Published paper

Brazier, J.E., Fukuhara, S., Roberts, J., Kharroubi, S., Yamamoto, Y., Ikeda, S., Doherty, J., Kurokawa, K. (2009) *Estimating a preference-based index from the Japanese SF-36*, Journal of Clinical Epidemiology, 62 (12), pp, 1323-1331
http://dx.doi.org/10.1016/j.jclinepi.2009.01.022

Title: Estimating a preference-based index from the Japanese SF-36

Abstract

Objective: The main objective was to estimate a preference-based SF-6D index from the SF-36 for Japan and compare to the UK results.

Study design and setting: The SF-6D was translated into Japanese. 249 health states defined by this version of the SF-6D were then valued by a representative sample of 600 members of the Japanese general population using standard gamble. These health state values were modelled using classical parametric random effects methods with individual level data and OLS on mean health state values, together with a new nonparametric approach using Bayesian methods of estimation.

Results: All parametric models estimated on Japanese data were found to perform less well than their UK counterparts in terms of poorer goodness of fit, more inconsistencies, larger prediction errors and bias, and evidence of systematic bias in the predictions. Non-parametric models produce a substantial improvement in out of sample predictions. The physical, role and social dimensions have relatively larger decrements than pain and mental health compared to the UK.

Conclusion: The differences between Japan and UK valuation of the SF-6D make it important to use the Japanese valuation data set estimated using the non-parametric Bayesian technique presented in this paper.

Key words: Preference-based measures, QALYs, SF-6D, Bayesian modelling, Japan, cross-cultural comparisons

Running title: Estimating a preference-based index from the Japanese SF-36

Abstract word count: 195

BACKGROUND

Preference-based measures of health standardized, multidimensional systems for classifying health states that generate scores on a scale where full health is one and zero is equivalent to death. [1] The use of preference-based measures of health has grown considerably over the last decade with the increasing use of economic evaluation to inform health policy.

The SF-36 is one of the most widely used generic health survey instruments. [2] There is substantial evidence of the ability of the eight dimension scores of the SF-36 to describe the health differences between patient groups and to track changes over time. However, the methods for scoring the SF-36 are not preference-based and do not provide a means to derive QALYs. The SF-6D was developed as a practical tool for obtaining a preference-based index from SF-36 data. [3]

Given the high cost of undertaking valuation surveys, early work using preference-based measures has tended to use the valuation results from just one or two countries, which for the EQ-5D [4] and SF-6D has been the UK and for the HUI [5] has been Canada. However, significant differences have been found between countries, for example values obtained in the UK EQ-5D surveys and those from Japan [6] and USA [7] and between Canadian values for the HUI2 and those obtained in the UK [8]. This paper presents results from undertaking a valuation of the SF-6D in Japan and compares it to results for the UK.

An important problem for preference-based measures has been their size and the consequent need to model health state values from valuation of a subset of possible states. Classical modelling with random effects estimated using generalised least squares (GLS) has met with some success, but has encountered major challenges due to the nature of the distribution of health state values which is generally skewed, truncated, hierarchical and non-continuous [10]. An alternative non-parametric model has been developed to estimate health state values by Bayesian methods [11]. The nonparametric method was applied to the UK valuation data set and was found to achieve a better predictive model and so has replaced the original UK preference-based scoring algorithm. For this reason, it was felt to be important to apply this approach to the valuation data from Japan and compare it to the classical random effects model.

METHODS

The methods closely follows those used the UK to derive the SF-6D preference-based measure of health from the SF-36. [3] The first stage in the UK study was to develop the SF-6D, a health state classification with six dimensions (physical functioning, role limitation, social functioning, bodily pain, mental health and vitality). The six dimensions have between four and six that define a total of 18000 health states. A selection of 249 SF-6D health states were valued by a representative sample of the UK general population using the standard gamble (SG) valuation technique. The resultant valuation data were modelled to generate a UK algorithm for valuing all SF-6D states. These methods required some modification for use in Japan.

Translation of the SF-6D

The SF-6D uses items of the SF-36, so it was possible to use an existing translation of the SF-36 undertaken by Fukuhara and colleagues [10]. The only item that required modification was the role dimension. This dimension was changed to ask the frequency of role limitation induced by general health states, for both role-emotional and role-physical showed same pattern in the SF-36 survey in Japan. Previous work with the SF-36 in Japan had found that the emotional dimension did not perform in the same way as in the west and reflected a different cultural perception of health [11]. The second modification was to use the item from version 2 of SF-36 since it has 5 levels per item rather than the two in version 1 and so provides a better spread of responses.

The resultant Japanese SF-6D is a multi-level six dimensional health state classification and this is presented in English on Table 1.

Valuation survey

A sub-set of health states defined by the SF-6D were valued by a representative sample of the general public in Japan. Each respondent was asked to rank and value seven states using a variant of the SG technique. The key design components of the survey were the selection of health states, the sampling of respondents and the conduct of the interviews.

Selection of health states

It was not possible to use the same health state sample as used in the UK study, since the Japanese version of the SF-6D had been modified. The sample of states was selected using an orthogonal design (by applying the orthoplan procedure of SPSS), which generates the states needed to estimate an additive model supplemented by an additional sample obtained at random. Two hundred and forty states were generated in this way. On inspection of the states, it was found that there were 10 infeasible states, such as ‘extreme pain’ alongside ‘no problems’ in the social and/or role dimensions. For these states, the levels of role and social were changed to level 3.

Respondents valued 6 intermediate states and the worst SF-6D health state. The 6 intermediate states for each interview were selected using a block system. The 240 states were divided into 6 severity categories using a total score based on the sum of dimension level scores. Five blocks of 48 were created by taking random samples of 8 states from each severity category. Each interviewer received one or two blocks of 48 cards. For each block the interviewers randomly sorted the states into 8 sets of 6 states for the interviews. This block system ensured that each person valued a range of health states across the space defined by the SF-6D and it was also designed to maximise the chance that each of the 240 cards would be valued by an equal number of respondents.

Selection of respondents

The aim was to ensure the sample reflected the variability in the population of characteristics such as age, socio-economic status and level of education. This was achieved by a two-stage cluster random design where 50 districts were randomly selected across Japan. Quota sampling was then applied using variables for age, education, income and type of occupation. The target was 600 respondents since this was found to be sufficient in the UK study.

The Interviews

Trained interviewers conducted the survey, interviewing respondents in their own home. The interview began with the respondent being asked to complete a short self-completion questionnaire about his or her own state of health, that included the SF-6D. This familiarised the respondent with the idea of describing health in terms of the SF-6D. Other self-report questionnaires used in the survey were the SF-36, SF-8, EQ-5D and EQ-5D-VAS.

The respondent was asked to rank a set of nine cards: one for each of the intermediate health states they would have to value (6), along with the best state defined by the SF-6D, the worst state and being dead. Next respondents were asked to value health states using a variant of the SG that uses a Chance Board to display probabilities. [12].

In the SG valuation task respondents were asked to value each intermediate SF-6D health states against the uncertain prospect of ending up in the best or worst SF-6D health state. For calculating QALYs it is necessary to transform the results onto a scale where 1 is full health and 0 is equivalent to death. To do this the worst SF-6D state was valued in a SG task where full health and dead was the uncertain prospect. This task yields a value 'p' for the worst state and this was used to transform the value of the six intermediate SF-6D health state valuations (SG) using the formula: SGADJ =SG+(1-SG)*P. The transformed value SGADJ is used in all subsequent analyses.

Modelling

There were two modelling approaches applied to the data. The classical parametric approach used by Brazier et al [3] and a more realistic and flexible nonparametric model [11].

A general model for health state valuations can be described by:

$$y_{ij} = f(\mathbf{x}_{ij}, \alpha_j) + \varepsilon_{ij}, \qquad (1)$$

where, for i = 1, 2, …, n_j and j = 1,2, …, m, $\mathbf{x}_{ij}$ is the i^{th} health state valued by respondent j and the dependent variable y_{ij} is the adjusted SG score given by respondent j for that health state. The general model has two sets of independent, zero-mean, random effect terms: ε_{ij} is a random error term associated with each observation and α_j is a term to allow for individual characteristics of respondent j.

The interpretation of $f(\mathbf{x}_{ij}, \alpha_j)$ is as the true indifference SG value that respondent *j* has for health state $\mathbf{x}_{ij}$. The objective is to obtain a health state utility measure for the population as a whole, and this is generally taken to be the mean of the respondent-level health state values across the population. In order to account for different populations, it is possible to model α_j in terms of respondent-level covariates such as age, gender or socio-economic factors, but the principal objective of this study was to estimate a health state preference function for the population as a whole.

The parametric approach

Kharoubi et al [9] specify the following model for respondent *j*'s health state utility:

$$f(\mathbf{x}_{ij}, \alpha_j) = \mu + \boldsymbol{\theta}'\mathbf{I}(\mathbf{x}_{ij}) + \alpha_j, \qquad (2)$$

where μ and $\boldsymbol{\theta}$ denote unknown parameters, $\mathbf{I}(\mathbf{x}_{ij})$ is a vector of dummy explanatory variables. In the simplest, no-interactions, case of this model, $\mathbf{I}(\mathbf{x}_{ij})$ is a vector of terms $I_{\delta\lambda}(\mathbf{x}_{ij})$ for each level $\lambda > 1$ of dimension δ of the SF-6D. For example, $I_{32}(\mathbf{x}_{ij})$ denotes dimension $\delta = 3$ (social functioning), level $\lambda = 2$ (health limits social activities a little of the time). For any given health state $\mathbf{x}_{ij}$, $I_{\delta\lambda}(\mathbf{x}_{ij})$ is defined as:

$I_{\delta\lambda}(\mathbf{x}_{ij})$ = 1 if, for state $\mathbf{x}_{ij}$, dimension δ is at level λ.

$I_{\delta\lambda}(\mathbf{x}_{ij})$ = 0 if, for state $\mathbf{x}_{ij}$, dimension δ is not at level λ

In all, there are 25 of these terms, with level $\lambda = 1$ acting as a baseline for each dimension. Hence the intercept parameter μ represents the health state utility value for state 111111, and summing the coefficients $\theta_{\delta\lambda}$ of the 'on' dummies derives the value of any other state.

More generally, $\mathbf{I}(\mathbf{x}_{ij})$ can include additional dummy variables to account for interactions between the levels of different dimensions, and the model selected by Brazier et al [2] included one such interaction term, MOST, which takes the value of 1 if any dimension in the health state is at one of the most severe levels[1], and 0 otherwise.

Estimation of this random effects model is via generalised least squares or maximum likelihood. Since α_j has zero mean, the population health state utility for state **x** in this model is simply $\mu + \boldsymbol{\theta}'\mathbf{I}(\mathbf{x})$. The other modelling approach has been to model mean health state values by OLS.

The nonparametric approach

Kharroubi and colleagues [9] built a new Bayesian statistical nonparametric model to describe the intrinsic characteristics of individual health state valuation data that is argued to be more theoretically appropriate than previous parametric models. For respondent *j*, the health state utility of state $\mathbf{x}_{ij}$ is

$$f(\mathbf{x}_{ij}, \alpha_j) = 1 - \exp(\alpha_j)\{1 - u(\mathbf{x}_{ij})\} . \qquad (3)$$

Note that the individual respondent term α_j enters multiplicatively rather than additively as in (2). The term *u*(**x**) is the *median* health state utility of health state **x**.[2] It is treated as an unknown function and in a nonparametric framework it therefore becomes a random variable. The model for *u*(**x**) is

[1] Most severe is defined as levels 4 to 6 for physical functioning, levels 3 and 4 for role limitation, 4 and 5 for social functioning, mental health and vitality, and 5 and 6 for pain.

[2] In the Kharroubi et al [11] model, the distribution of α_j is normal, so it has zero median as well as zero mean, and the median of $\exp(\alpha_j)$ is therefore 1.

$$u(\mathbf{x}) \sim \mathrm{N}(\gamma + \boldsymbol{\beta}'\mathbf{x}, \sigma^2), \qquad (4)$$

and furthermore the values of $u(\mathbf{x})$ and $u(\mathbf{x}')$ for two different states $\mathbf{x}$ and $\mathbf{x}'$ have a correlation $c(\mathbf{x}, \mathbf{x}')$ which decreases as the distance between $\mathbf{x}$ and $\mathbf{x}'$ increases. The effect of this is to assert that if $\mathbf{x}$ and $\mathbf{x}'$ describe very similar health states (in the sense that their levels are the same or close in all dimensions) their utilities will be approximately the same, and so the preference function varies smoothly as the health state changes.

Note that the mean health state utility in (3) is

$$\bar{u}(\mathbf{x}) = 1 - \bar{\alpha}\{1 - u(\mathbf{x})\} \ ,$$

where $\bar{\alpha}$ is the mean value of $\exp(\alpha)$ over the whole population. This will not in general be 1, and so the population (mean) health state utility is not the same as the median health state utility $u(\mathbf{x})$. More details of the nonparametric modelling and evaluation of $\bar{\alpha}$ are given elsewhere [11].

Comparisons of models

The parametric models have been judged in classical ways in order to allow comparison with the original UK study: the adjusted R^2 (i.e. explanatory power); inconsistencies in the coefficient estimates (i.e. the main explanatory dummy variables are expected to be negative and increasing in absolute size and an inconsistent result occurs where a coefficient on the main effects dummies decreases in absolute size with a worse level); the mean absolute prediction error when the model is used to predict the actual mean values of each of the 241 states valued (MAE) ; the number of prediction errors that are larger than 0.05 or 0.10 in absolute value: a t-test of the null hypothesis that the mean prediction error is zero (this is a test for bias in the predictions); Jarque-Bera test (JBPRED) for normality of the prediction errors; Ljung-Box (LB) test for autocorrelation in the prediction errors. The LB test is for systematic variation in the prediction errors, signifying misspecification of the underlying model.

The nonparametric model is compared to the parametric model in a number of ways. The first is a plot of sample mean health state values alongside predicted mean values and residuals. A better test of the validity of the model is to investigate its ability to predict the values for states that have not been used in the estimation. Data relating to 12 health states were removed from the estimation data, and the models fitted on data for the remaining 237 states. These 12 states were selected to be well spread over the health state space. For both the parametric and non-parametric model results are compared in terms their ability to predict 12 out of sample states, including the predicted mean health state values and their estimated standard deviations and mean root mean square error across the 12 states. Finally Q-Q plots of standardised predictive errors for the out of sample mean health states values will be presented. The RE model has been selected as the parametric comparator since its specification is closest to the non-parametric model and because it is the most theoretically correct of the parametric models since it takes account of between and within respondent variation.

THE DATA

The total number of interviews conducted in the Japanese survey was 600. To ensure comparability with the UK data set, the same exclusion criteria have been used: respondents have been excluded who gave all states the same value (on the grounds that they did not understand the task) or where they did not value the worst state (since it is not possible to appropriately adjust their values). Applying these criteria resulted in a total of 135 exclusions. The main analyses are based on the responses from the remaining 465 respondents.

The characteristics of those respondents included in the analysis are compared to those of the excluded cases in Table 2. There is little difference in the characteristics of cases included and those excluded in terms of their mean age, sex, marital status and proportion in full time employment. The main difference is that excluded cases had a higher proportion with a poor understanding of the task, which is not surprising. The characteristics of included respondents are comparable to those from the 2005 Japanese census for 18 to 75 year olds: mean age 46 (vs. 48), females 52% (vs. 54%) and employed was 56% (vs. 51%), although there was some difference in the proportion who were married 67 (vs. 75). [13]

A further 185 health state values were missing, leaving a total of 3070 health state valuations. There were on average 12 valuations per state (excluding PITS which is valued by every respondent). This is slightly lower than the 14 per state achieved in the UK. The distribution of individual values (SG) is similar in both countries and highly skewed and bi-modal. Mean values from the Japanese survey range from 0.29 for worst state to 0.87 for 122122 with large standard deviations. Median values usually exceed the mean health value reflecting the negative skewness in the data.

In Japan the mean value for the worst state is 0.29 (SD 0.35), with only 20 respondents rating PITS as equal to being dead and 39 as worse than being dead (out of 465). The UK mean value was 0.21 (sdSD 0.43) with 73 respondents rating this state as equal to being dead and 165 as worse (out of 611).

RESULTS

Parametric model

Results are presented for Japan for the OLS mean and RE individual level models with and without the intercept being restricted to unity. The unrestricted models have intercepts significantly less than 1, which is contrary to theoretical expectations of the value of state 111111. There are also a number of inconsistencies where worsening levels on each dimension do not attract larger decrements from full health. In some cases the coefficient estimates are approximately equal (e.g. PAIN5 and PAIN6), and in others the scale for that dimension does not have an obvious ordering (PF). However, the ROLE scale is overall ordinal, although RL 5 has a smaller coefficient that RL4.

Restricting the intercept to unity (while theoretically justifiable) worsens the performance of the models by introducing more inconsistencies, creating larger prediction errors and prediction bias. In addition this restriction does not solve the autocorrelation problem with the errors which is a feature of all of the models reported here. In both cases (restricted and unrestricted) the RE specification has fewer inconsistencies but larger prediction errors than the mean model. Figure 1 shows the impact of these problems by reporting the actual mean health state values alongside the predictions and errors from the random effects model. This shows a systematic tendency for the parametric model to over predict for a large proportion of health states

The models for the UK data generally perform better than the Japanese equivalents (Table 3). The unrestricted intercept is higher (though still significantly less than one), and there are fewer inconsistencies. Restricting the intercept to unity introduces a few more inconsistencies and systematic error in the predictions persists but in contrast to the Japanese data it does not significantly increase the size of the prediction errors. The UK models also suffered from autocorrelation problems.

Nonparametric

The non-parametric model does not provide coefficients for parameters in the same way as shown on Table 3, so the model is reported in terms of its ability to predict health state values. Figure 2 presents the resulting predicted mean health state valuations for the Bayesian nonparametric model, along with actual mean health state valuations and the residuals. Figure 1 presents the corresponding plots for the classical parametric model (1). These graphs indicate a significant improvement in predictive ability from the non-parametric model. In particular, the nonparametric model does not over predict the value for a large proportion of health states.

There are other important differences between the models. The parametric model estimates the health state utility for the worst state to be 0.3917, even though the observed average for this state is 0.2903, whereas the nonparametric model achieves 0.292. Across the 249 states that were used in the study, the predictive performance of the nonparametric model is better than the parametric model overall, with a root mean square error (RMSE) of 0.089 for the nonparametric model compared to 0.133 for the parametric model.

Data relating to 12 health states were removed from the estimation data, and the parametric and nonparametric models fitted on data for the remaining 237 states. Table 4 presents the true sample means for the 12 omitted states, together with their predicted mean and standard deviation values from the parametric and nonparametric models estimated on the reduced data set. It can be seen that the Bayesian model predicts the omitted data quite well, and substantially better than the parametric model with root mean squared prediction errors of 0.0485 compared to 0.1186. The classical standard errors are larger than the Bayesian ones, primarily because the Bayesian analysis is able to make use of other evaluations by the same respondent to estimate their individual random effects. Figures 3 and 4 show the Q-Q plots of standardised predictive errors for the 12 health states sample means, for the parametric and nonparametric models respectively. In each figure the straight line corresponds to the theoretical *N* (0,1) distribution. Figure 3 suggests that the standard model is not well validated by its predictive performance. In contrast, it is clear from Figure 4 that Bayesian model predictions are well validated as we are estimating the 12 out of sample health states more accurately than we were expecting. Our claim is fully justified by the results in Table 4. The mean of the standardised residuals for the parametric model is -0.834 compared to 0.29 for the Bayesian one, and 11 out of the 12 residuals from the standard model are negative.

These 12 states were selected to be well spread over the health state space, but almost identical overall predictive results were obtained when omitted states were chosen at random.

DISCUSSION

This study has replicated the methods developed for estimating a preference-based measure for health in the UK for use in Japan. The SF-6D had to be modified and so the sample of states used in the valuation survey were different. The resultant data set has been analysed using classical parametric modelling methods and non-parametric models estimated using Bayesian methods developed on the UK data.

While restricting the intercept to unity appears to worsen the statistical performance of the models, these restricted models are preferred for theoretical reasons concerning the value of full health. Despite a detailed examination of alternative models, it was found the additive form could not be improved upon using parametric methods. All parametric models had systematic prediction errors since the models under predict at the upper end and over predict values at the lower end. Overall the findings are that the parametric models have not performed as well as their UK counterparts in terms of proportion of variation explained, the size of the errors and number of inconsistent pairs of coefficients. However, the models have produced significant coefficients on the main effects and many of the inconsistencies with the SF-6D are either understandable i.e. (PF5 is not necessarily worse than PF6) or the coefficients are actually very close. Indeed looking at model 1d, there are only 5 inconsistencies of more than 0.01 (excluding PF5).

It is not clear why the Japanese models do not perform as well. It may be that the respondents in Japan found the SG questionnaire more difficult than their UK counterparts due to cultural differences. Interestingly, this was not found in a recent SF-6D valuation study undertaken in Hong Kong. [14]

There might be concerns about generalisability since the age of the sample only went up to 75. It has been shown that in the UK study that older people tend to give lower values to SF-6D health states than younger people, particularly for more severe states [15], but given over 75s are just 10% of the general population this should not result in substantial differences. The exclusion of 22.5% of the sample might also be concerning, but these respondents were found to be comparable to those included respondents for all variables except understanding.

Given the Japanese models did not perform as well as the UK models, and the previous findings that a non-parametric approach might improve the model [9], we decided to apply this later method to the Japanese data. This was shown to substantially improve the predictive performance of the model and so is the recommended method for generating a preference-base index from SF-36 data for use in Japan (available from authors).

Nonetheless the classical parametric results are helpful in understanding some of the differences between UK and Japanese valuation. The results clearly showed differences between the UK and Japanese samples. The Japanese valuation of the worst state was significantly higher than for the UK (0.29 vs. 0.21). Furthermore, the modelling revealed health dimensions had different relative weight underlying these overall differences. The physical functioning, role limitation and social functioning dimensions have larger decrements, whereas pain and mental health have smaller decrements than the UK models. It is noticeable that there is little difference between levels 2 to 5 of the mental health dimension.

The reasons for these differences with the UK are not clear, but are similar to the findings of the Japanese valuation of the EQ-5D [6]. The differences may arise from the translation not being linguistically equivalent, though this was minimised in earlier work with the SF-36. People in Japan may have a different attitude to risk rather than health per se, and this could explain the differences. It is interesting that the differences between the UK and a Hong Kong general population sample were less and did not follow the same pattern [14]. These cross-cultural issues in health state valuation form an important research agenda for the future.

REFERENCES

1. Brazier JE, Ratcliffe J, Tsuchiya A, Solomon J. *Measuring and valuing health for economic evaluation*. Oxford: Oxford University Press 2007.
2. Ware, J. E., Snow, K. K., Kolinski, M., Gandek, B. *SF-36 Health Survey manual and interpretation guide*. Boston: The Health Institute, New England Medical Centre, Boston, 1993.
3. Brazier, J.E., 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).
4. Dolan, P. Modelling valuation for Euroqol health states. *Medical Care* 1997; 35:351-363
5. Feeny, D. H., Furlong, W. J., Torrance, G. W., Goldsmith, C.H., Zenglong, Z., Depauw, S., Denton, M., and Boyle, M., 2002. Multi-attribute and single-attribute utility function for the Health Utility Index Mark 3 system. *Medical care 2002*, 40(20), 113(128).
6. Tsuchyia A, Idea S, Ikegami N, Nishimura S, Sakai I, Fukuda T, Hamashima C, Hisashige A, Tmura M (2002) Estimating an EQ-5D population value set: the case of Japan. *Health Economics* 11(4):341-354.
7. Shaw J, Johnson JA, Coons S.US valuation of the EQ-5D health states: Development and Testing of the D1 valuation model. Medical care 2005,43(3):203-220.
8. McCabe C, Stevens K, Roberts J, Brazier JE. Health state values from the HUI-2 descriptive system: results from a UK survey *Health Economics* 2005;14(3):231-244.
9. Kharroubi S, Brazier J, Roberts J, O'Hagan A. Modelling SF-6D health state preference data using a nonparametric Bayesian method. *Journal of health Economics 2007*; 26(3): 597-612.

10. Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K. Translation, adaptation and validation of the SF-36 Health Survey for use in Japan. Journal of Clinical Epidemiology 1997, 51 (11):1037-1044.

11. Fukuhara S, Ware JE, Kosinski M, Wada S, Gandek B: Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. Journal of Clinical Epidemiology, 51(11): 1045-1053, 1998

12. Furlong W, Feeny D, Torrance GW, et al. Guide to design and development of health state utility instrumentation. Hamilton, Ontario, Canada: McMaster University; 1990. Centre for Health Economics and Policy Analysis Paper 90-9.

13. Population of Japan Census 2005 http://www.stat.go.jp/english/data/kokusei/2005/poj/mokuji.htm

14. Lam CLK, Brazier J, McGhee SM. Valuation of the SF-6D health states is feasible, acceptable, reliable and valid in a Chinese population. *Value in Health* 2008;11:295-303.

15. Kharroubi SA, Brazier JE, O'Hagan A. Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. *Social Science and Medicine* 2007; 64:1242-1252.

Table 1: The Short Form 6D

Level	**Physical Functioning**	Level	**Pain**
1	Your health does not limit you in vigorous activities	1	You have no pain
2	Your health limits you a little in vigorous activities	2	You have pain but it does not interfere with your normal work (both outside the home and housework)
3	Your health limits you a little in moderate activities	3	You have pain that interferes with your normal work (both outside the home and housework) a little bit
4	Your health limits you a lot in moderate activities	4	You have pain that interferes with your normal work (both outside the home and housework) moderately
5	Your health limits you a little in bathing and dressing	5	You have pain that interferes with your normal work (both outside the home and housework) quite a bit
6	Your health limits you a lot in bathing and dressing	6	You have pain that interferes with your normal work (both outside the home and housework) extremely
	Role limitations		**Mental health**
1	Due to your health, doing your work and other regular activities is not difficult	1	You feel tense or downhearted and low none of the time
2	Due to your health, doing your work and other regular activities is seldom difficult	2	You feel tense or downhearted and low a little of the time
3	Due to your health, doing your work and other regular activities is sometimes difficult	3	You feel tense or downhearted and low some of the time
4	Due to your health, doing your work and other regular activities is almost always difficult	4	You feel tense or downhearted and low most of the time
5	Due to your health, doing your work and other regular activities is always difficult	5	You feel tense or downhearted and low all of the time
			Vitality
	Social functioning		
1	Your health limits your social activities none of the time	1	You have a lot of energy all of the time
2	Your health limits your social activities a little of the time	2	You have a lot of energy most of the time
3	Your health limits your social activities some of the time	3	You have a lot of energy some of the time
4	Your health limits your social activities most of the time	4	You have a lot of energy a little of the time
5	Your health limits your social activities all of the time	5	You have a lot of energy none of the time

Footnote: The SF-36 items used to construct the SF-6D are as follows: physical functioning items1, 2 and 10; role limitation due to physical problems item 3; role limitation due to emotional problems item 2; social functioning item 2; both bodily pain items; mental health items 1 (alternate version) and 4; and vitality item 2.

Table 2: Survey information

	Japan	
Interviews	600	
Excluded		
For not valuing the worst state	63 (11%)	
For giving same value to all states	72 (12%)	
For only valuing one state	0	
Included v. excluded	Inc.	Exc.
Mean age	48	44
Female %	54	49
Married %	75	73
In FT employment %	51	49
Poor understanding of task[a]	1	10

[a] judged by interviewer

Table 3: Main valuation models for UK and Japan

	Japan					**UK**			
	1a	**1b**	**1c**	**1d**		**2a**	**2b**	**2c**	**2d**
	Mean	**RE**	**Mean**	**RE**		**Mean**	**RE**	**Mean**	**RE**
C	**0.740**	**0.746**	**1.000**	**1.000**		**0.827**	**0.833**	**1.000**	**1.000**
PF2	-0.033	**-0.027**	**-0.084**	**-0.066**		-0.014	-0.021	**-0.060**	**-0.058**
PF3	-0.002	-0.007	**-0.057**	**-0.041**		0.008	-0.026	-0.020	**-0.051**
PF4	**-0.065**	**-0.069**	**-0.108**	**-0.093**		-0.027	**-0.065**	**-0.060**	**-0.088**
PF5	-0.028	**-0.036**	**-0.078**	**-0.068**		**-0.043**	**-0.044**	**-0.063**	**-0.061**
PF6	**-0.092**	**-0.105**	**-0.144**	**-0.137**		**-0.096**	**-0.135**	**-0.131**	**-0.160**
RL2	-0.013	-0.010	**-0.062**	**-0.039**		-0.019	**-0.027**	**-0.057**	**-0.056**
RL3	-0.001	-0.019	**-0.044**	**-0.044**		**-0.043**	**-0.055**	**-0.068**	**-0.076**
RL4	**-0.075**	**-0.074**	**-0.122**	**-0.102**		**-0.036**	**-0.055**	**-0.066**	**-0.078**
RL5	**-0.049**	**-0.067**	**-0.097**	**-0.092**					
SF2	-0.005	-0.022	**-0.070**	**-0.064**		-0.027	**-0.034**	**-0.071**	**-0.066**
SF3	-0.028	-0.032	**-0.088**	**-0.074**		**-0.049**	-0.022	**-0.084**	**-0.048**
SF4	**-0.044**	**-0.050**	**-0.101**	**-0.089**		**-0.057**	**-0.041**	**-0.093**	**-0.066**
SF5	**-0.072**	**-0.098**	**-0.128**	**-0.128**		**-0.073**	**-0.089**	**-0.105**	**-0.109**
PAIN2	-0.006	-0.007	**-0.069**	**-0.049**		0.008	-0.001	**-0.048**	**-0.042**
PAIN3	-0.007	**-0.019**	**-0.064**	**-0.059**		-0.001	-0.018	-0.034	**-0.046**
PAIN4	-0.016	**-0.042**	**-0.089**	**-0.087**		-0.032	**-0.026**	**-0.070**	**-0.055**
PAIN5	**-0.056**	**-0.062**	**-0.115**	**-0.104**		**-0.062**	**-0.068**	**-0.107**	**-0.103**
PAIN6	**-0.051**	**-0.093**	**-0.111**	**-0.129**		**-0.149**	**-0.155**	**-0.181**	**-0.178**
MH2	**-0.040**	**-0.022**	**-0.099**	**-0.064**		-0.026	-0.019	**-0.057**	**-0.043**
MH3	-0.021	-0.017	**-0.072**	**-0.050**		-0.022	**-0.032**	**-0.051**	**-0.055**
MH4	**-0.036**	**-0.045**	**-0.090**	**-0.077**		**-0.095**	**-0.093**	**-0.121**	**-0.115**
MH5	**-0.038**	**-0.049**	**-0.084**	**-0.072**		**-0.114**	**-0.106**	**-0.140**	**-0.125**
VIT2	-0.016	-0.002	**-0.077**	**-0.046**		-0.044	-0.006	**-0.094**	**-0.040**
VIT3	-0.019	-0.014	**-0.068**	**-0.049**		-0.037	-0.008	**-0.069**	**-0.030**
VIT4	**-0.065**	**-0.023**	**-0.112**	**-0.057**		-0.029	-0.011	**-0.069**	**-0.040**
VIT5	**-0.044**	**-0.031**	**-0.088**	**-0.055**		**-0.076**	**-0.068**	**-0.106**	**-0.087**
n	241	3070	241	3070		611	3518	611	3518
Adj R2	0.305	0.123	0.018	b		0.583	0.200	0.508	[b]
Inconsistencies	5	4	10	6		2	2	5	4
MAE	0.067	0.068	0.082	0.111		0.071	0.073	0.074	0.078
% > |0.05|	55	56	63	74		47	49	47	49
%o > |0.10|	21	23	30	52		21	21	21	24
t (mean=0)	a	-0.103	**-2.962**	**-15.560**		[a]	0.250	[a]	**-6.717**
JBPRED	2.816	3.146	0.135	0.741		0.737	1.178	0.681	2.461
LB	**944.20**	**779.02**	**129.42**	**288.72**		**520.71**	**386.63**	**169.57**	**185.3**

All models are estimated with White's heteroscedasticity consistent standard errors.
Estimates in bold are significant at $t_{0.05}$.
[a] Mean zero by definition.
[b] No R^2 statistics, GEE estimation.

Table 4: Out of sample predictions for 12 health states

		Nonparametric posterior inference		Parametric inference	
Omitted state	true sample mean	mean	(s.d.)	mean	(s.d.)
122211	0.7761	0.7208	0.0694	0.8430	0.1135
141651	0.5302	0.5556	0.0692	0.7107	0.0935
212445	0.5954	0.6092	0.0654	0.6416	0.1034
233551	0.5680	0.5041	0.0795	0.6425	0.1224
312254	0.6824	0.6078	0.0609	0.7084	0.0946
332122	0.5550	0.5832	0.0588	0.7444	0.0907
411121	0.6355	0.6478	0.0658	0.8504	0.0930
425212	0.6323	0.5682	0.0728	0.6446	0.1147
511243	0.7502	0.7375	0.0682	0.7578	0.1142
542524	0.3785	0.3982	0.0694	0.5417	0.0948
623144	0.5711	0.5813	0.0653	0.5978	0.1033
653444	0.5706	0.4759	0.0740	0.4527	0.3002

Figure 1. Sample mean and predicted health states valuations for the parametric model.

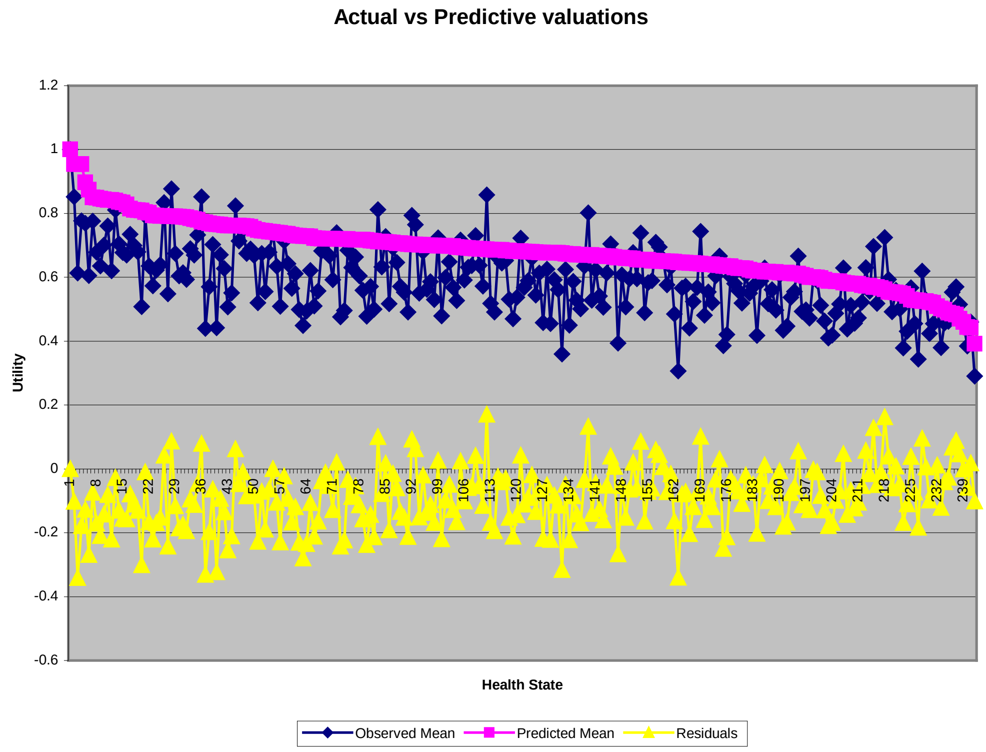


Figure 2. Sample mean and predicted health states valuations for the nonparametric model.

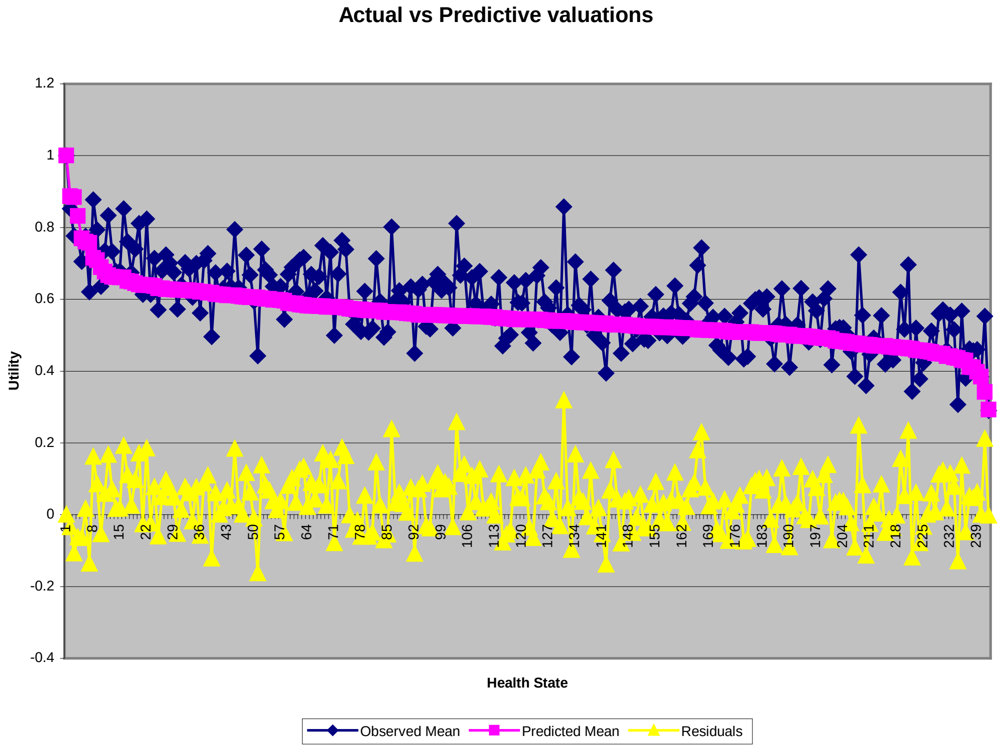


Figure 3. Q-Q plot of standardised predictive errors for the parametric model for the 12 out of sample health states.

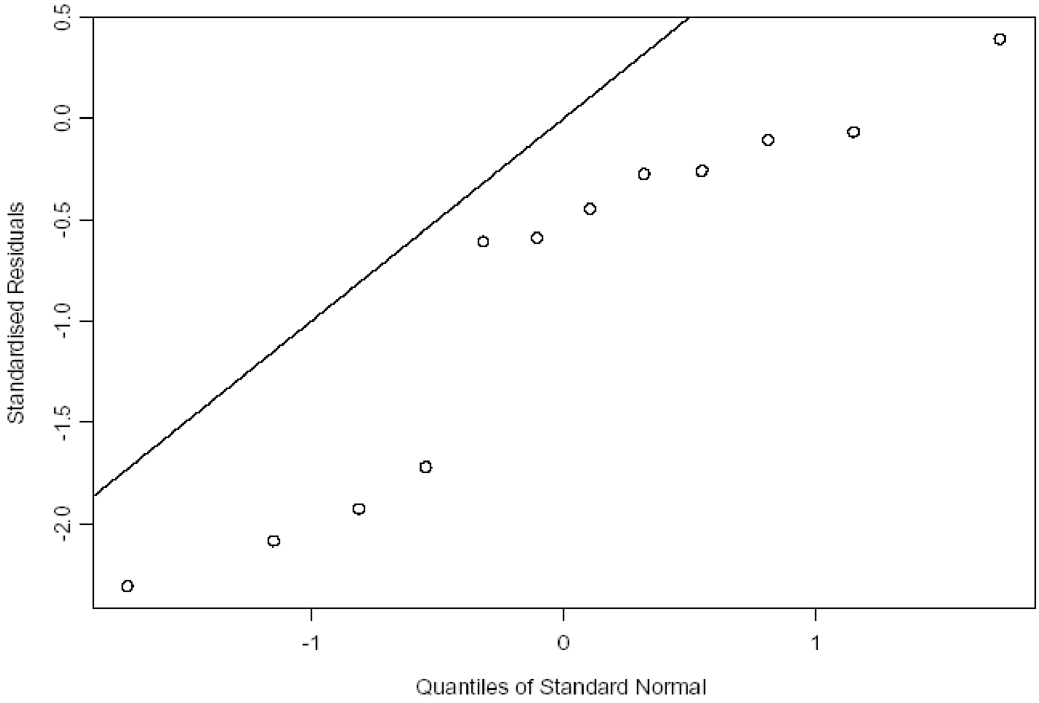
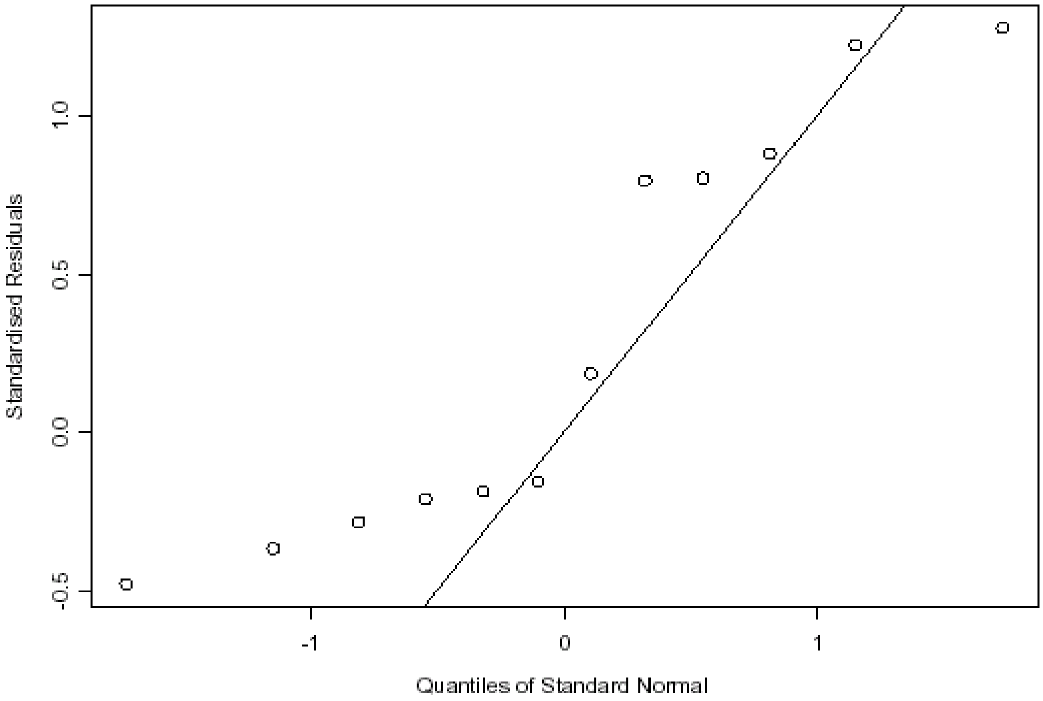


Figure 4. Q-Q plot of standardised predictive errors for the nonparametric model for the 12 out of sample health states.


Standardised Residuals
Quantiles of Standard Normal