Monograph:
Rowen, D., Brazier, J. and Roberts, J. (2008) Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? Discussion Paper. (Unpublished)

HEDS Discussion Paper 08/14

HEDS Discussion Paper 08/14


ScHARR
School of Health and
Related Research

The University Of Sheffield.

ScHARR

Health Economics and Decision Science Discussion Paper Series

No. 08/14

Mapping SF-36 onto the EQ-5D index: how reliable is the relationship?

Donna Rowen[a]*, John Brazier[a] and Jennifer Roberts[b]

[a] Health Economics and Decision Science, University of Sheffield, Regent Court, 30Regent Street, Sheffield, S1 4DA, UK
[b] Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield, S1 4DT, UK
* Correspondence to: Donna Rowen, Health Economics and Decision Science,University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK.
Telephone: +44114 222 0728.
Fax: +44114 272 4095.
Email:d.rowen@sheffield.ac.uk

Summary

Mapping from health status measures onto generic preference-based measures is becoming a common solution when health state utility values are not directly available for economic evaluation. However the accuracy and reliability of the models employed is largely untested, and there is little evidence of their suitability in patient datasets. This paper examines whether mapping approaches are reliable and accurate in terms of their predictions for a large and varied UK patient dataset. SF-36 dimension scores are mapped onto the EQ-5D index using a number of different model specifications. The predicted EQ-5D scores for subsets of the sample are compared across inpatient and outpatient settings and medical conditions. This paper compares the results to those obtained from existing mapping functions. Our results suggest that models mapping the SF-36 onto the EQ-5D have similar predictions across inpatient and outpatient setting and medical conditions. However, the models overpredict for more severe EQ-5D states; this problem is also present in the existing mapping functions.

Key words: health status; SF-36; SF-12; EQ-5D; utility; mapping.

Acknowledgements: We would like to thank Cardiff Research Consortium for use of the HoDAR data.

Funding sources: John Brazier is funded by the UK Medical Research Council.

Introduction

Clinical trials use a multitude of health status measures in order to measure health and health related quality of life. However, most of these measures cannot be used in assessments of cost effectiveness using cost per Quality Adjusted Life Year (QALY). Preference-based measures such as the EQ-5D are commonly used to do this, but are not always used in clinical studies. One solution to this problem is to apply a mapping function to convert non-preference based health data into one of the generic preference-based measures; this is helpful to those submitting evidence to agencies such as NICE (2008). However the accuracy and reliability of the mapping models employed is largely untested, and there is little evidence of their suitability in patient datasets.

A recent review of mapping non-preference-based measures onto generic preference-based measures (Brazier et al, 2008) found 29 studies. However, most of these used simple OLS modelling procedures on comparatively small data sets. Further, existing studies have neglected to investigate the robustness of the models across patient data sets.

The purpose of this paper is to examine whether mapping models are reliable and accurate in terms of their predictions for a large and varied patient dataset. The mapping relationship examined here is between the EQ-5D index, a generic preference-based measure of health related quality of life and the SF-36, a generic non-preference-based health status measure commonly used in clinical trials. A mapping relationship is estimated using a range of techniques and statistical specifications. We examine the mapping relationship across inpatient and outpatient settings and medical conditions according to ICD classification. Furthermore, we compare the mapping approach used here to the existing models of Franks et al. (2004) and Gray et al. (2006) in terms of predictive performance.

Methods

The model

The SF-36 assesses health across eight dimensions using 36 items. The SF-36 produces a score on a 0-100 scale for each of the eight dimensions, which are specific health domains such as physical functioning, social functioning and vitality. These scores are not comparable across dimensions and are not based on individual

preferences, therefore they cannot be used to generate QALYs. The SF-36 can be used to generate a preference-based index via the SF-6D (Brazier et al., 2002).

The EQ-5D is the most widely used generic preference-based measure of health-related quality of life which produces utility scores anchored at 0 for death and 1 for perfect health. The utility scores represent preferences for particular health states. The descriptive system has 5 dimensions (mobility, self-care, usual activity, pain/discomfort and anxiety/depression) and 3 levels (no problems, some problems, extreme problems) which create 243 unique health states. This study uses the UK TTO value set in its main analysis (Dolan, 1997). The EQ-5D valued using the UK TTO value set is preferred by NICE (NICE, 2008). The SF-6D has been found to differ from the EQ-5D (Brazier et al., 2004) and so to achieve comparability between studies this paper explores an alternative strategy of mapping.

Model specifications

Regression analysis is used to examine the relationship between the EQ-5D utility score and the SF-36 using the 8 dimension scores; physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional and mental health, squared dimension scores and interaction terms derived using the product of two dimension scores. The dependent variable, the EQ-5D utility score, is measured on a -1 to 1 scale. The 8 dimension scores of the SF-36 are rescaled onto a 0-1 scale to enable easier interpretation of the results and the squared terms and interaction terms are generated using the rescaled scores.

Three models are estimated: (1) all dimensions; (2) all dimensions and squared terms; (3) all dimensions, squared terms and interactions. The general model is defined as

$$y_i = \alpha + \boldsymbol{\beta} x_{ij} + \boldsymbol{\theta} r_{ij} + \boldsymbol{\delta} z_{ij} + \varepsilon_{ij} \quad (1)$$

where $i = 1, 2, ..., n$ represents individual respondents and $j = 1, 2, ..., m$ represents the 8 different dimensions. The dependent variable, y, represents the EQ-5D utility score, x represents the vector of SF-36 dimensions, r represents the vector of squared terms, z represents the vector of interaction terms and ε_{ij} represents the error term. This is an additive model which imposes no restrictions on the relationship between dimensions. The squared terms are designed to pick up non-linearities in the relationship between

dimension scores and the EQ-5D index. There is no reason for it to be linear and there is evidence in physical functioning, for example, that the same differences in scores at the lower end of the scale indicate larger differences in functioning than at the upper end (Brazier et al., 1998). Interaction terms are important since there is evidence from other measures that dimensions are not additive (Feeny et al, 2002). Statistical measures of explanatory power, predictive ability, and model specification are reported.

The sample used here is a patient dataset (described below) where respondents are included each time they are treated, and hence some respondents have multiple observations. Random effects models are used to take account of this data structure. The estimated models are used to generate predicted EQ-5D scores. Predictive ability is assessed using line graphs of the observed and predicted EQ-5D utility scores ordered by observed tariff value of EQ-5D state, mean error, mean absolute error and mean squared error.

EQ-5D utility scores are known to exhibit a ceiling effect, where a large proportion of subjects rate themselves in full health with a utility score of 1, and hence the data can be interpreted as being bounded or censored at 1. Ignoring the bounded nature of the EQ-5D will result in biased and inconsistent estimates, and hence the random effects tobit model is an appropriate alternative (Sullivan and Ghushchyan, 2006). The tobit model with an upper censoring limit of 1 is defined as

$$y_i^* = \alpha_i + \boldsymbol{\beta} x_{ij} + \boldsymbol{\theta} r_{ij} + \boldsymbol{\delta} z_{ij} + \varepsilon_{ij}$$

$$y_i = \begin{cases} y_i^* \text{ if } y_i^* < 1 \\ 1 \text{ if } y_i^* \geq 1 \end{cases} \qquad (2)$$

where y_i^* is the observed EQ-5D utility score and y_i is the bounded measure of the EQ-5D score.

However, the tobit model also produces biased estimates in the presence of heteroscedasticity or non-normality (see Greene, 2000, and Sullivan and Ghushchyan, 2006). The censored least absolute deviations (CLAD) model is also used here since it produces consistent estimates in the presence of heteroscedasticity and non-normality (see Powell, 1984, and Sullivan and Ghushchyan, 2006).[1]

Reliability and robustness

In order to examine whether the estimated relationships are reliable and robust across inpatient and outpatient setting and medical conditions, we estimate model (3) as outlined above for subsets of the sample data[1]. The model is estimated for inpatients and outpatients and for the medical conditions of neoplasms, diseases of the circulatory system and diseases of the digestive system as measured according to ICD classifications C, I and K respectively.

Comparison to existing mapping functions

Our models are compared to the approaches of Franks et al. (2004), Gray et al. (2006) and Sullivan and Ghushchyan (2006) to determine whether their mapping approaches are more or less reliable for a patient dataset. The existing models from the literature are estimated using the published results and algorithms rather than re-estimating the models using our dataset. We take this approach because mapping is used in economic evaluations to estimate the EQ-5D using the SF-36 (or SF-12) when this is the only health status measure that has been included in the trial. Therefore in practical applications the published results and algorithms are used and it is not feasible to re-estimate the model.

Franks et al. (2004) regress the EQ-5D utility score on PCS-12 and MCS-12, squared terms and cross-products using OLS. PCS and MCS are the physical and mental component summary scores estimated using factor analysis and shown to contain most of the information contained in the 8 dimensions of the SF-36 (Ware et al., 1995). In accordance with this approach PCS-12 and MCS-12 are centred on the means used by Franks et al. (2004) and the published coefficients are used to produce predicted EQ-5D utility scores.[2]

Gray et al. (2006) use a response mapping approach that uses a multinomial logit model to estimate the probability that a respondent will choose a particular level for each dimension of the EQ-5D using responses to the 12 items included in the SF-12 (general health, climbing stairs, moderate activities, accomplish less due to physical health, work limitations, accomplish less due to emotional problems, work carefully, pain interference, calm, energy, down-hearted and low, interference with social

[1] The estimation results are not reported here but are available from the authors.

activities). Subsequently predicted EQ-5D level responses for each dimension are generated using Monte Carlo simulation methods and the corresponding EQ-5D utility score for that health state is calculated. We use the Gray et al. (2006) algorithm to predict EQ-5D utility scores.[3]

Sullivan and Ghushchyan (2006) regress the US EQ-5D utility score on PCS-12 and MCS-12, the product of PCS-12 and MCS-12 and sociodemographic variables using OLS, tobit and CLAD. It is not appropriate to use the exact model of Sullivan and Ghushchyan (2006) as they use the US-based EQ-5D values developed by Shaw et al. (2005) rather than the UK-based values developed by Dolan (1997) and further only report models including sociodemographic variables unavailable in our dataset. Instead we have used the tobit and CLAD estimation techniques suggested by Sullivan and Ghushchyan (2006) as outlined above and re-estimated the model using our dataset.

The data

The Health Outcomes Data Repository, HODaR, is a dataset collated by Cardiff Research Consortium. The data is collected from a prospective survey of inpatients and outpatients at Cardiff and Vale NHS Hospitals Trust, which is a large University hospital in South Wales, UK. The survey is linked to existing routine hospital health data to provide a dataset with sociodemographic, health related quality of life and ICD classification data[2]. The survey includes all subjects aged 18 years or older and excludes individuals who are known to have died. The survey also excludes people with a primary diagnosis on admission of a psychological illness or learning disability. As well as information on inpatients, the survey includes outpatient clinics on a rotational basis where all patients within the selected clinic are surveyed. The response rate in HODaR prior to October 2003 was 36% and subsequently strategies have been implemented to improve response rates.

The inpatient sample has 31,236 eligible observations across 27,620 individuals from August 2002 to November 2004, and of these there are 25,783 complete responses across 23,179 individuals for SF-36 and EQ-5D questions and hence this is the sample used here. The outpatient sample has 9,081 eligible observations across 8,610

[2] Ssee Currie et al. (2005) for further details on HODaR.

individuals collected from June 2002 to November 2004, and of these there are 7,465 complete responses across 7,122 individuals.

Results

Table 1 provides descriptive statistics on health status. The inpatient and outpatient samples in the HODaR dataset demonstrate substantial health problems according to the EQ-5D, the SF-36 dimension scores and the SF-12 summary scores in comparison to UK population norms. Health appears similar between inpatients and outpatients. In comparison to the inpatient sample the outpatient sample has a larger proportion of females and a lower mean age.

Table 1 Descriptive data for the inpatient and outpatient samples

	Inpatients				Outpatients				UK population norms[4]	
	Mean	SD	Median	Inter-quartile range	Mean	SD	Median	Inter-quartile range	Mean	SD
EQ-5D index	0.68	0.31	0.73	0.413	0.69	0.31	0.73	0.38	0.86	0.23
SF-36 dimension scores										
Physical functioning	58.90	33.53	65.00	60.00	62.29	33.39	70.00	60.00	88.40	17.98
Social functioning	63.43	33.16	66.67	66.67	66.35	32.02	77.78	55.56	88.01	19.58
Role physical	28.74	41.90	0.00	75.00	34.21	44.11	0.00	100.00	85.82	29.93
Role-emotional	51.14	47.14	66.67	100.00	54.32	46.99	66.67	100.00	82.93	31.76
Mental health	69.54	23.13	76.00	32.00	69.58	22.54	76.00	32.00	73.77	17.24
Vitality	45.36	25.73	45.00	40.00	45.60	25.37	45.00	40.00	61.13	19.67
Bodily pain	58.13	28.68	55.56	44.44	58.86	28.84	55.56	55.56	81.49	21.69
General health	52.80	26.28	52.00	47.00	53.29	25.91	52.00	47.00	73.52	19.90
SF-12 summary scores										
Physical component score	38.25	12.18	36.68	21.49	39.51	12.34	38.47	22.50	50.00	10.00
Mental component score	44.85	11.69	46.21	19.38	45.03	11.45	46.92	19.07	50.00	10.00
Mean age	58.14				55.55					
Female	52%				61%					
N	25,783				7,465					

Inpatients

Table 2 shows the results of the regression analyses using dimensions, squared terms and interaction terms for the inpatient dataset. The results show that all dimensions are always significant with the exception of role physical, vitality and role emotional and are positive with the exception of role physical and vitality. The results indicate that the squared terms for physical functioning, bodily pain, social functioning and

mental health are always significant and negative and many interaction terms are also significant with mixed signs. Statistical measures reported in Table 2 of within, between and overall R-squared, root mean squared error, rho and Wald chi-squared indicate that models (2) and (3) perform better than model (1). Table 3 reports mean error, mean absolute error (MAE) and mean squared error (MSE) of predicted compared to actual utility scores by EQ-5D utility range for all models estimated in Table 2. Table 3 indicates that the estimation techniques of tobit and CLAD do not clearly improve the accuracy of the generated predictions as MAE and MSE are not reduced. Model (3) estimated using random effects GLS and random effects tobit have the most accurate predictions as indicated by MAE and MSE.

Figure 1 shows the observed and predicted EQ-5D utility scores, ordered by observed tariff value of the EQ-5D state. The sample used is the inpatient dataset and the predictions are generated using model (3) estimated using random effects GLS. Figure 1 and MAE and MSE reported in table 3 suggest that the model predicts well for milder health states, but overpredicts the value of more severe EQ-5D states. All models estimated in Table 2 suffer from the same problem.

Table 2 Prediction models for inpatients using dimensions, squared terms and interaction terms

	Random effects GLS			Tobit	CLAD
	(1)	(2)	(3)	(4)	(5)
Dimensions					
Physical functioning (PF)	0.332*	0.548*	0.559*	0.559*	0.663*
Role physical (RP)	-0.060*	-0.021	-0.146*	-0.146*	-0.475*
Bodily pain (BP)	0.303*	0.747*	0.715*	0.713*	0.733*
General health (GH)	0.169*	0.322*	0.407*	0.407*	0.325*
Vitality (VIT)	-0.039*	0.007	0.017	0.017	-0.142*
Social functioning (SF)	0.115*	0.256*	0.293*	0.293*	0.525*
Role-emotional (RE)	0.010*	0.014	0.067*	0.067*	-0.024
Mental health (MH)	0.237*	0.577*	0.483*	0.483*	0.527*
Dimensions squared					
Physical functioning (PF)		-0.250*	-0.227*	-0.227*	-0.082*
Role physical (RP)		0.043*	0.001	0.001	-0.056*
Bodily pain (BP)		-0.378*	-0.330*	-0.329*	-0.171*
General health (GH)		-0.137*	0.032	0.031	0.167*
Vitality (VIT)		-0.014	-0.012	-0.012	0.063
Social functioning (SF)		-0.179*	-0.163*	-0.163*	-0.182*
Role-emotional (RE)		0.017	0.034	0.034	0.058*
Mental health (MH)		-0.321*	-0.242*	-0.242*	-0.152*
Interaction terms					
PF x RP			0.022	0.022	0.185*
PF x BP			-0.032	-0.031	-0.192*
PF x GH			0.073	0.073	-0.009
PF x VIT			-0.132*	-0.132*	-0.078
PF x SF			-0.023	-0.023	-0.246*
PF x RE			0.047*	0.047*	0.045*
PF x MH			-0.014	-0.013	-0.054
RP x BP			0.019	0.019	0.097*
RP x GH			0.068*	0.068*	0.215*
RP x VIT			0.050	0.049	0.031
RP x SF			0.067*	0.067*	0.108*
RP x RE			-0.012	-0.012	0.013
RP x MH			0.022	0.022	0.154*
BP x GH			-0.217*	-0.217*	-0.208*
BP x VIT			-0.002	-0.002	0.120*
BP x SF			0.055	0.055	-0.070*
BP x RE			-0.038	-0.038	0.039*
BP x MH			0.131*	0.131*	-0.075
GH x VIT			-0.066	-0.066	-0.200*
GH x SF			-0.157*	-0.158*	-0.144*
GH x RE			-0.033	-0.033	-0.019
GH x MH			-0.084	-0.084	-0.114*
VIT x SF			0.143*	0.143*	0.174*
VIT x RE			-0.020	-0.019	-0.021
VIT x MH			0.023	0.022	0.095
SF x RE			-0.023	-0.023	-0.024
SF x MH			-0.065	-0.065	-0.133*
RE x MH			-0.048	-0.048	-0.035
Constant	0.0071	-0.2493*	-0.256*	-0.256*	-0.289*
Within R-squared	0.18	0.21	0.22	-	-
Between R-squared	0.67	0.70	0.71	-	-
Overall R-squared	0.67	0.70	0.71	-	-
Root MSE	0.15	0.15	0.15	-	-
Rho	0.28	0.24	0.24		
Wald Chi-squared	48380.12	56129.39	57195.96		

Note: * significant at 1%

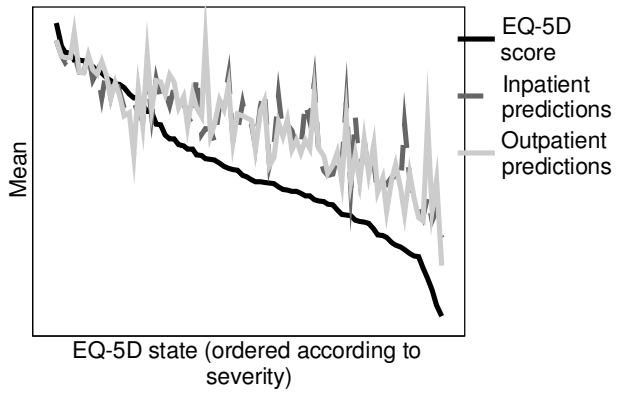

Figure 1 Observed and predicted EQ-5D scores: Inpatients and outpatients random effects GLS model

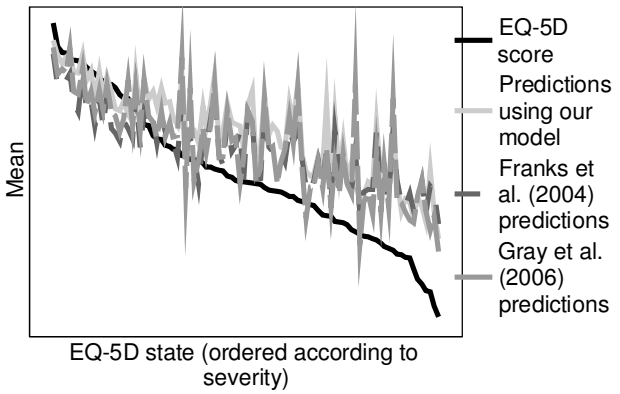

Figure 2 Observed and predicted EQ-5D scores: Comparison to existing mapping functions

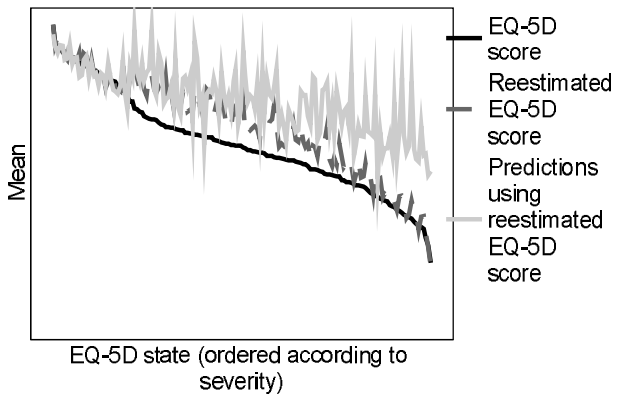

Figure 3 Observed and predicted EQ-5D scores: Using EQ-5D tariff re-estimated without an N3 term using the MVH data

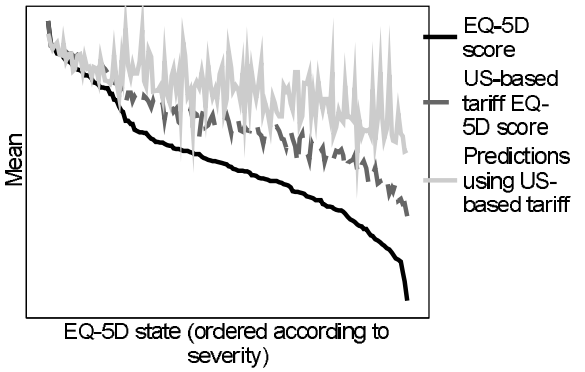

Figure 4 Observed and predicted EQ-5D scores: Using the US-based EQ-5D tariff

Table 3 Mean error, mean absolute error and mean squared error of predicted compared to actual utility scores by EQ-5D utility range for random effects GLS models, random effects tobit models, CLAD model, Franks et al. model and Gray et al. model

EQ-5D utility score	Random effects GLS (1)	(2)	(3)	Random effects tobit (4)	CLAD (5)	Franks et al.	Gray et al.
Mean error							
<0	-0.340	-0.266	-0.260	-0.260	-0.269	-0.252	-0.213
0-0.249	-0.241	-0.219	-0.217	-0.218	-0.237	-0.144	-0.144
0.25-0.499	-0.191	-0.189	-0.191	-0.191	-0.219	-0.064	-0.081
0.5-0.6.99	0.098	0.072	0.070	0.070	0.052	0.201	0.135
0.7-0.799	-0.004	-0.024	-0.024	-0.024	-0.044	0.095	0.056
0.8-0.899	0.041	0.034	0.034	0.034	0.004	0.167	0.114
0.9-1.0	0.064	0.086	0.085	0.085	0.025	0.154	0.123
Full index	-0.001	0.000	0.000	-0.001	-0.031	0.101	0.059
Mean absolute error							
<0	0.340	0.271	0.266	0.266	0.278	0.254	0.272
0-0.249	0.244	0.238	0.238	0.238	0.260	0.175	0.278
0.25-0.499	0.202	0.215	0.219	0.219	0.247	0.136	0.282
0.5-0.699	0.138	0.131	0.130	0.130	0.122	0.211	0.210
0.7-0.799	0.105	0.098	0.095	0.095	0.102	0.147	0.145
0.8-0.899	0.106	0.088	0.085	0.085	0.092	0.183	0.172
0.9-1.0	0.086	0.086	0.085	0.085	0.092	0.154	0.123
Full index	0.138	0.129	0.127	0.127	0.133	0.178	0.186
Mean squared error							
<0	0.132	0.099	0.097	0.097	0.110	0.082	0.135
0-0.249	0.078	0.080	0.080	0.080	0.095	0.048	0.123
0.25-0.499	0.061	0.066	0.067	0.067	0.085	0.032	0.102
0.5-0.699	0.028	0.028	0.028	0.028	0.026	0.060	0.094
0.7-0.799	0.017	0.015	0.014	0.014	0.018	0.034	0.052
0.8-0.899	0.019	0.015	0.014	0.014	0.016	0.051	0.065
0.9-1.0	0.015	0.013	0.013	0.013	0.013	0.037	0.042
Full index	0.033	0.030	0.030	0.030	0.033	0.048	0.076

Inpatients and outpatients

Figure 1 shows the observed and predicted EQ-5D scores for inpatients and outpatients. The predictions are generated using model (3) estimated using random effects GLS. The mapping relationship follows the same pattern across inpatient and outpatient settings and both overpredict for more severe EQ-5D states. Wald test statistics calculated to determine whether the estimated coefficients for inpatients are equal to the estimated coefficients for outpatients for models with exactly the same specification indicate that the estimated coefficients are not equal and hence the models are not robust to different samples. However, differences in predictions are small with mean absolute difference at the state level of 0.069 and mean squared difference of 0.012. Wald test statistics were also calculated for subsets of the inpatient sample according to medical condition for the ICD classifications with the largest number of observations in the dataset, which are the medical conditions of neoplasms (n=2,574), diseases of the circulatory system (n=3,522) and diseases of the digestive system (n=3,114) as measured according to ICD classifications C, I and K respectively. The test statistics again indicate that the estimated coefficients are not equal and hence are not robust across subsets of the inpatient sample according to medical condition, but differences in predictions are small with highest mean absolute difference at the state level of 0.054 and highest mean squared error of 0.005.

Comparison to existing mapping

Figure 3 shows observed and predicted EQ-5D utility scores for model (3) and for the models of Franks et al. (2004) and Gray et al. (2006). The mapping relationship is similar across all approaches and they all overpredict for more severe EQ-5D states. Table 3 shows mean error, mean absolute error and mean square error of predicted compared to actual utility scores by EQ-5D utility range for the Franks et al. (2004) model and the Gray et al. (2006) model. As indicated by Figure 3, the errors are higher for more severe health states for all models. Our model performs better than the existing models as reported by mean error, mean absolute error and mean square error.

Re-estimation of the EQ-5D

One hypothesis is that the predictions may be poor for more severe EQ-5D states because they all have at least one dimension at the most severe level and the EQ-5D

model uses an 'N3' term, a dummy variable for states with at least one dimension at the most severe level. The 'N3' term was used in the original UK modelling (Dolan, 1997), but has not been included in all the models of other EQ-5D valuation studies (see for example the US valuation study, Shaw et al. (2005)). The inclusion of the N3 term may be a reason why the utility score is overpredicted for the more severe states which have at least one dimension at the most severe level. We re-estimated the EQ-5D tariff without the N3 term using the same data and methods as Dolan (1997). The re-estimated tariff and the original Dolan (1997) tariff produce similar scores for mild and very severe health states but deviate for more moderate health states, with mean difference in tariff values at the state level of 0.134 and mean squared difference of 0.026. Figure 4 plots the observed and predicted EQ-5D utility scores using a re-estimated version of the EQ-5D and plots this alongside the UK-based values developed by Dolan (1997). The predicted values for the re-estimated EQ-5D scores still overpredict for more severe states, but not as much as previously, with MAE of 0.106 and MSE of 0.021 in comparison to MAE of 0.127 and MSE of 0.030 for the predictions based on Dolan (1997) tariff. However the PITS state is overpredicted by 0.63 for the re-estimated EQ-5D scores and 0.61 for the predictions based on Dolan (1997) tariff.

US-based EQ-5D

The re-estimated UK-based tariff and Dolan (1997) tariff produce similar scores for mild and very severe health states and hence the preferences regarding more severe health states may be a property of the dataset rather than the estimation technique used for the valuation. The US-based EQ-5D tariff has a smaller range from 1 to -0.11 and hence has higher scores for very severe states, suggesting that the mapping relationship between the US-based EQ-5D index and the SF-36 may not suffer from overprediction for more severe health states. Figure 5 plots the observed and predicted EQ-5D scores using the US-based EQ-5D values developed by Shaw et al. (2005) alongside the UK-based values developed by Dolan (1997). This demonstrates that the predicted values for the US-based EQ-5D values still overpredict for more severe states, but the estimates are more reliable than those plotted in figure 3 with MAE of 0.110 and MSE of 0.022 in comparison to MAE of 0.127 and MSE of 0.030 for the predictions based on UK Dolan (1997) tariff. The PITS state is overpredicted by 0.38

for the US-based EQ-5D values and 0.86 for the predictions based on UK Dolan (1997) tariff.

Discussion

The patient dataset used here is much better than general population datasets in terms of diversity of conditions and severity of health. Our results suggest that the mapping relationship between the EQ-5D index and the SF-36 for a large and varied UK patient dataset is reliable and accurate across inpatient and outpatient settings and medical conditions. However, our results indicate that the mapping relationship is not accurate and reliable for more severe EQ-5D health states. The inclusion of squared and interaction terms in the models improves diagnostics, mean error, MAE and MSE, suggesting that the mapping relationship is non-linear and dimensions are additive. The mapping approach used here is compared to the existing approaches of Franks et al. (2004) and Gray et al. (2006) and all suffer from overprediction for more severe EQ-5D health states. The added complexity of the response mapping approach used by Gray et al. (2006) does not seem to improve the predictability for all health states in comparison to our approach.

One potential reason for the overprediction for more severe health states are the floor effects of the SF-36. We have tried to account for these floor effects by using squared terms and interaction terms in our model, but, as the figures illustrate, this does not resolve the problem. We also tried re-estimating the EQ-5D utility tariff using the original dataset used by Dolan (1997) but omitting the N3 term Although Figure 4 demonstrates better predictions for more severe health states, the problem of overprediction is still evident. Indeed, if the preferences regarding more severe health states is a property of the dataset rather than the estimation technique, then the valuation produced here will still demonstrate the same properties. We also estimated our model using the US-based EQ-5D values, and although Figure 5 demonstrates better predictions for more severe health states, again the problem of overprediction is still evident. The importance of the problem of overprediction in economic evaluations is difficult to measure, since it depends on the patient group and the effect of treatments. Ara and Brazier (2008) predict mean cohort EQ-5D utility values using mean cohort scores for the dimensions of the SF-36 from published datasets. They find mean errors of 0.285 and 0.158 in prediction for the 5 out of 63 cohorts in an out

of sample dataset with mean EQ-5D utility value below 0.175 and between 0.175 and 0.35 respectively. The impact at the group level may be less important since few patients have EQ-5D utility values below 0.5, and the inpatient and outpatient datasets used here each have 17% of observations with an EQ-5D utility value below 0.5, suggesting that not many observations will be affected by the overprediction for more severe states that is presented here. Therefore for most studies this may not matter, only where many patients have EQ-5D utility values below 0.5.

The results suggest that there are differences in the EQ-5D and SF-36 health status measures for more severe health states which make mapping unreliable for these states. Another finding is that the vitality, role physical and role-emotional dimensions of the SF-36 did not significantly effect the EQ-5D index, hence interventions aimed at improving these dimensions will not be reflected in the mapping model. However, these domains were found to be important to members of the public in the valuation of the SF-6D (Brazier et al., 2002). Mapping is increasingly being used between condition specific measures and generic measures of health (refer to Brazier et al. (2007)). However, the lack of overlap in the dimensions covered by many condition specific measures and EQ-5D limit the usefulness of this approach as these problems may be worsened if the health domains included in the measures are different.

Conclusions

Mapping enables utility scores to be estimated in trials where a non-preference based health status measure has been used but no generic preference-based measure. Our results suggest that approaches mapping the SF-36 onto the EQ-5D are robust across setting and medical condition but overpredict for more severe EQ-5D states. Our results raise doubt over the suitability of mapping for patient datasets which have a proportion of subjects with poorer health or where dimensions are not represented in the target measure. Potential policy implications are that mapping the SF-36 onto the EQ-5D can be useful, but may not be suitable for all populations.

References

Ara, R., Brazier, J., 2008. Deriving an Algorithm to Convert the Eight Mean SF-36 Dimension Scores into a Mean EQ-5D Preference-Based Score from Published Studies (Where Patient Level Data Are Not Available). Value in Health, forthcoming.

Brazier JE, Harper R, Thomas K, Jones N, Underwood T, 1998. Deriving a preference based single index measure from the SF-36 J.Clinical Epidemiology 51 (11):1115-1129

Brazier, J., Roberts, J., Deverill, M., 2002. The estimation of a preference-based measure of health from the SF-36. Journal of Health Economics 21, 271-292.

Brazier, J., Roberts, J., Tsuchiya, A., Busschbach, J., 2004. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Economics 13, 873-884.

Brazier, J., Yang, Y., Tsuchiya, A., 2007. Review of methods for mapping between condition specific measures onto generic measures of health. Report prepared for the Office of Health Economics, May 2007.

Chay, K. Y., Powell, J. L., 2001. Semiparametric Censored Regression Models. The Journal of Economic Perspectives 15, 29-42.

Currie, C. J., McEwan, P., Peters J. R., Patel, T.C., Dixon, S., 2005. The Routine Collation of Health Outcomes Data from Hospital Treated Subjects in the Health Outcomes Data Repository (HODaR): Descriptive Analysis from the First 20,000 Subjects. Value in Health 8, 581-590.

Dolan, P., 1997. Modeling Valuations for EuroQol Health States. Medical Care 35, 1095-1108.

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., Denton, M., Boyle, M., 2002. Multiattribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. Medical Care 40, 113-128

Franks, P., Lubetkin, E. I., Gold Marthe R, Tancredi, D. J., Haomiao, J., 2004. Mapping the SF-12 to the EuroQol EQ-5D Index in a National US Sample. Medical Decision Making 24, 247-254.

Gray, A. M., Rivero-Arias, O., Clarke, P. M., 2006. Estimating the Association between SF-12 Responses and EQ-5D Utility Values by Response Mapping. Medical Decision Making 26, 18-29.

Greene, W.H., 2000. Econometric Analysis. New Jersey: Prentice Hall.

Jenkinson, C., Layte, R., Wright, L., Coulter, A., (1996) The UK SF-36: An analysis and interpretation manual. Oxford: Health Services Research Unit 1996.

Kind, P., Hardman, G., Macran, S., 1999. UK Population Norms for EQ-5D. Centre for Health Economics Discussion Paper 172, University of York, York.

Lawrence, W. F., Fleishman, J. A., 2004. Predicting EuroQoL EQ-5D Preference Scores from the SF-12 Health Survey in a Nationally Representative Sample. Medical Decision Making 24, 160-169.

McCabe, C., Stevens, K., Roberts, J., Brazier, J., 2005. Health state values for the HUI 2 descriptive system: Results from a UK survey. Health Economics 14, 231-244.

NICE (2008) Guide to the methods of technology appraisal. NICE, London. http://www.nice.org.uk/aboutnice/howwework/devnicetech/technologyappraisalprocessguides/guidetothemethodsoftechnologyappraisal.jsp

Shaw, J. W., Johnson, J. A., Coons, S. J., 2005. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. Medical Care 43, 203-220.

Sullivan, P. W., Ghushchyan, V., 2006. Mapping the EQ-5D Index from the SF-12: US General Population Preferences in a Nationally Representative Sample. Medical Decision Making 26, 401-409.

Ware, J. E., Kolinski, M., Keller SD. 1995. How to score the SF-12 physical and mental health summaries: a user's Manual. Boston: The Health Institute, New England Medical Centre, Boston, MA.

[1] CLAD was performed in STATA using programs written for Chay and Powell (2001).

[2] Franks et al. (2004) estimate other models but these are not analysed here as these models use demographic variables not available in the dataset used here. Furthermore Franks et al. found that more complex models explained only minimally additional variance. Lawrence and Fleishman (2004) use similar variables and estimation techniques to Franks et al (2004) in order to predict EQ-5D scores from the SF-12 and hence their model is not analysed here separately.

[3] The Gray et al. (2006) algorithm is available from the HERC website http://www.herc.ox.ac.uk/downloads/supp_pub/sf12eq5d

[4] EQ-5D population norms obtained from Kind et al. (1999) for the Measurement and Valuation of Health survey and SF-36 population norms obtained from Jenkinson et al. (1996) for the Oxford Healthy Life Survey.