Article:
O'Brien, B.J., Spath, M., Blackhouse, G. et al. (2 more authors) (2003) A view from the Bridge: agreement between the SF-6D utility algorithm and the Health utilities Index. Health Economics, 12 (11). pp. 975-982. ISSN 1057-9230

https://doi.org/10.1002/hec.789

HEALTH ECONOMICS
Health Econ. **12**: 975–981 (2003)
Published online 27 February 2003 in Wiley InterScience (www.interscience.wiley.com). **DOI:**10.1002/hec.789

A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index

Bernie J. O'Brien[a,b]*, Marian Spath[b,c], Gordon Blackhouse[a,b], J.L. Severens[d], Paul Dorian[e] and John Brazier[f]
[a] *Department of Clinical Epidemiology and Biostatistics, McMaster University, Canada*
[b] *Centre for Evaluation of Medicines, St Joseph's Hospital, Canada*
[c] *Department of Medical Technology Assessment, University Medical Centre, Nijmegen, Netherlands*
[d] *Department of Health Organisation, Policy & Economics, University of Maastricht, Netherlands*
[e] *Department of Medicine, St. Michael's Hospital, Toronto, Canada*
[f] *Sheffield Health Economics Group, University of Sheffield, UK*

Summary

Background: The SF-6D is a new health state classification and utility scoring system based on 6 dimensions ('6D') of the Short Form 36, and permits a ''bridging'' transformation between SF-36 responses and utilities. The Health Utilities Index, mark 3 (HUI3) is a valid and reliable multi-attribute health utility scale that is widely used. We assessed within-subject agreement between SF-6D utilities and those from HUI3.

Methods: Patients at increased risk of sudden cardiac death and participating in a randomized trial of implantable defibrillator therapy completed both instruments at baseline. Score distributions were inspected by scatterplot and histogram and mean score differences compared by paired *t*-test. Pearson correlation was computed between instrument scores and also between dimension scores within instruments. Between-instrument agreement was by intra-class correlation coefficient (ICC).

Results: SF-6D and HUI3 forms were available from 246 patients. Mean scores for HUI3 and SF-6D were 0.61 (95% CI 0.60–0.63) and 0.58 (95% CI 0.54–0.62) respectively; a difference of 0.03 ($p < 0.03$). Score intervals for HUI3 and SF-6D were (-0.21 to 1.0) and (0.30–0.95). Correlation between the instrument scores was 0.58 (95% CI 0.48–0.68) and agreement by ICC was 0.42 (95% CI 0.31–0.52). Correlations between dimensions of SF-6D were higher than for HUI3.

Conclusions: Our study casts doubt on the whether utilities and QALYs estimated via SF-6D are comparable with those from HUI3. Utility differences may be due to differences in underlying concepts of health being measured, or different measurement approaches, or both. No gold standard exists for utility measurement and the SF-6D is a valuable addition that permits SF-36 data to be transformed into utilities to estimate QALYs. The challenge is developing a better understanding as to why these classification-based utility instruments differ so markedly in their distributions and point estimates of derived utilities.

Keywords SF-6D; HU13; health utilities

*Correspondence to: Centre for Evaluation of Medicines, 105 Main St. E., Level P1, Hamilton, Ontario, L8N 1G6 Canada.
E-mail: obrienb@mcmaster.ca

Received 13 June 2002
Accepted 11 November 2002

Introduction

The Short Form 36 (SF-36) health survey is a standardized questionnaire used to assess health-related quality of life (HRQL) across eight dimensions of physical functioning, role limitations (physical), bodily pain, general health, vitality, social functioning, role limitations (emotional) and mental health. Developed originally for the Rand Medical Outcomes study, the SF-36 is now one of the most widely used measures of HRQL, with application to evaluation of specific interventions and survey assessment of population health over time and between groups [1]. The scale limits of each dimension of SF-36 are 0 and 100 with higher scores indicating higher levels of HRQL. Although there is a facility for an aggregate SF-36 score across dimensions, this is a simple arithmetic aggregation across the scales and assigns them all equal weight in the total score. This assumption limits the ability of SF-36 in making assessments of the net impact of an intervention on HRQL. For example, assume intervention A increases scores on the first 4 of the SF-36 dimensions and decreases scores on the last 4 dimensions, with the opposite being observed for intervention B. In this circumstance one can only say that interventions A and B are different in their profiles of HRQL and not that one is better or worse than the other.

Establishing the 'net' effect of an intervention on HRQL is particularly important in economic evaluation where the goal is to compare the added cost of a treatment to it's added effectiveness in composite health units such quality-adjusted life-years (QALYs) [2]. The QALY measure is a quantitative framework for combining data on survival and HRQL into a single metric; survival time adjusted for HRQL using a utility scale with anchors of full health (=1) to death (=0). Based on principles of health state utility measurement [3], utility scores can be measured either directly, using preference trade-off exercises such as standard gamble or time trade-off, or indirectly using multi-attribute health status utility classification systems such as the Health Utilities Index.

In our experience, many circumstances arise where SF-36 data have been collected in a study and researchers are interested to derive 'Q' weights for quality-adjusted survival and estimating QALYs. The question is whether it is feasible to derive a valid and reliable method for creating a 'bridge' between a respondent's eight dimensional SF-36 score and a corresponding health state utility weight in the interval 0-1.

There are two general approaches to the empirical bridging between SF-36 and utility. The first is to use regression analysis on a dataset where subjects have completed both SF-36 and a utility measure. Nichol [4] used this method with the Health Utilities Index (HUI) and Fryback [5] with the Quality of Well Being Index (QWB); both studies fit linear additive models by ordinary least squares with dimension scores of SF-36 as independent variables explaining between 50 to 60% of the variance in utility score. The second approach is to define and value a series of health states using combinations of response levels (e.g. 'a little of the time', 'most of the time') over SF-36 dimensions. This approach draws directly from the conceptual and empirical logic of multi-attribute utility theory (MAUT) [6] used in the construction of HUI [7] and EQ5D [8] where an additive or multiplicative utility function is estimated based on a fractional factorial design from the universe of all possible health states. The 'bridge' back to SF-36 is formed via the beta coefficients on the utility scoring formula and the corresponding levels on SF-36 dimensions.

This second approach has been adopted by Brazier *et al.* who have reported both pilot work [9] and a complete survey based on 836 respondents in the UK [10] using an algorithm formed from 6 dimensions of the SF-36 and referred to as the SF-6D. In this study we assess the within-subject agreement between the SF-6D utility algorithm of Brazier *et al.* [10] and the HUI (mark 3) using baseline assessments in patients at high risk of sudden cardiac death who are subjects in a randomized trial comparing drug therapy with an implantable cardioverter defibrillator (ICD).

The key pragmatic value of 'bridging the gap' between utility measures is that it enables the comparison of studies that have been conducted with either the SF-36 or HUI3. But it should be stressed that our motivation is the study of agreement between alternative utility measurement scales and not the implied validation of the SF-6D system on the presumption that the HUI3 is a gold standard. Given that there is no criterion standard for health state utility, the comparison of alternative instruments and their underlying stimuli should improve our understanding of construct validity in this field.

Methods

The SF-6D utility algorithm

The SF-6D algorithm is described in detail elsewhere [9,10]. In brief, the algorithm is based on 6 of the 8 dimensions of SF-36 – 'General Health' is omitted and 'role limitation (physical)' and 'role limitation (emotional)' are combined. Each dimension has a number of levels such as 'limited a lot' and 'limited a little'. The combination of levels over dimensions defines a universe of 18 000 unique health states ($=6\times4\times5\times6\times5\times5$). The infeasibility of measuring utilities for all 18 000 states warrants a fractional factorial design and therefore 249 health states were valued by 836 respondents from the UK general public. The method of standard gamble was used to elicit utility values using a two-stage 'cascade' [2] technique: (1) a total of 6 states per subject were valued relative to an upper anchor of no dysfunction (level 1 on all dimensions) and a lower anchor of the lowest levels for all dimensions; (2) the lower anchor was then valued against an upper anchor of full health (level 1 on all dimensions) and a lower anchor of death. The second lottery values were then used to normalize the first lottery values on the conventional (0,1) scale of dead to full health. Finally a linear additive utility model was fit by ordinary least squares and with SF-6D item-levels and interactions as covariates. The model preferred by the authors had an R^2 of 0.53.

The DINAMIT trial dataset

The Defibrillator in Acute Myocardial Infarction Trial (DINAMIT) is a randomized, open label, parallel group randomized trial of implantable cardioverter defibrillator (ICD) therapy versus usual care in patients with recent myocardial infarction and risk factors for sudden cardiac death. Details of the study protocol are available elsewhere [11]. The study will recruit 525 patients based on the hypothesis that ICD therapy can reduce the annual risk of all-cause mortality from 30% to 20%. DINAMIT has sub-studies addressing both cost-effectiveness and quality of life. For the latter, and of relevance to this study, patients complete both the HUI and the SF-36 at baseline, 6 months and 1 year. Recruitment into DINAMIT is continuing, and the present study is based on within-subject comparison of the available sample ($n=246$) at June 2000 who had completed baseline (pre-randomization) assessments for HUI and SF-36.

Statistical methods

Data on classification into HUI3 and SF-6D were converted into utility scores based upon the scoring algorithms from Feeny *et al.* [12] and Brazier *et al.* [10] with imputation for missing item values for SF-6D based on the guidance for SF-36 [13]. Descriptive statistics for the sample were computed along with mean, standard deviation and 95% confidence intervals for utilities. The within-subject difference in mean utility score was tested by paired *t*-test. Utility data from the two instruments were presented as a scatterplot. Pearson correlation was computed and bivariate linear regression was fit by ordinary least squares.

A limitation of simple correlation and regression is that two measures can be perfectly correlated (i.e. fall on a straight line) but have poor agreement, such that the straight line through the X–Y scatterplot is not at 45° with a zero intercept. For analysis of agreement we computed the intra-class correlation coefficient (ICC) by two-way analysis of variance with factors of subject and instrument. The ICC is equal to 1 only when values lie on a 45° straight line through the origin and there is perfect agreement within-subjects for the two instruments. Finally, because independence between utility attributes is an important property of multi-attribute utility measures we computed a correlation matrix between quality-of-life dimensions for both HUI3 and SF-6D.

Results

A total of 310 patients had completed baseline evaluation in the trial at the time of our study. Patients had a mean (SD) age of 62.1 (11.2) years and 26.4% were female. There were 239 SF-6D and 263 HUI3 forms with no missing items. After following the guidance for imputing missing item values for SF-36 [13] and using the same rule for HUI3, there were 267 SF-6D forms and 281 HUI3 forms eligible for analysis. Finally, the joint set of patient forms with complete and imputed data for both SF-6D and HUI3 was 246.

The distributions of utility scores by SF-6D and HUI3 are presented in Figure 1. The mean utility score for HUI3 was 0.61 (95% CI 0.60–0.63) and

the mean score for SF-6D was 0.58 (95%CI 0.54–0.62); the mean difference in score was 0.03 ($p = 0.03$ by paired t-test). The SF-6D scores passed the Kolmogorov-Smirnoff test for normality at the 5% level but this was not true for HUI3 scores which show a skewed and bi-modal distribution (Figure 1). The SF-6D had a minimum value of 0.30 and a maximum value of 0.95; this was a much smaller range than HUI3 which had a minimum of −0.21 and a maximum of 1.00.

A scatterplot of SF-6D against HUI3 is presented in Figure 2. Also plotted in Figure 2 are results from a bivariate regression indicating a positive linear association between the two measures with an R^2 goodness-of-fit value of 0.34. This corresponds to a Pearson Correlation Coefficient of 0.58 (95% CI 0.48–0.68). Also shown in Figure 2 is the 45° line of perfect agreement between the two instruments. The intraclass correlation coefficient for agreement was 0.42 (95% 0.31–0.52).

Correlation matrices between dimensions of functioning for both SF-6D and HUI3 are presented in Table 1. All dimensions of SF-6D have significant ($p<0.05$) positive correlation ranging from 0.12 (mental vs physical) to 0.40 (vitality vs social). In contrast, for HUI3 of 28 correlations between dimensions only 14 are significant ($p<0.05$) and these range from 0.11 (hearing vs emotion) to 0.34 (ambulation vs dexterity).

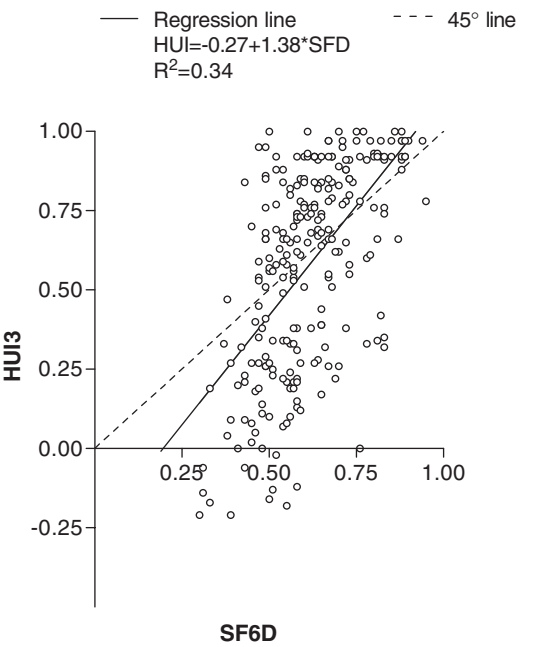

Figure 2. Scatterplot of HUI3 and SF-6D utility scores with linear regression line and 45° (perfect agreement) line

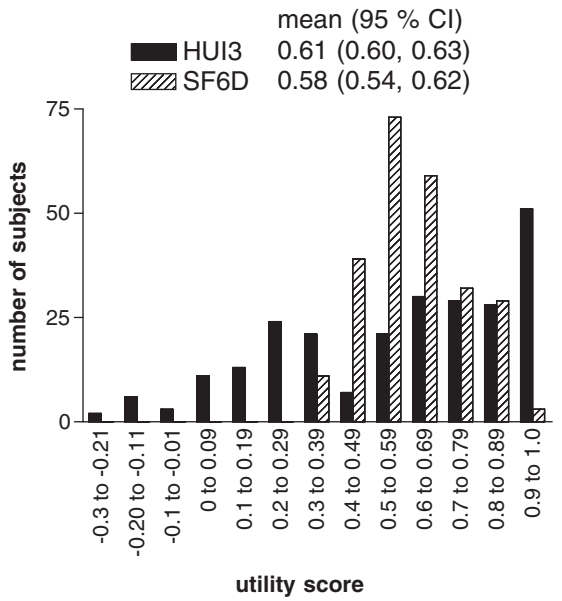

Figure 1. Frequency distributions for HUI3 and SF-6D utility scores with mean and 95% CI

Discussion

Bridging the gap between health profile measures such as SF-36 and health state utilities for construction of QALYs has obvious appeal. For a resource-constrained researcher designing a prospective evaluative study such as a clinical trial, being able to use SF-36 to obtain both health-related quality of life scores and health state utilities for QALYs is clearly desirable. In this study we assessed agreement between the new SF-36-derived utility measure known as SF-6D and HUI3, an established and widely used utility measure.

Our results generally indicate poor agreement between utilities derived by SF-6D and HUI3. The intra-class correlation coefficient was 0.42 and it is clear from the scatterplot and regression line (Figure 2) that the data do not cluster around the 45° line of perfect agreement. The suggestion from Figure 2 and related analysis is of systematic

Table 1. Correlation matrix (Kendall's Tau) for Health Utilities Index and Short-Form 6D

	Vision	Hearing	Speech	Emotion	Pain	Ambulation	Dexterity	Cognition
(a) Health Utilities Index:								
Vision	1.00							
Hearing	0.09	1.00						
Speech	0.04	0.07	1.00					
Emotion	0.12[a]	0.11[a]	0.09	1.00				
Pain	0.11[a]	−0.02	0.03	0.24[b]	1.00			
Ambulation	0.19[b]	−0.02	0.01	0.21[b]	0.29[b]	1.00		
Dexterity	0.09	−0.08	0.10	0.27[b]	0.21[b]	0.34[b]	1.00	
Cognition	0.09	0.08	0.08	0.18[b]	0.20[b]	0.27[b]	0.27[b]	1.00
	Physical	Role	Social	Pain	Mental	Vitality		
(b) Short-Form 6D:								
Physical	1.00							
Role	0.19[b]	1.00						
Social	0.24[b]	0.36[b]	1.00					
Pain	0.26[b]	0.37[b]	0.39[b]	1.00				
Mental	0.12[a]	0.19[b]	0.29[b]	0.24[b]	1.00			
Vitality	0.24[b]	0.39[b]	0.40[b]	0.30[b]	0.17[b]	1.00		

[a] Correlation is significant at the 0.05 level (2-tailed).
[b] Correlation is significant at the 0.01 level (2-tailed).

disagreement between SF-6D and HUI3; for HUI3 values greater than 0.75, corresponding SF-6D values are markedly *lower* and lie to the left of the 45° line; for HUI3 values below 0.4 the corresponding SF-6D utility is markedly *higher*.

The distributional differences between the SF-6D and HUI3 are evident from the frequency distributions in Figure 1. Data for HUI3 covers the interval from −0.21 to 1.0 whereas the SF-6D interval is only 0.30–0.95. The scatterplot shows a much wider range of HUI3 values for any given SF-6D value and the SF-6D distribution is normal whereas the HUI3 distribution is not. Consideration of scale extremes is also informative; for example, 11 patients scored less than zero on HUI3 ('worse than death') and on SF-6D these same patients all scored 0.25 or greater. In aggregate, the difference between measures in mean score is 0.03 and statistically significant at the 5% level. Finally, we considered the important principle of independence between utility dimensions and showed that the correlations between dimensions are greater for SF-6D than HUI3.

Three questions arise from this analysis; (1) are the observed differences in mean utilities quantitatively 'important'?; (2) can differences between the utility scores be explained in terms of differences in the instruments; (3) what are the implications of these findings for prospective utility measurement and further research?

The scoring basis of both HUI3 and SF-6D is the standard gamble where strength of preference for a health state is quantified in terms of the risk of death a person would accept to avoid living in the specified health state. For a utility score difference of 0.03 not to be important, one would have to be indifferent about a 3% difference in mortality risk. In a recent review by Drummond [14] a difference of 0.03 is cited as being used in sample size calculations as a minimally important difference. Of note, Drummond also cites the same minimal value of 0.03 as being used in other utility systems such as EQ-5D. Furthermore, Samsa *et al.* [15] reviewed the topic of minimally important differences in health status measures and noted that differences in HUI scores as small as 0.01 have been considered to be important. Ultimately, whether a difference of 0.03 is important will depend upon the evaluative context.

Can differences in utility scores be explained by differences in the SF-6D and HUI3 instruments? Although the two instruments appear to be measuring similar utility-based concepts of strength-of-preference for health states, the concept and characterization of health differs between the two measures. The HUI3 classification is based on a 'within-the-skin' definition of health that

focuses on impairments in vision, hearing, speech, and cognition. The SF-36 is based on the broader World Health Organization definition of health, and addresses concerns with physical, mental and social functioning. Therefore, the SF-6D measures the impact of health on aspects of HRQOL such as role and social functioning. Although the two instruments do overlap in dimensions such as pain, emotion or mental health, and ambulation or physical functioning – there is no one-to-one match on the dimensions of health being measured. The SF-36 also includes positive as well as negative aspects of health and therefore the definition of 'full health' is not the same between the two measures.

There are several differences between the instruments in how health states are valued. Valuation of health states for both SF-6D and HUI3 were done using the standard gamble technique in different ways. SF-6D health states were directly valued by respondents using standard gamble; in constrast, HUI3 states were directly valued by respondents by visual analogue scale, with standard gamble scores being forecast by a statistical power transformation. Furthermore, the utility-theoretic framework for measurement is different for the two instruments. In developing HUI3, Feeny *et al*. [12] reduced the health-state sampling demands of their valuation task by invoking simplifying assumptions from multi-attribute utility theory about the relationship between dimensions. For example, the multiplicative form of utility function, which is the basis of HUI3 scoring, permits a form of interaction between dimensions that is assumed to be the same between all dimensions and for all levels of each dimension. In contrast, the SF-6D was valued using regression techniques, with a linear additive model chosen on the basis of goodness-of-fit and parsimony. The model estimates decrements for each movement away from the highest level of functioning for each dimension. It also contains an interaction term which assumes the value of one when any dimension in the health state is at the most severe level, and zero otherwise.

What are the implications of our study for future research? The first issue is that any comparison between HUI3 and SF-6D utilities and QALYs should be made with caution. What we cannot, as yet, determine from our data is the extent to which the instrument comparability problem still persists if one compares within-instrument *differences* in utility scores. For example, the difference in utility – over time or between groups – may be the same for SF-6D and HUI3 measured utilities, even though our study suggests that the absolute utility scores will differ. If this were shown to be true then it would provide some comfort for inter-study comparability of QALYs. Subsequent data from our DINAMIT trial will permit us to address this question, because we will have repeated measures within subjects and the randomization blinding will be broken to permit comparison between treatment groups. It would also be instructive, although difficult, to try to disentangle the competing reasons as to why the instrument scores differ; is this due to the different concepts of health being measured or the extent to which they invoke the assumptions of multi-attribute utility theory? The challenge is that these multiple explanatory factors are confounded in current datasets and new experimental evidence would be costly to generate.

There are several limitations to this study. First, we have studied a single population with a specific disease and the extent to which our observations can be generalized is unclear. Second, we have compared the new SF-6D utility algorithm only to HUI3 – a comparison that arose by opportunity not design. A broader study would be helpful, comparing SF-6D with other multi-attribute systems such as EQ-5D and with direct measurement of utility by standard gamble.

In conclusion, our study casts doubt on the whether utilities and QALYs estimated via SF-6D are comparable with those from HUI3. At this stage it is not clear whether such differences arise mainly from differences in underlying concepts of health being measured or different utility-theoretic measurement approaches. There is no gold standard for health state utility measurement and the SF-6D algorithm is a valuable addition that permits SF-36 data to be transformed into utilities to estimate QALYs. The challenge is developing a better understanding as to why these classification-based utility instruments differ so markedly in their distributions and point estimates of derived utilities.

Acknowledgements

This work was initiated while Marian Spath was a visiting graduate student from the University of Nijmegen to the Centre for Evaluation of Medicines,

McMaster University and completed while Bernie O'Brien was a Visiting Professor at the Centre for Health Economics Research and Evaluation, University of Sydney, Australia. We are grateful to both institutions for their hospitality. George Torrance, David Feeny and Bill Furlong all provided helpful comments and suggestions. We are grateful to the Steering Committee of the DINAMIT trial for granting early access to baseline utility data for this methodological study.

References

1. Ware JE, Donald C. The MOS 36-item short-form health survey (SF-36) conceptual framework and item selection. *Med Care* 1992; **30**: 473–481.
2. Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. (2nd edn) Oxford medical publications: New York; 1997.
3. Torrance GW. Measurement of health state utilities for economic appraisal: A review. *J Health Econ* 1986; **5**: 1–30.
4. Nichol MB, Sengupta N, Globe DR. Evaluating quality-adjusted life years: estimation of the Health Utility Index (HUI2) from the SF-36. *Med Decision Making* 2001; **21**: 105–112.
5. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BEK. Predicting Quality of Well-being scores from the SF-36: results from the Beaver Dam Health Outcome Study. *Med Decision Making* 1997; **17**(1): 1–9.
6. Keeney RL, Raiffa RL. *Decisions with Multiple Objectives. Preferences and Value Tradeoffs*. (1st edn). Cambridge University Press: Cambridge; 1993.
7. Feeny D, Furlong W, Boyle MH, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. *PharmacoEconomics* 1995; **7**(6): 490–502.
8. Kind P, Dolan P, Gudex C, Williams A. Variations in population health status: results from a United Kingdom national questionnaire survey. *BMJ* 1998; **316**(7133): 736–741.
9. Brazier JE, Usherwood T, Harper R, Thomas KJ. Deriving a preference-based single index from the UK SF-36 health survey. *J Clin Epidemiol* 1998; **51**(11): 1115–1128.
10. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ* 2002; **21**: 271–292.
11. Hohnloser SH, Connolly SJ, Kuck KH *et al*. The defibrillator in Acute Myocardial Infarction Trial (DINAMIT): study protocol. *Am Heart J* 2000; **140**: 735–739.
12. Feeny D, Furlong W, Torrance GW *et al*. Multi-Attribute and Single-Attribute Utility Functions for the Health Utilities Index Mark 3 System. Medical Care, 2002.
13. Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey: manual and interpretation guide*. The Health Institute, New England Medical Centre, Nimrod Press: Boston, MA, 1993.
14. Drummond MF. Introducing economic and quality of life measurements into clinical studies. *Ann Med* 2001; **33**: 344–349.
15. Samsa G, Edelman D, Rothman M, Rhys Williams G, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: A general approach with illustration using the health utilities index Mark II. *PharmacoEconomics* 1999; **15**: 141–155.

