DERIVING AN 'ENHANCED' EUROQOL FROM SF-36

Work in Progress: Not to be quoted or reproduced without permission of the authors

Authors: Mirella Longo[1], David Cohen[1], Kerry Hood[2], Mike Robling[2]

1. University of Glamorgan, Business School

2. University of Wales College of Medicine, Department of General Practice

Paper presented at the Health Economists Study Group meeting.

Nottingham, July 2000

Introduction:

Many HESG members will be familiar with requests from clinicians preparing a research proposal who have (or, rather, think they have) already decided on the design of their study and now want to invite an economist to join the team to "do the costing". Such a lack of awareness of both economics and economic appraisal can lead to considerable conflict when the economist begins challenging some of the most fundamental aspects of the study such as the choice of comparator, the primary outcome measures to be used or even the study question itself.

It is becoming increasingly common in clinical research to include measures of health related quality of life (HRQoL) among the clinical outcomes. Since clinicians look to research findings to help inform the clinical management of individual patients, they naturally tend to favour condition specific measures which fully encompass and focus on those features of HRQoL which are of greatest relevance to patients with the condition in question and which are likely to be most responsive to relatively small changes in these specific attributes. Increasingly, however, the advantages of generic measures which can be applied across all conditions are being recognised and the economist is often told that the team has already decided to include a generic HRQoL measure an outcome.

The problem for economic evaluation, however, is that none of the available generic health status measures are preference based and are therefore not capable of producing the utility weighted single index scores which are needed for cost utility analyses. When the invited economist suggests that a preference based measure should be included in the study, the most likely reaction is one of opposition. This is partly because such measures are of little direct relevance to clinicians and partly because these measures tend to be rather insensitive and not sufficiently responsive to small changes in HRQoL - or small differences in HRQoL between intervention and control groups- even when these differences are clinically important. Using a preference based measure in addition to (as opposed to substituting for) non-preference based HRQoL measures is often felt to add unnecessary complexity and cost of the trial - not to mention increasing the burden on patients (Drummond and Davies, 1991).

Given the large number of other potential conflicts which the economist and the clinical researchers will have to sort out, it would clearly be advantageous if one of the more common generic HRQoL measures could be used not only for its originally intended purpose of measuring health status, but also if it could somehow indicate preferences and therefore generate the sought after utility weighted index.

One way of doing this could be to try revalue the content of the original non-preference based measure using one of the well established preference elicitation techniques such as standard gamble or time trade off. Brazier and colleagues are currently attempting to do this with the UK version of SF-36. Their preliminary work on this involved 'restructuring' the original 36 dimensions down to 6 (the SF-6D) and assigning values to a sample of health states using a convenience sample of health professionals, managers and patients with preferences elicited by standard gamble and visual analogue methods (Brazier et al, 1998). While preliminary results

have been encouraging, a number of difficulties were encountered, and results of the full project are keenly awaited.

A second alternative could be to see whether the dimensions of a non-preference measure could be 'restructured' to the dimensions of a preference based health measure where the values attached to different health states have already been assigned. The predetermined values attached to the latter's defined health states could then be used to generate the sought after weighted single index and be used for the subsequent calculation of QALYs.

Such 'mapping' of health state descriptors from a non-preference to a preference measure has been attempted. Gudex (1986) mapped scores from the "Reusch Social Disability Rating Scale" onto the Rosser matrix. using a set of logical but fairly arbitrary decision rules. This produced only limited success in producing QALYs and has been criticised *inter alia* for the lack of attempt to test the predictive accuracy of the translation procedure (Brazier et al, 1999).

Our study, although adopting a different approach to Gudex, is similarly attempting create a preference based weighted index from responses elicited from a non-preference health status questionnaire. Using multiple regression analysis we assessed the extent to which a preference based index can be generated from non-preference data and secondly to assess if the greater responsiveness to changes in health status of the former can increase the responsiveness of the latter.

Choice of measures.

The SF-36 health profile is derived from a standardised, non-disease specific, self-completion questionnaire containing 36 items covering 8 dimensions of health; general health, bodily pain, mental health, physical functioning, role emotional, role physical, social functioning and vitality (Ware and Sherbourne, 1992). Within each dimension, responses are coded and summed to provide raw scores which are then transformed onto a 0 (worst) to 100 (best) scale. SF-36 however, does not attach weights (utilities) to the dimensions and thus generates a health profile rather than an index score. It is thus normally not possible to say whether someone scoring high on dimension X and low on dimension Y has a more preferred or less preferred health state than someone scoring low on dimension X and high on dimension Y. Accordingly, if both can be returned to perfect health through an effective intervention, it is not possible to say which intervention has yielded the greater health (QALY) gain.

At the same time, there is extensive evidence on the descriptive validity of the SF-36 as a measure of health status for a large range of medical conditions. Accordingly, it has become one of the most widely used measures of health status in clinical research (Garrett et al, 1993). A UK version (substituting British expressions for American) has been validated (Brazier et al 1992). Importantly, the SF-36 appears to meet very well the criteria used in assessing psychometric tools; practicality, internal consistency, reliability, validity (content, face and construct) and responsiveness. (Brazier and Devrill,1999).

The EuroQol health index, like SF-36, is derived from a standardised non-disease specific, self-completion questionnaire. Health is described across 5 (originally 6) dimensions; mobility, self-care, usual activity, pain/discomfort and anxiety/depression and over 3 level per dimension (EuroQol Group, 1990). A tariff of values for each of the resulting health states, was obtained from a large random sample of the adult population of the UK using time trade methods. (Dolan et al, 1995). Anecdotal evidence suggests that EQ-5D is now the most commonly used measure in ongoing cost utility studies.

Comparisons of SF-36 with EQ-5D from the literature.

Brazier et al (1999) provide a review of the 'relatively few' studies that have attempted to explore the relationship between preference based and non-preference based measures. Their conclusion was that "Overall it would appear that there are only low-to-moderate correlations between [non-preference health measures] and preference measures, but this was not consistent between or within methods" (p:89)

The only study they identified that directly compared EuroQol with SF-36 in a whole population (i.e. not restricted to patients with any specified conditions) was Brazier et al (1993). This tested the construct validity of EQ-6D (the earlier version of EQ-5D) by comparing the two measures in terms of how the pattern of health between groups fit predictions from a set of hypotheses (e.g. that there would be an inverse relationship between age and functioning or between recent use of health services and perceived health). This was done via a postal survey of patients from two practices in Sheffield. They found evidence for the construct validity of the EuroQol dimension responses and the derived score in terms of distinguishing between groups with expected health differences. The EuroQol, however was shown to be less responsive than SF-36 throughout the range of health states and particularly less responsive in the case of for those states representing relatively low levels of health. Spearman rank coefficients between the two varied from 0.48 to 0.60.

Other studies have compared EQ-5D with SF-36 in studies of patients with specific conditions. Brazier et al (1996) tested construct validity and responsiveness of these two measures together with the OPCS Disability Survey on a sample of elderly women (age >75) who took part in a randomised controlled trial. Results were not dissimilar to the general population study above. Substantial agreement was shown between the three instruments and evidence for their construct validity against age and recent use of health services was demonstrated. SF-36 was again found to be more responsive to changes in health status (on the basis of hypothetical changes) and again this was particularly so at lower levels of health.

More recently, Dorman et al (1999) and Myers and Wilks (1999) compared EQ-5D with SF-36 scores in stroke patients and patients with chronic fatigue syndrome respectively. Both studies found good correlations generally between the two measures, but there were exceptions in both cases. In the case of stroke patients, the mental health domain of SF-36 correlated only poorly with the psychological functioning domain of EQ-5D, while in the case of patients with chronic fatigue syndrome, poor correlation was shown with the physical limitation domain.

The overall message from these and other similar studies (e.g. Hollingworth et al, 1995, Brazier et al, 1999b) is that EQ-5D shows generally good agreement with SF-36, but the degree of agreement can vary between clinical condition. At the same time, it is has been shown to be less responsive than SF-36 in most cases.

The current study.

Method:

We have recently completed a randomised controlled clinical trial examining the effects of alternative ways of introducing guidelines on the management of breast disorders in primary care (the BRIDGE study) (BRIDGE Study Group, 1999). Women who consulted their general practitioner with breast problems were asked to complete the SF-36 and a condition specific questionnaire (the Cardiff Breast Scales) 6 months after their initial consultation.

In order to assess test-retest reliability and responsiveness, one in nine of the women who responded to the 6 month questionnaire were again sent the SF-36 and condition specific questionnaire at 7 months. A transition question was added to the 7 month questionnaire asking whether they felt their health had changed in the past month. The group who indicated that there had been no change in health status were used for assessing reliability and those who indicated a improvement or deterioration in health status were used to assess responsiveness.

Twelve months after the initial consultation, all study women were again sent the SF-36 and condition specific questionnaire, but this time EQ-5D was added. Evidence from a cross-over study (Dorman et al, 1999) suggests the absence of ordering effect i.e. that completing one questionnaire after completing of the other, does not affect patients' subsequent responses.

We randomly divided the 12 month data into a training subset (75%) and a validation subset (25%). EuroQol scores are known to be skewed (Dolan,1997). To alleviate some of the skewness present in the data the EuroQol scores were transformed using the formula $ln(2 - \text{EuroQol})$ which produced the best fit.

A multiple regression model was then fitted to the training subset with the transformed EuroQol scores as the dependent variable and all eight subscales of the SF-36 as explanatory variables. Non significant variables were then removed so that an optimal model could be produced. Goodness of fit was assessed for the optimal model on the complete dataset. The quality of the model's predictions was then assessed using the validation subset who were not included in the model fitting.

Since EQ-5D was not included in the 6/7 month surveys, an 'enhanced' EuroQol score was calculated from the subscales of the SF-36 using the model derived from the 12 month dataset. This made it possible to compare the responsiveness of SF-36 with the 'enhanced' EuroQol scores, along with the condition specific scales. Our hypothesis was that the SF-36 and the 'enhanced' EuroQol scores would have similar levels of responsiveness (i.e. the 'enhancement' would increase the responsiveness of EuroQol), whilst the condition specific scale would be more responsive than either.

The index of responsiveness (Guyatt et al, 1987) was used to assess the sensitivity of these scales.

Results

Questionnaires were returned from 638 women at 12 months. Of these, 20 had not completed the EQ-5D. The remaining 618 women were the randomly divided into the training subset (n = 468) and the subset validation set (n = 150).

Because the questionnaires were sent out and returned as a 'pack' it was not possible to compare response rates between questionnaires. In terms of missing data, however, the EQ-5D performed much better. While 20 respondents did made no attempt to complete the EQ-5D questionnaire, those who did answered all questions. This compares with SF-36 where missing data by subscale were as follows; general health 27, vitality 12, mental health 11, bodily pain 36, role emotional 54, role physical 45, social functioning 45, physical functioning 43. This better performance of EQ-5D in terms of missing data is consistent with other studies (e.g Dorman et al, 1999, Brazier et al, 1996)

Only 11 women in the training subset had given themselves a negative score on EQ-5D which indicates a quality of life worse than death. Given these small numbers it was felt to be inappropriate to attempt to make predictions within this region of scores.

1. Predicting EuroQol from SF-36

A multiple regression model was fitted to this group with all eight subscales as explanatory variables. The coefficients from this model are given in Table 1 below. This model explained 70.0% of the variation in EuroQol scores. It can be seen that three of the subscales of the SF-36 did not significantly predict the transformed EuroQol scores. These are mental health, role-emotional and role-physical. A reduced model only including the remaining five subscales was then fitted. This model explained 70.1% of the variation in EuroQol scores. The coefficients for this model are given in Table 2 below.

Table 1 Coefficients from the full model

Model	Unstandardised coefficient	Standardised coefficient	Significance
Constant	0.61600		<0.001
Bodily Pain	-0.00200	-0.34	<0.001
General Health	-0.00076	-0.11	0.012
Mental Health	-0.00038	-0.05	0.286
Physical Functioning	-0.00126	-0.019	<0.001
Role – Emotional	-0.00015	-0.04	0.374
Role – Physical	0.00040	0.01	0.814
Social Functioning	-0.00108	-0.18	0.001
Vitality	-0.00092	-0.12	0.007

Table 2 Coefficients from the reduced model

Model	Unstandardised coefficient	Standardised coefficient	Significance
Constant	0.60700		<0.001
Bodily Pain	-0.00204	-0.35	<0.001
General Health	-0.00087	-0.12	0.004
Physical Functioning	-0.00113	-0.17	<0.001
Social Functioning	-0.00126	-0.21	<0.001
Vitality	-0.00109	-0.15	<0.001

This reduced model was then used to calculate EuroQol scores for the validation subset. When the model was used to predicted transformed EuroQol scores, it explained 67.7% of the variability in these scores. The fact that this is only a slight reduction on the 70.1% observed in the training subset indicates a relatively robust fitting model.

2. Assessing responsiveness

Having shown that SF-36 can be used to predict EuroQol scores relatively well, the next step was to assess the responsiveness of the new scores. The table below shows the Guyatt index scores for the enhance EuroQol, the SF-36 and the condition specific Cardiff Breast Scales. The index is determined by dividing the mean change in the group who stated that they had experience a change in health by the standard deviation of the group who stated no change in health. Higher index scores represent greater responsiveness.

Responsiveness of Enhanced EuroQol, SF-36 sub-scales and the two condition – specific sub-scales

	Number in group reporting a change	**Number in group reporting no change**	**Index of Responsiveness**
Enhanced EQ	18	32	0.55
SF-36			
Bodily Pain	22	36	0.34
General Health	23	37	0.27
Mental Health	22	35	0.09
Physical Functioning	22	37	0.38
Role Emotional	21	30	0.29
Role Physical	21	36	0.21
Social Functioning	21	36	0.44
Vitality	22	35	0.35
Cardiff Breast Scale			

General Wellbeing	22	37	0.55
Relationships	14	32	0.35

The enhanced EuroQol has reached the same level of responsiveness as the condition specific measure and is greater than all of the subscales of the SF-36. Whilst we do not have an assessment of the responsiveness of the original EuroQol on this dataset, previous studies have shown in both general and condition specific populations that EuroQol is less sensitive than the SF-36. Therefore the results on this single dataset are encouraging for the new 'enhanced' EuroQol.

Discussion:

This study has shown that in women presenting in primary care with breast conditions it is possible to predict a more sensitive version of the EuroQol using the SF-36 subscales. This group contains a relatively heterogeneous set of conditions, albeit site specific. Due to the small number of very serious ratings of quality of life (negative EuroQol scores), it has not been possible to test the prediction in this region. Further work on a more serious condition would be required to study this.

In general, it would be useful to take a similar group of women, but with different conditions and test to see if the weights derived in this study are comparable. This would allow us to move towards being able to produce economic assessments that are sensitive to changes in quality of life. A study which allows us to directly compare the responsiveness of these two versions of EuroQol would also be required to move this research forward.

A further refinement on the above modelling using standard least squares regression techniques would be to fit a model which accounted for measurement error in the SF-36 subscales (Cheng and Van Ness, 1999). However, this measurement error modelling should only add to the quality of the prediction, which has already been shown to be reasonable.

References

Brazier J, Harper R, Jones N et al. Validating the SF-36 health survey questionnaire: a new outcome measure for primary care. Br Med J 1992; 305:160-164

Brazier J, Jones N, Kind P. Testing the validity of the EuroQol and comparing it with the SF-36 health survey questionnaire. Quality of Life Research 1993;2:169-180.

Brazier J, Walters S, Nicholl J, Kohler B. Using the SF-36 and EuroQol on an elderly population. Quality of Health Research 1996;5:195-204

Brazier J, Usherwood T, Harper R Thomas K. Deriving a preference-based single index from the UK SF-36 health Survey. J Clin Epidemiol 1998; 51(11):1115-1128

Brazier J, Deverill M. A checklist for judging preference base measures of health related quality of life: Learning from psychometrics. Health Economics 1999;8(1) 41-52

Brazier J, Deverill M, Green C, Harper R, Booth A. a review of the use of health status measures in economic evaluation Health Technol Asses 1999; 3(9).

Brazier J, Harper R, Munro SJ, Snaith ML. Generic and condition specific outcome measures for people with osteoarthritis of the knee. Rheumatology 1999:38:870-877.

BRIDGE Study Group. The presentation and management of breast symptoms in general practice in South Wales. Br J Gen Pract 1999;49:811-812

Cheng C, Van Ness JW. Statistical regression with measurement error. Kendall's Library of Statistics 6. Arnold (London) 1999.

Dolan P. Aggregating health state valuations. J Health Serv Res Policy 1997;2(3) 160-165.

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a population survey. Centre for Health Economics Discussion Paper 138. University of York. York.1995.

Dorman PJ, Dennis M, Sandercock P. How do scores on the EuroQol relate to scores on the SF-36 after stroke? Stoke 1999;30:2146-2151.

Drummond MF, Davies L. Economic analysis alongside clinical trials: revising the methodological issues. In. J. Tech Ass Health Care 1991; 7(4):561-673

EuroQol Group. EuroQol - a facility for the measurement of health related quality of life. Health Policy 1990; 16:199-228

Garrett AM, Ruta DA, Abdalla MI et al. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS Br Med J 1993:306:1440-1444.

Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-8

Gudex C. QALYs and their use by the health service. Discussion paper 20. York:Centre for Health Economics, University of York. 1986.

Hollingworth W, Mackenzie R, Todd CJ, Dixon AK. Measuring changes in quality of life following magnetic resonance imaging of the knee: SF-36, EuroQol or Rosser index? Quality of Life Research 1995:4:325-334

Myers C, Wilks D. Comparison of EuroQol EQ-5D and SF-36 in patients with chronic fatigue syndrome. Quality of Life Research 1999; 8:9-16

Ware J, Sherbourne CF. The SF-36 Short Form health status survey 1, conceptual framework and item selection. Med Care 1992;30:473-483