model fit Chi-2	5657		2661		6463		7547		5509	
Model fit sig.	0.000		0.000		0.000		0.000		0.000	
Nagelkerke Pseudo R2	0.609		0.565		0.607		0.575		0.511	

Table 4: Frequency distribution of actual EQ-5D response categories and predicted categories using ordered logit, Health Survey for England data

	Actual category distribution (%)			Predicted category distribution (%)			Overall % correctly categorised
	1	2	3	1	2	3	
Using 8 SF-36 domain scores as predictors:							
Mobility	84.1	15.8	0.1	87.4	12.6	0	91.2
Self-care	95.3	4.4	0.4	96.7	3.3	0	96.9
Usual activities	82.5	15.2	2.3	86.5	12.3	1.2	88.9
Pain/discomfort	65.3	31.6	3.1	71.8	26.3	1.8	80.0
Anxiety/Depression	78.4	20.1	1.6	84.7	14.6	0.7	84.7
Using all 36 SF-36 questions as predictors:							
Mobility	84.1	15.8	0.1	86.7	13.3	0	92.4
Self-care	95.3	4.4	0.4	96.0	4.0	0	97.3
Usual activities	82.5	15.2	2.3	86.0	13.0	1.0	89.4
Pain/discomfort	65.3	31.6	3.1	68.4	29.9	1.7	81.3
Anxiety/Depression	78.4	20.1	1.6	83.1	16.5	0.4	85.3

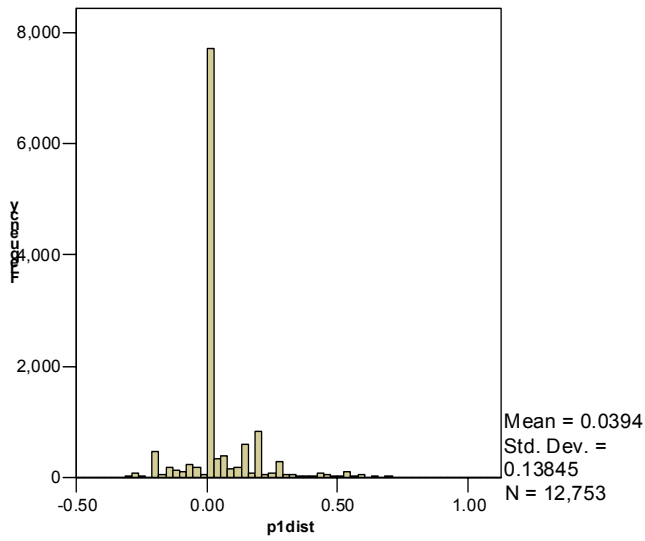
Table 5: Actual EQ-5D tariff and estimated quality of life tariffs using ordered logit & two part models, Brazier algorithm and Lundberg algorithm (n=12,753), Health Survey for England data

	Mean	S.D.	Range	Min	Max
Actual EQ-5D tariff	0.85	0.22	1.36	-0.36	1
<i>Tariffs based on predicted item responses using ordered logit</i>					
Predicted tariff 2 (using domain scores)	0.9	0.19	1.24	-0.24	1
Predicted tariff 1 (using all SF-36 Qs)	0.89	0.16	0.97	0.03	1
<i>Tariffs estimated directly using two part models</i>					
Predicted tariff 3 (using domain scores & linear two-part model)	0.85	0.17	0.89	0.11	1
Predicted tariff 1 (using all SF-36 Qs & linear two-part model)	0.85	0.20	1.20	-0.20	1
Predicted tariff 5 (using domain scores & non-linear two part model)	0.87	0.17	1.19	-0.19	1
Predicted tariff 6 (using all SF-36 Qs & non-linear model)	0.86	0.02	1.24	-0.24	1
Brazier utility estimate	0.75	0.12	0.62	0.30	0.92
Lundberg utility estimate	0.88	0.11	0.62	0.38	1

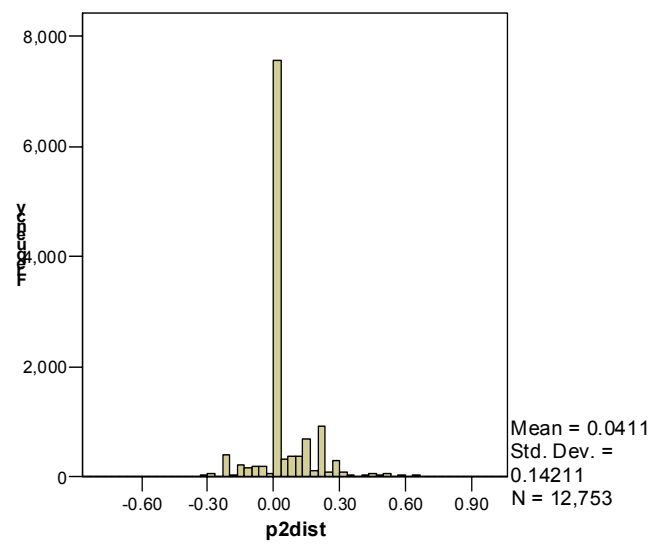
Table 6: Two part model SF-36 domain score coefficients and significance for each EQ-5D question, using Health Survey for England data

	Probit model		Linear Regression		Non-linear Regression	
	Estimate	Sig.	Estimate	Sig.	Estimate	Sig.
Constant	-2.529	0.119	0.892	0.014	0.914	0.000
sf36pfs – physical functioning	0.005	0.000	-0.002	0.000	0.000	0.004
sf36rps - role physical	0.001	0.000	0.000	0.000	0.000	0.694
sf36res - role emotional	0.001	0.000	0.000	0.000	0.002	0.128
sf36scs - social functioning	0.002	0.000	-0.001	0.000	0.001	0.000
sf36mhs – mental health	0.006	0.001	-0.001	0.000	-0.001	0.001
sf36evs - energy/vitality	0.002	0.000	0.000	0.000	0.006	0.084
sf36ps - bodily pain	0.011	0.001	-0.003	0.000	0.002	0.000
sf36hps – gen. health percep.	0.003	0.000	-0.001	0.000	0.001	0.000
sf36chs - change in health	0.000	0.000	-0.001	0.000	0.914	0.001
Λ					-0.631	
Γ	-0.013	0.001				
Model fit Chi-2	455		349			
Model fit sig.	0		0		0	
R2	0.395		0.528		0.539	

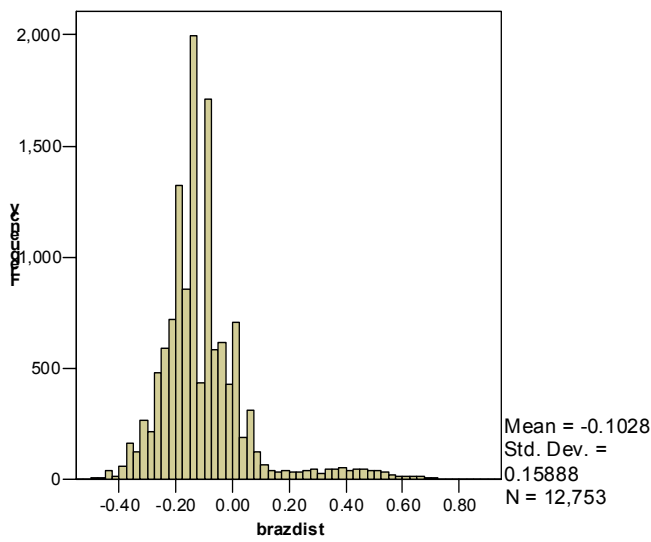
Figure 1 a-d: Distribution of differences (predicted minus actual) between actual Eq-5D tariffs and predicted tariffs using ordered logit with a) SF-36 questions (P1dist), b) ordered logit with SF-36 domains (p2dist), c) Brazier algorithm (Brazdist) and d) Lundberg algorithm (Lunddist).



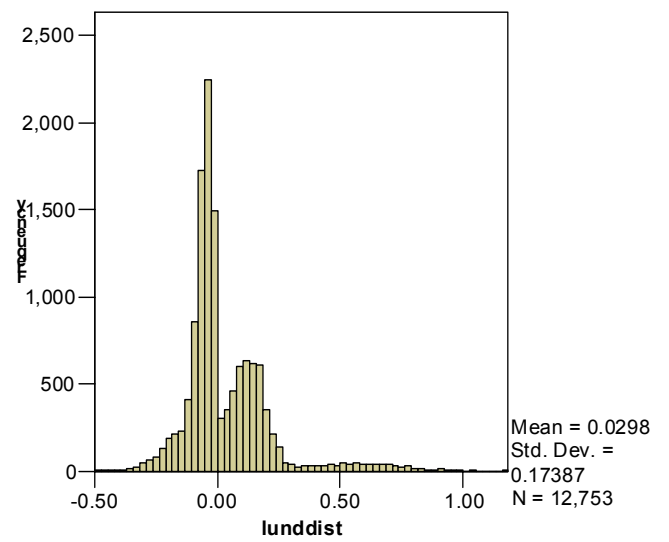
(a)



(b)

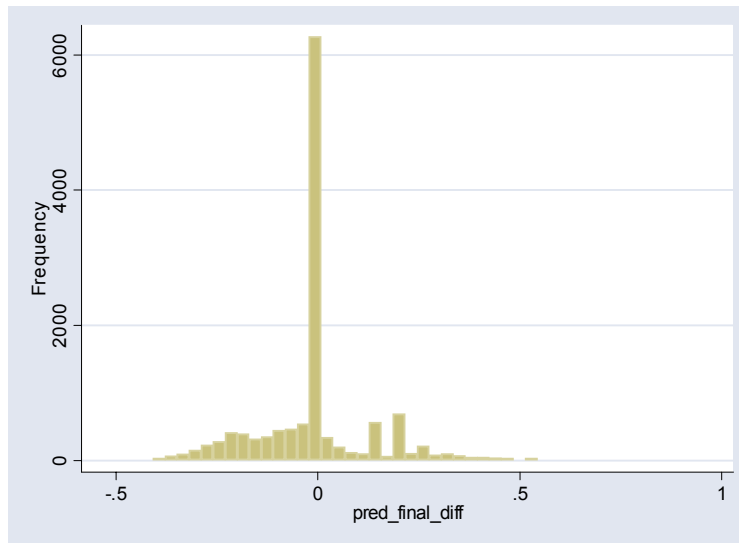


(c)

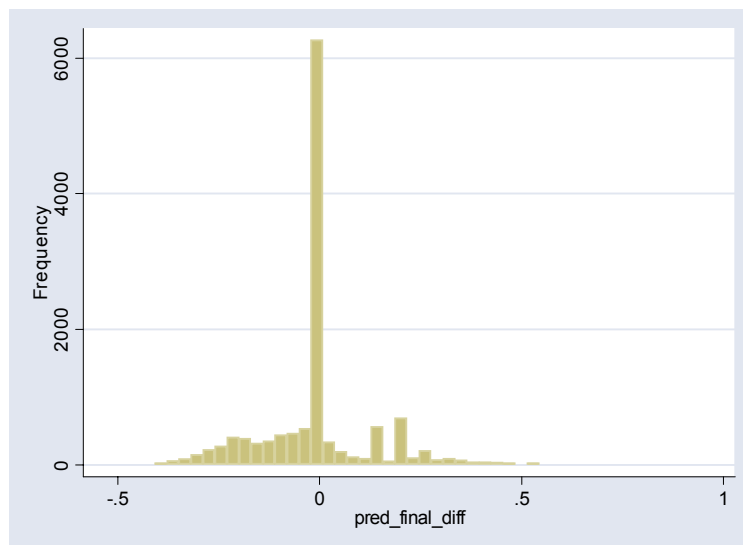


(d)

Figure 2 a-b: Distribution of differences (predicted minus actual) between actual Eq-5D tariffs and predicted tariffs using two-part models with SF-36 questions.

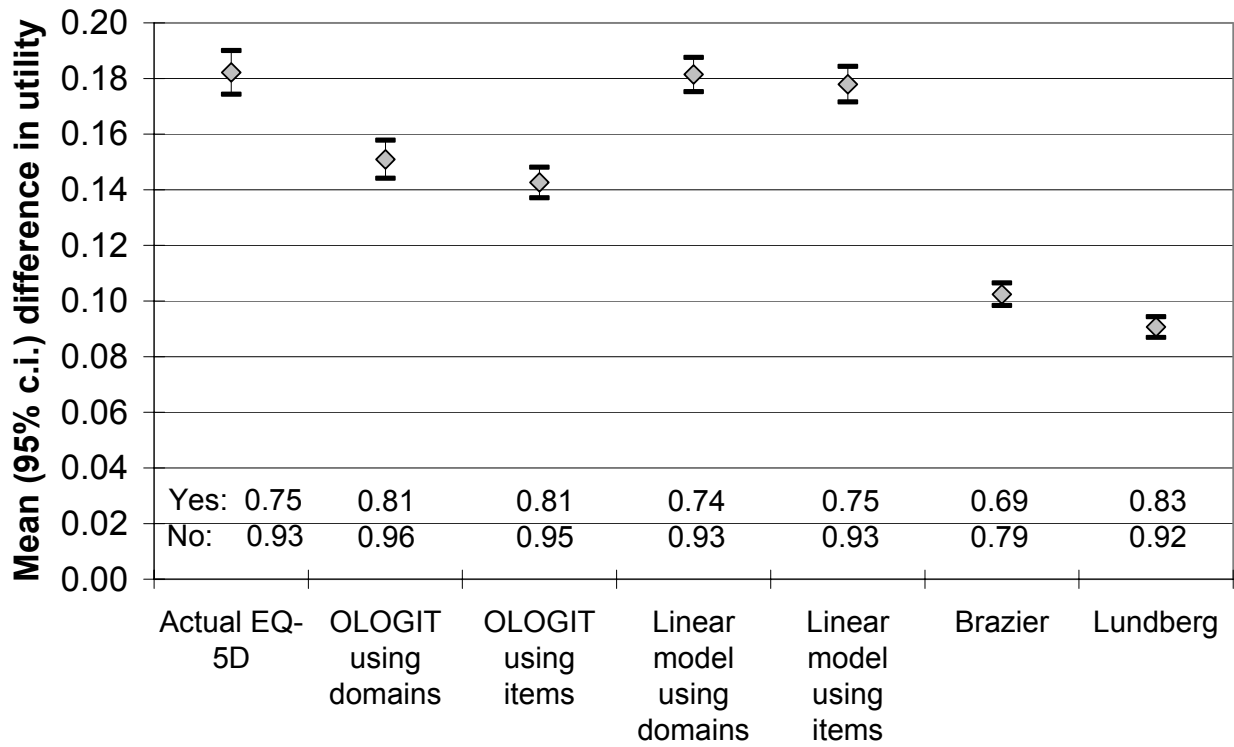


(a)



(b)

Figure 3: Mean (95% c.i.) utility values for individuals with (n=5,316) or without (n=7,437) long-standing illness, by actual EQ-5D tariff, ordered logit using SF-36 domain scores and individual items, linear model using domain scores and individual items, and Brazier algorithm and Lundberg algorithm.



References

- Brazier, J., Roberts, J., & Deverill, M. 2002, "The estimation of a preference-based measure of health from the SF-36", *J.Health Econ.*, vol. 21, no. 2, pp. 271-292.
- Brazier, J., Usherwood, T., Harper, R., & Thomas, K. 1998, "Deriving a preference-based single index from the UK SF-36 Health Survey.", *J.Clin.Epidemiol.*, vol. 51, no. 11, pp. 1115-1128.
- Dolan, P., Gudex, C., Kind, P., & Williams, A. 1996, "The time trade-off method: results from a general population study", *Health Econ.*, vol. 5, no. 2, pp. 141-154.
- EuroQol Group 1990, "EuroQol - a new facility for the measurement of health-related quality of life", *Health Policy*, vol. 16, pp. 199-208.
- Fryback, D. G., Lawrence, W. F., Martin, P. A., Klein, R., & Klein, B. E. 1997, "Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcomes Study.", *Medical Decision Making*, vol. 17, no. 1, pp. 1-9.
- Greene, W. H. 1997, *Econometric Analysis*, 4 edn, Prentice-Hall, London.
- Harvey A 1976, " Estimating Regression Models with Muliiplicative Heteroscedasticity", *Econometrica*, vol. 44, pp. 461-465.
- Health Survey 1998, *Health Survey for England 1996* The Stationery Office, London.
- Hollingworth, W., Deyo, R. A., Sullivan, S. D., Emerson, S. S., Gray, D. T., & Jarvik, J. G. 2002, "The practicality and validity of directly elicited and SF-36 derived health state preferences in patients with low back pain", *Health Econ.*, vol. 11, no. 1, pp. 71-85.
- Lee, T. A., Hollingworth, W., & Sullivan, S. D. 2003, "Comparison of directly elicited preferences to preferences derived from the SF-36 in adults with asthma", *Med Decis.Making*, vol. 23, no. 4, pp. 323-334.
- Long, J. S. 1997, *Regression models for categorical and limited dependent variables*. Sage, Thousand Oaks.
- Lorgelly PK "Mapping SF-36 to Utilities: How can it be done?", *paper presented at HESG meeting*, August 2001..
- Lundberg, L., Johannesson, M., Isacson, D. G., & Borgquist, L. 1999, "The relationship between health-state utilities and the SF-12 in a general population", *Medical Decision Making*, vol. 19, no. 2, pp. 128-140.
- Nerlove, M. & Press, S. 1973, *Univariate and multivariate log-linear and logistic models*. RAND-R1306-EDA/NIH Rand Corporation, Santa Monica, Calif.
- Shmueli, A. 1999, "Subjective health status and health values in the general population", *Med Decis.Making*, vol. 19, no. 2, pp. 122-127.

Ware, J.-E. J. & Gandek, B. 1998, "Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project", *J.Clin.Epidemiol.*, vol. 51, no. 11, pp. 903-912.

Yatchew, A. & Griliches, Z. 1985, "Specification Error in Probit Models", *Review of Economics and Statistics*. February 1985, vol. 67: 134-39 39.

Zavoina, R. & McElvey, W. 1975, "A statistical model for the analysis of ordinal level dependent variables", *Journal of Mathematical Sociology*, vol. Summer, pp. 103-120.