Title:	**The Interchangeability of Utility Measures: a Systematic Review of Contemporaneous EQ-5D and SF-6D Group Mean Scores**
Running Head:	Group-Level Interchangeability of the EQ-5D and SF-6D
Keywords:	EQ-5D, SF-6D, utility, preference-based measures of health, review
Word Count:	3979
Table Count:	2
Figure Count:	4

Authors: Mr David G T Whitehurst (1,2)
Prof. Stirling Bryan (2)
Dr Martyn Lewis (1)

Affiliations: (1) Arthritis Research Campaign National Primary Care Centre, Primary Care Sciences, Keele University, Staffordshire, UK
(2) Department of Health Economics, University of Birmingham, Birmingham, UK

Contact Information: Mr David Whitehurst
Arthritis Research Campaign National Primary Care Centre,
Primary Care Sciences,
Keele University,
Staffordshire,
ST5 5BG,
UK

Telephone number: +44 (0) 1782 584711
Fax number: +44 (0) 1782 583911
E-mail: d.whitehurst@cphc.keele.ac.uk

Abstract

Aims

To be suitable for analysis in a cost-utility study, a health outcome measure must provide a single index value that reflects preferences for different health states. Currently, a number of measures are routinely used in economic evaluations. However, the EQ-5D and SF-6D are the only preference-based measures that have utility scores derived from a representative sample within the UK. Reviews of health status measures have highlighted the need for head-to-head comparisons across a range of clinical conditions in order to consider the interchangeability of different instruments and the associated implications for cost-effectiveness evidence.

The objective of this review is to explore the interchangeability of the EQ-5D and SF-6D health state classification questionnaires using contemporaneous group mean scores identified through a systematic search of 13 electronic databases. Previous comparative research indicates that there may be a systematic pattern to the observed differences in patient-level utility estimates, with the EQ-5D providing higher scores for milder health states and the SF-6D providing higher scores for more severe health states. This review explores this common conception in a systematic manner, drawing on group level mean data.

Methods

The review was restricted to papers published in peer-reviewed journals between January 2002 and October 2007. The search strategy consisted of the following terms; (EQ5D OR EQ-5D OR EuroQ$) AND (SF6D OR SF-6D), where $ represents a truncation facility. Given the paucity of high-quality research of alternative preference-based outcome measures, a broad selection criteria was used in order to incorporate a wide range of what may be considered 'comparative' research. All (sub-)group comparisons identified in the retrieved papers were included, provided arithmetic mean scores were identifiable from the text, tables or figures presented.

Analysis of the retrieved mean EQ-5D and SF-6D scores consisted of 4 parts, (i) an examination of the distribution of scores, (ii) assessment of correlation and agreement statistics, (iii) exploration of the patterns of differences in mean scores using cut-off values, and (iv) separate consideration of 'change' scores, where baseline and follow-up data were reported (NUMBER 4 IS NO LONGER BEING PURSUED).

Data & Results

Thirty papers were identified from 89 unique abstracts, which provided data on 523 group and sub-group comparisons. Despite high values for both correlation and absolute agreement statistics, the graphical depiction of the paired values highlighted the hypothesised systematic variation in group level scores. A number of further analyses were undertaken to deal with clustering issues (i.e. 523 paired data values from only 30 studies), with the same trends identified. Further analysis is ongoing.

Conclusions

Results from this systematic review raise concerns about the cross-study comparability of cost-effectiveness studies that differ in their choice of utility measure.

1. Introduction

1.1. Background

Economic evaluation is playing an increasing role in the health policy decision-making process. In the United Kingdom (UK), all technology appraisals considered by the National Institute for Health and Clinical Excellence (NICE) are required to incorporate an economic evaluation undertaken by an independent academic centre (Raftery, 2001). Similar mandates exist in Australia (Commonwealth Department of Human Services and Health, 1995) and the Canadian Province of Ontario (Canadian Coordinating Office for Health Technology Assessment, 1997) regarding the public subsidy of health care products, along with authoritative guidelines regarding the conduct of economic evaluations (Gold et al, 1996; Ziekenfondsraad, 1999). An inevitable consequence of these explicit requirements is the need for strong, evidence-based, methodological techniques to address areas of uncertainty. Given the increasing support for economic evaluation within a cost-utility framework, one such area of uncertainty concerns the role of preference-based health state classification questionnaires.

Traditionally, health-related quality-of-life (HRQoL) measures are classified as either disease-specific or generic (Guyatt et al, 1993). Disease-specific measures are considered to be more sensitive to subtle changes in health, although they have limited generalisable qualities because they do not permit comparisons to be made across conditions. Conversely, generic measures are purposely designed for use across different populations. In concordance with published guidelines, preference-based measures are typically a subset of generic health instruments (NICE, 2004). This is because the primary objective of an economic evaluation is to assist in the broader decision-making process, where there is responsibility for budgetary decisions regarding health care provision for a multitude of conditions. However, disease-specific preference-based measures have been developed in some clinical areas (Torrance et al, 2004; Brazier et al, 2005).

To be suitable for analysis in a cost-utility study, an outcome measure must provide a single index value that reflects respondents' preferences for different health states. Preferences can be measured directly using preference-elicitation techniques such as time trade-off, standard gamble, rating scales, magnitude estimation or person trade-off, although such approaches are rare within clinical research because of the logistical and financial restrictions of administering face-to-face interviews and/or interactive computer programs (Dolan, 2001;

Kopec and Willison, 2003). An alternative approach is to indirectly capture health state preference values, or utility scores, via preference-based health state classification questionnaires (also known as multi-attribute utility scales (MAUS), or preference-based HRQoL instruments). Utility scores are essential components within a cost-utility analysis as they are used to calculate quality-adjusted life years (QALYs), which have become the primary outcome measure for economic evaluations of health care technologies.

The descriptive systems of preference-based health state classification questionnaires define respondents' HRQoL as one of a finite number of health states. As such measures are developed for use across the complete spectrum of clinical conditions, the descriptive content must capture a broad range of health dimensions. Indeed, there is an implicit assumption that the dimensions and items included in a preference-based measure define the contents of an individual's utility function, i.e. the things the individual values (Brazier and Deverill, 1999). In addition to the descriptive system, classification systems have a procedure for scoring each health state defined by the questionnaire, usually based on community-derived preferences, which generates the single index score. These index scores fall on a scale where 1 indicates full health and 0 represents a health state equivalent to death. Negative scores can represent health states with a valuation considered to be worse than death. The purpose of the single index score is to represent the relative value that society places on living in different health states.

Many preference-based measures of health have been developed and evaluated since the late 1970s. Currently, a number of measures are routinely used in economic evaluations, such as the Health Utilities Index Mark II (HUI2) and Mark III (HUI3) (Furlong et al, 2001), the Assessment of Quality of Life (AQoL) instrument (Hawthorne et al, 1999), the EQ-5D (Brooks, 1996) and the SF-6D (Brazier *et al*, 2002; Brazier and Roberts, 2004). There is a growing literature that explores the consequences of methodological differences between alternative preference-based measures, such as the effect of using different elicitation techniques (Tsuchiya et al, 2006) or the relationship between patient and general population health state valuations (Ratcliffe et al, 2007). Reviews of health status measures have highlighted the need for head-to-head comparisons across an extensive range of clinical conditions in order to consider the interchangeability of different instruments (Brazier et al, 1999; McDonough et al, 2007), which has led to an ever-increasing body of evidence that

examines specific measurement properties of alternative preference-based instruments using patient-level data.

1.2. Preference-based measures: the EQ-5D and SF-6D

The EQ-5D was developed by an international group of researchers, The EuroQol Group, which initially comprised of seven research centres within Finland, The Netherlands, Norway, Sweden and the UK (Rabin and de Charro, 2001). The primary component of the EQ-5D is a 5-dimension self-classification system (covering mobility, self-care, usual activities, pain/discomfort and anxiety/depression), which can be self-completed or administered by an interviewer. Each dimension contains 3 levels, which provides 243 distinct health states that can be defined by a unique five-digit number. For example, state 22231 indicates level two (some problems) on the mobility, self care and usual activities dimensions, level three (extreme problems) on the pain/discomfort dimension and level 1 (no problems) for anxiety/depression. Health states for 'unconscious' and 'dead' have been added, resulting in a total of 245 health states.

Despite the number of valuation studies for the descriptive system, the EQ-5D element of this study focused on the UK-derived scoring algorithm, the so-called 'York A1 tariff', which was derived from the valuation study performed by the Measurement and Valuation of Health (MVH) group at the University of York (Dolan et al, 1995). This was the first large-scale EQ-5D valuation study and is still the most widely-used scoring algorithm (Räsänen, 2006). Preferences were elicited from a representative sample of 3395 members of the non-institutionalised UK adult population using time trade-off methodology. Econometric modelling was used to interpolate valuations for all EQ-5D health states based on the directly observed valuations for 42 states, which were chosen to ensure a sufficient spread across the valuation space. The resultant algorithm provides utility scores within a range of -0.594 (state 33333, the lowest level on each dimension) to 1.000 (state 11111, the highest level on each dimension).

The SF-6D is a preference-based health state classification questionnaires that can be derived from two widely-used generic health profile measures, the Short Form 36 (SF-36) (Ware et al, 1992) and the Short Form 12 (SF-12) (Ware et al, 1996), which assess HRQoL across 8 dimensions of health; physical functioning, role limitations (physical), bodily pain, general

health, vitality, social functioning, role limitations (emotional) and mental health. The SF-12 is an abridged version of the SF-36, with the 12 items selected on the basis of their psychometric performance.

The conventional scoring procedures for the SF-36 and SF-12 are not suitable for use in cost-utility studies because they do not incorporate information regarding preferences for different health states, nor do they provide a single index score. Given the widespread use of these instruments in clinical studies, researchers at the University of Sheffield (UK) conducted a valuation survey to derive scoring algorithms for estimating preference-based single index values from SF-36 and SF-12 data (Brazier et al, 1998; Brazier et al, 2002; Brazier and Roberts, 2004). As the term 'SF-6D' is used to represent the preference-based measure derived from either the SF-36 or SF-12, the two measures are denoted as the SF-6D (SF-36) and SF-6D (SF-12) throughout this paper, where appropriate.

The SF-6D comprises of 6 dimensions (following the exclusion of the general health dimension and a combination of the two dimensions measuring role limitations), which is constructed using 11 items from the 36-item instrument for the SF-6D (SF-36) and 7 items from the 12-item instrument for the SF-6D (SF-12), each with between 2 and 5 levels of severity. These descriptive systems define 18,000 and 7,500 health states for the SF-6D (SF-36) and SF-6D (SF-12) respectively. Preferences for the scoring function were measured using standard gamble methodology, based on a representative sample of members of the non-institutionalised UK adult population (n=611). As with the EQ-5D, econometric modelling was used to interpolate valuations for all SF-36 and SF-12 health states based on directly observed valuations for 249 health states. The range of index values covered by the scoring algorithms for the SF-6D (SF-36) and SF-6D (SF-12) are 0.301 to 1.000 and 0.345 to 1.000 respectively. Although the minimum values permitted by the SF-6D do not include zero, the scoring range is still interpreted on a scale where 1 indicates full health and 0 represents a health state equivalent to death.

The comparability of utility scores is compromised, to some extent, because of the sample of participants used in their respective valuation study. Unsurprisingly, such samples correspond to the country where an instrument was developed, such as the Health Utilities Index (Canada) and the Assessment of Quality-of-Life (Australia) instruments. The EQ-5D and SF-6D are the only widely-used preference-based measures that have utility scores derived from a

representative sample within the same country - the United Kingdom (UK). Previous research has addressed the proposition that healthier states tend to have higher utility scores according to the EQ-5D, with the SF-6D providing higher utility scores for poorer health states (Joore and Brunenberg, 2007). Using data from 43 different populations reported across 19 articles, and an arbitrarily defined cut-off of 0.64 on the EQ-5D to distinguish between 'moderate and poor' health states and 'good' health states, Joore and Brunenberg concluded that the SF-6D consistently provided higher utility scores for moderate and poor health states, while there was no systematic difference between the two measures for relatively good health states.

The objective of this study was to further explore the relationship between contemporaneous EQ-5D and SF-6D group mean scores reported in the literature and consider the implications regarding the interpretation and comparability of economic evaluations that employ different utility measures. Ultimately, it is group mean data that drives the cost and QALY components of an analysis. Therefore, the interchangeability of group mean data is an important consideration when addressing the practical implications of using different measures.

2. Methods

2.1. Review Methodology

Studies were identified via a systematic search of 13 online electronic databases: Medline, PsycINFO, EMBASE, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Allied and Alternative Medicine (AMED), EconLit, the Cochrane Library (Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register (CMR), NHS Economic Evaluation Database (NHSEED), and the Health Technology Assessment database (HTA)), and the database held by the Patient-reported Health Instruments (PHI) group at Oxford University (http://phi.uhce.ox.ac.uk). In addition to these searches, the reference lists of included papers were checked and key journals were hand searched to identify papers that may have been overlooked due to inaccurate indexing (*Quality of Life Research*, *Health Economics*, *The Journal of Health Economics* and *Health and Quality of Life Outcomes*).

The search strategy was designed to acknowledge the heterogeneity of terms used to report the EQ-5D and SF-6D, such as the common error of reporting the EQ-5D and SF-6D without a hyphen and misspellings of the 'EuroQol' search term. Accordingly, the two components of the search strategy were: (i) EQ5D **OR** EQ-5D **OR** EuroQ$ **AND** (ii) SF6D **OR** SF-6D, where $ represents a truncation facility to account for variations of the search term. The strategy was modified, where appropriate, to satisfy the specific requirements of host databases.

Given the paucity of high-quality research of alternative preference-based outcome measures (Kopec and Willison, 2003), a broad selection criteria was used in order to incorporate a wide range of data sources. A two-stage study selection process was employed. The first stage involved the identification of papers that reported original research (i.e. not review articles), written in English and published post-2001. In addition, there had to be an indication, within the title or abstract, of empirical data regarding actual (i.e. non-hypothetical) health state valuations for both the EQ-5D and SF-6D. The timeframe for the review consisted of papers published in peer-reviewed journals between January 2002 and October 2007 in order to coincide with the publication year for the first population-based scoring algorithm for the SF-6D (Brazier et al, 2002). Although both the publication date and language restrictions could have been explicitly incorporated into the search strategy, it was considered unnecessary to impart such restrictions at the search stage given the anticipated number of identified papers.

Upon completion of the first stage, which was performed independently by two of the authors (DGTW and ML), full text versions of the remaining articles were obtained. The purpose of the second stage was to verify that the initial criteria had been met and to establish whether the EQ-5D health state scores presented in the articles were derived using the York A1 tariff. This latter requirement was considered inappropriate for the first stage as it is unlikely that such detail would be apparent in an abstract. The reasons for exclusion were documented at each stage. Studies retrieved through the bibliographic searches were assessed by the lead author only (DGTW).

2.2. Data extraction and statistical analysis

All (sub-)group comparisons were included in the statistical analysis, provided contemporaneous arithmetic mean EQ-5D and SF-6D health state scores were identifiable

from the text, tables or figures presented in the retrieved papers. Data extraction for this phase of the review was performed by the lead author (DGTW).

Analysis of the retrieved mean scores consisted of 3 parts; (i) an examination of the descriptive statistics and distribution of scores, (ii) assessment of correlation and agreement statistics and (iii) exploration of the patterns of differences in mean scores using cut-off values. Given the broad entry criteria, it was anticipated that multiple comparisons would be retrieved from single studies, e.g. authors may provide EQ-5D and SF-6D data on specific subgroups as well as the total sample. Sensitivity analyses were performed in order to consider the influence of a 'clustering' effect, where the same individuals are contributing to a number of subgroup comparisons in the analysis. Pre-planned sensitivity analyses involved the assessment of only one group comparison per study (based on the comparison with the largest sample size) and the exclusion of all comparisons with sample sizes less than 50.

The distributions of the retrieved mean EQ-5D and SF-6D scores were examined and checked for normality using visual inspection and interpretation of frequency distribution histograms, box plots and normal probability plots. To test the strength of the relationship between scores, Pearson's and Spearman's correlation coefficients were considered. Pearson's correlation coefficient is a parametric measure of the strength and direction of a linear relationship between two variables, which requires both variables to be approximated by a normal distribution. Spearman's correlation coefficient is a non-parametric measure that assesses how well an arbitrary monotonic function can describe the relationship between two variables. This statistic does not require the assumption of normally distributed variables and is often appropriate for analysing EQ-5D datasets of patient-level data, which typically exhibit negatively-skewed or bimodal distributions. For HRQoL measures, it has been suggested that parametric statistical methods are robust to violations of the normality assumption and that further research is necessary to support this claim (Walters and Campbell, 2004). Accordingly, both correlation coefficients were calculated.

Establishing a monotonic relationship between utility scores, whether linear or not, does not mean that two measures 'agree' (Bland and Altman, 1986), meaning that evidence of strong correlation is of limited benefit when considering the interchangeability of measures for use in economic evaluation. A high level of agreement implies a high correlation coefficient but a high correlation coefficient does not imply a high level of agreement. Assessment of the level

of agreement between the mean EQ-5D and SF-6D scores was based on graphical and statistical approaches, using Bland and Altman plots (Bland and Altman, 1986) and the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979).

Simple scatter diagrams and lines of equality (45^o lines) would provide an indication of the level of agreement between two sets of scores and identify any systematic trends in the data. However, a plot of the difference between two sets of scores and their mean, known as a Bland and Altman plot, will be more informative as it allows for the identification of a relationship between the measurement error and the best estimate of the true value, i.e. the mean of the two measurements. For the quantification of absolute agreement, a single measure ICC based on a two-way mixed analysis of variance model was calculated, where the two outcome measures are treated as a source of variability. For this review, the following benchmarks were used to interpret correlation co-efficients; 0.00 to 0.10 representing virtually no correlation/agreement, 0.11 to 0.40 slight, 0.41 to 0.60 fair, 0.61 to 0.80 moderate and 0.81 to 1.00 substantial correlation/agreement (Shrout, 1998).

Given the absence of objectively-derived published cut-off values for categorising preference-based utility scores into mild/moderate/severe groups, the retrieved mean scores were examined with respect to the 'cross-over' value derived by Barton et al. (2006), which estimated the point (on the EQ-5D) below which scores on the SF-6D would be higher than scores on the EQ-5D. A value of 0.754 was predicted from their regression analysis, using responses from 1865 individuals aged ≥ 45 years in one UK general practice. The mean EQ-5D and SF-6D data retrieved in this review were used to assess the predictive ability of this cross-over value. Although Barton et al. attempted to validate their cross-over point with data previously reported in the literature, their strategy for identifying and reporting comparative utility scores was narrower than the methodologies adopted in this review.

3. Results

3.1. Database and bibliographic searches

Table 1 reports the results of the database and bibliographic searches. Eighty-nine unique abstracts were identified via the search strategy, which resulted in 32 initial inclusions

following application of the 2-stage selection criteria. A further xx papers were identified from the bibliographic searches (yet to be done). Columns 3 and 5 in Table 1 identify the unique contributions of each database, which are treated in a hierarchical manner based on the number of abstracts identified in the initial search (for example, 3 of the 15 titles identified in the CINAHL database were not present in Medline, PsycINFO or EMBASE). For this review, all included papers were identified from Medline.

The 32 initial papers provided data on 538 (sub-)group comparisons. Only 2 of the 32 papers reported utility scores for the SF-6D (SF-12). Due to the relatively small proportion of SF-6D (SF-12) data (3% of the 538 comparisons), and in order to ensure the analysis included 'like-for-like' comparisons, the review focused on SF-6D (SF-36) data only. The number of comparisons per study varied between 1 and 231. A breakdown of the study selection process and reasons for exclusion are presented as a flow diagram in Figure 1.

3.2. Frequency distributions, correlation and agreement

The frequency distributions of the retrieved mean scores are shown in Figure 2. SF-6D scores followed a normal distribution, while the EQ-5D scores were negatively skewed. A major reason for the marked difference in the distributions is because of the variability in the range of possible scores on the instruments. These differences in the distributions were also clear from inspection of box plots and normal probability plots (not shown).

Table 2 provides details of descriptive, correlation and agreement statistics for the 523 retrieved comparisons. On average, the SF-6D generated a higher utility score than the EQ-5D, although the reverse was found for median values. The non-parametric (Spearman's ρ) and parametric (Pearson's *r*) correlation coefficients both indicated 'substantial' correlation, with values of 0.940 and 0.920 respectively. The two-way mixed intraclass correlation coefficient was 0.757, which suggests 'moderate' agreement between the two measures. Graphical depiction of absolute agreement identified a systematic pattern between the scores, where a low (high) average of the paired scores was associated with lower (higher) EQ-5D scores (see Figure 3).

3.3. Assessment of patterns of differences

Figure 4 shows a ranked scatter graph to assess the patterns of differences between EQ-5D and SF-6D scores. The 523 comparisons were ranked by their EQ-5D and this ranking was then plotted against the corresponding utility estimates from the EQ-5D and SF-6D. This approach identified a relationship between scores, showing a systematic pattern at both ends of the utility scale. For poorer health states, the SF-6D provides higher utility estimates, whereas the EQ-5D provides higher estimates for healthier states. In addition to the pre-planned sensitivity analyses, the same analysis was repeated after eliminating the study that contributed the highest number of group mean comparisons (231 comparisons were provided by Petrou and Hockley, 2005). In all sensitivity analyses the same patterns were identified using ranked scatter graphs (not shown). Selecting to rank the comparisons by their EQ-5D score was an arbitrary choice. To ensure greater rigour in the analysis and to enhance the robustness of the results, the same analyses were repeated after ranking the comparisons by their SF-6D score and the same trends were identified (not shown).

The EQ-5D and SF-6D provided the same utility estimate in 11 (2%) of the retrieved comparisons. Of the remaining 512, there were 240 with EQ-5D scores above the cross-over value derived by Barton *et al* (2006). For this subset, Barton's prediction rule states that the EQ-5D would provide a higher score than the SF-6D, which was correct in 236 of the 240 paired data values (98%). For the 272 comparisons with EQ-5D scores below the cross-over value, 190 (70%) were in agreement with the prediction that the SF-6D would provide a higher score than the EQ-5D.

4. Discussion

This study has shown the existence of important differences between group mean scores generated by the EQ-5D and SF-6D at both ends of the utility scale. The phenomenon is more evident at the lower end of the scale because of the differences in the range of scores attainable from the two instruments.

The units of analysis in this study are means scores as opposed to individual-level data, which could be seen as a limitation when considering the interpretation of conventional analytical

approaches. The analysis of 523 data pairs took no account of the sample sizes of the respective group comparisons, i.e. each comparison had equal weight in the analysis regardless of whether it consisted of 10 respondents or 1,000 respondents. A second potential limitation concerns the possibility of 'clustering', where the same individuals were present in multiple comparisons within the same study. A number of sensitivity analyses were performed to address this. The same trends were identified after eliminating comparisons based on small samples, considering only one comparison per study, and after excluding a potentially dominant study.

High correlation statistics between utility scores provides little, if any, evidence to support the interchangeability of different instruments. Alternatively, it is considered more appropriate to quantify absolute agreement using the intraclass correlation coefficient. The ICC value in this study indicated moderate agreement between scores. However, over reliance on this statistical methodology would have failed to identify the important differences highlighted by the Bland and Altman plot (Figure 3). For future research in this area it is important to acknowledge the different information that numerical and graphical analysis can provide.

Previous research has indicated that the SF-6D provides higher utility scores for poorer states, i.e. towards the lower end of the utility scale. Therefore, the main contribution of the current study was to formally demonstrate the existence of important differences between EQ-5D and SF-6D group mean estimates for milder health states. These findings raise serious concerns about the cross-study comparability of economic evaluations that base their utility measurement on different standardised measures. Given that utility scores are principally used to generate QALYs, the next stage in this area of research is to investigate whether the systematic differences manifest as important differences in QALY gains for competing interventions.

Table 1: Results of the systematic and bibliographic searches

Search mechanism	No. of abstracts identified	Unique abstracts identified [a]	No. of articles included in the review	Unique articles included in in the review [a]
Databases				
Medline	54	54	32	32
EMBASE	47	1	26	0
PsycINFO	45	29	10	0
CINAHL	15	3	9	0
Cochrane - NHS EED	11	1	7	0
EconLit	10	0	7	0
AMED	5	0	3	0
PHI Bibilography	8	0	5	0
Cochrane - CENTRAL	7	0	6	0
Cochrane - CMR	4	0	1	0
Cochrane - CDSR	1	1	0	0
Cochrane - HTA	1	0	1	0
Cochrane - DARE	0	0	0	0
Bibliographic search				
Retrieved articles	-	-	-	???
Key journals	-	-	-	???
Total	208	89	-	32

[a] Databases treated in a hierarchical manner as listed in the table

Table 2: Descriptive, correlation and agreement statistics for the 523 retrieved comparisons

Statistic	EQ-5D	SF-6D
Mean	0.683	0.696
Median	0.731	0.704
Minimum	-0.084	0.405
Maximum	0.981	0.900
Interquaritle range	0.258	0.168
Pearson's *r*	0.920	
Spearman's ρ	0.940	
ICC (two-way mixed)	0.757	

Figure 1: Flow diagram to describe the study selection process

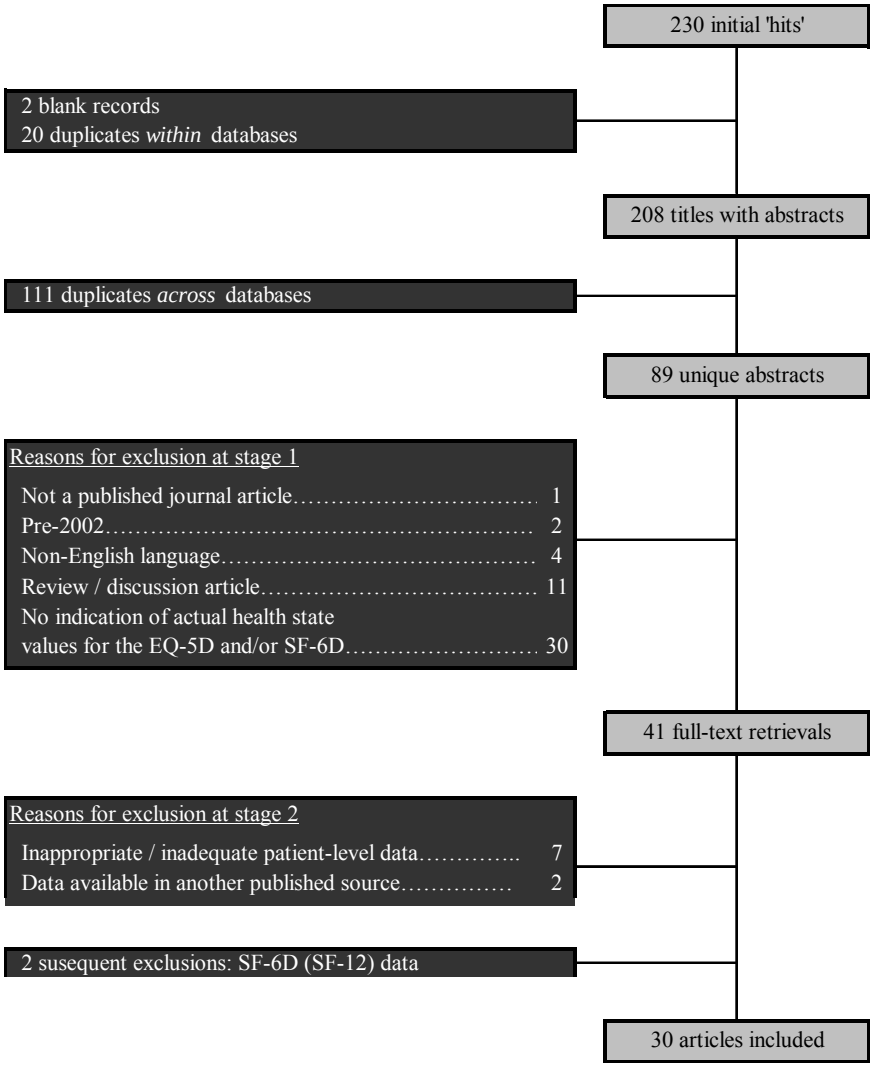


Figure 2: Frequency distributions of the retrieved mean EQ-5D and SF-6D scores from the 523 (sub-)group comparisons

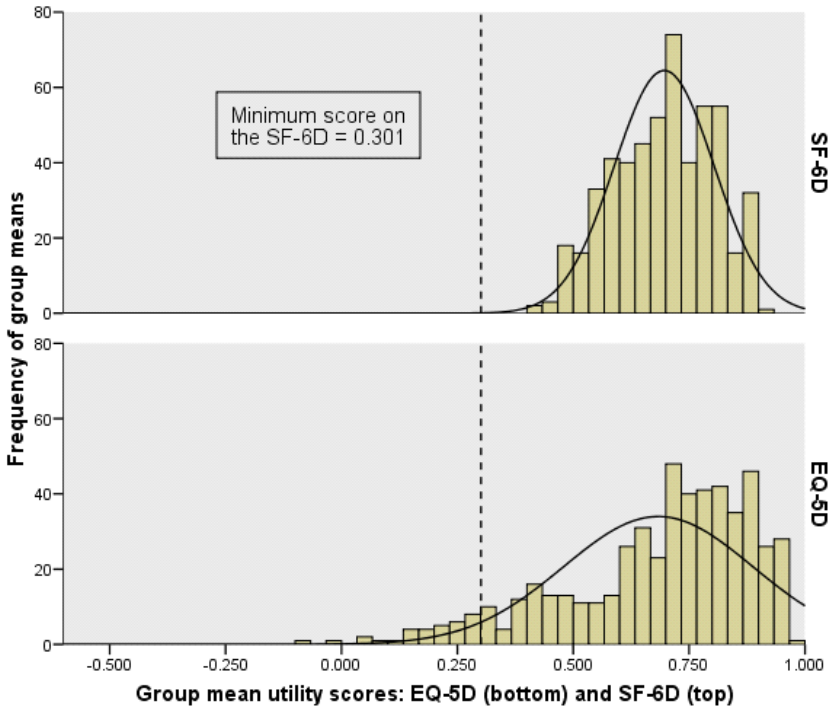


Figure 3: Bland and Altman plot, with limits of agreement, for the 523 paired EQ-5D and SF-6D scores

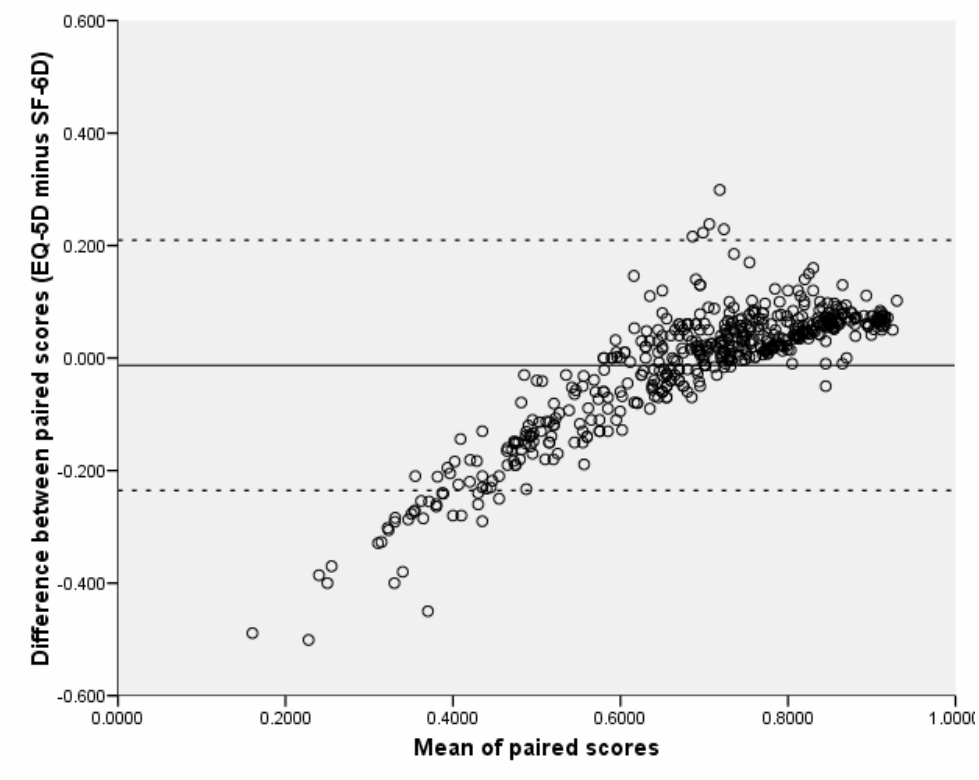
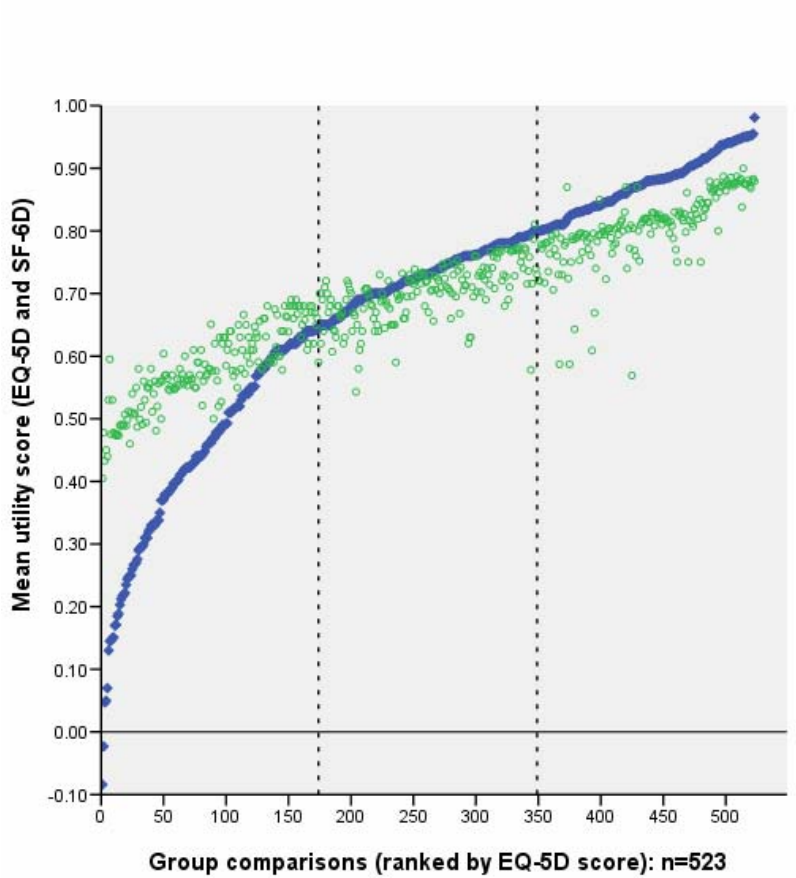


Figure 4: Ranked scatter graph showing the paired EQ-5D and SF-6D scores (ranked by EQ-5D score)



References

Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, Muir KR. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >/= 45 years. Health Econ. 2007; [Epub ahead of print]

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476): 307-10

Brazier J, Usherwood T, Harper R, Thomas K. Deriving a Preferecnce-Based Single Index from the UK SF-36 Health Survey. J Clin Epidemiol. 1998; 51(11): 1115-28

Brazier J, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. Health Econ. 1999; 8(1): 41-51

Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technol Assess. 1999; 3(9): 1-164

Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002; 21(2): 271-92

Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004; 42(9): 851-9

Brazier JE, Roberts J, Platts M, Zoellner YF. Estimating a preference-based index for a menopause specific health quality of life questionnaire. Health Qual Life Outcomes. 2005; 3: 13

Brooks R. EuroQol: the current state of play. Health Policy. 1996; 37(1): 53-72

Canadian Co-ordinating Office for Health Technology Assessment (CCOHTA). Guidelines for economic evaluation of pharmaceuticals: Canada. 2nd ed. Ottawa: CCOHTA; 1997

Commonwealth Department of Human Services and Health. Guidelines for the pharmaceutical industry on preparation of submissions to the Pharmaceutical Benefits Advisory Committee: including major submissions involving economic analyses. Canberra: Australian Government Publishing Service; 1995

Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: results from a UK general population survey. York: Centre for Health Economics (Discussion Paper No. 138); 1995

Dolan P. Output measures and valuation in health. In: Drummond MF, Maguire A, editors. Economic evaluation in health care: merging theory and practice. New York: Oxford University Press; 2001

Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Ann Med. 2001; 33(5): 375-84

Gold M, Siegel J, Russell L, Weinstein M. Cost-effectiveness in health and medicine. New York: OUP; 1996

Guyatt GH, Feeny DH, Patrick DL. Measuring health related quality of life. Ann Internal Med. 1993; 118(8): 622–629

Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual Life Res. 1999; 8(3): 209-24

Joore MA, Brunenberg D. Exploring patterns of differences between EQ-5D and SF-6D: a systematic review. Presented at the 6th International Health Economics Association World Congress; Copenhagen, 2007

Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol. 2003; 56(4): 317-25

McDonough CM, Tosteson AN. Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. Pharmacoeconomics. 2007; 25(2): 93-106

National Institute for Health and Clinical Excellence. Guide to the Methods of Technology Appraisal. London: National Institute for Health and Clinical Excellence; 2004

Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. Health Econ. 2005; 14(11): 1169-89

Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001; 33(5): 337-43

Raftery J. NICE: faster access to modern treatments? Analysis of guidance on health technologies. BMJ. 2001; 323(7324): 1300-3

Räsänen P, Roine E, Sintonen H, Semberg-Konttinen V, Ryynänen OP, Roine R. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. Int J Technol Assess Health Care. 2006; 22(2): 235-41

Ratcliffe J, Brazier J, Palfreyman S, Michaels J. A comparison of patient and population values for health states in varicose veins patients. Health Econ. 2007; 16(4): 395-405

Shrout PE. Measurement reliability and agreement in psychiatry. Stat Methods Med Res. 1998; 7(3): 301-17

Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bulletin. 1979; 86: 420-427

Torrance GW, Keresteci MA, Casey RW, Rosner AJ, Ryan N, Breton MC. Development and initial validation of a new preference-based disease-specific health-related quality of life instrument for erectile function. Qual Life Res. 2004; 13(2): 349-59

Tsuchiya A, Brazier J, Roberts J. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. J Health Econ. 2006; 25(2): 334-46

Walters SJ, Campbell MJ. The use of bootstrap methods for analysing Health-Related Quality of Life outcomes (particularly the SF-36). Health Qual Life Outcomes. 2004; 2: 70

Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996; 34(3): 220-33

Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30(6): 473-83

Ziekenfondsraad. Dutch guidelines for pharmacoeconomic research. Amstelveen: Health Insurance Council (Ziekenfondsraad); 1999