HEALTH ECONOMICS
Health Econ. **17**: 815–832 (2008)
Published online 24 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hec.1298

A COMPARISON OF THE PERFORMANCE OF THE EQ-5D AND SF-6D FOR INDIVIDUALS AGED $\geq$ 45 YEARS

GARRY R. BARTON[a,b,*], TRACEY H. SACH[c,d], ANTHONY J. AVERY[c], CLAIRE JENKINSON[c], MICHAEL DOHERTY[e], DAVID K. WHYNES[a] and KENNETH R. MUIR[c]

[a] *School of Economics, University of Nottingham, Nottingham, UK*
[b] *Health Economics Group, School of Medicine, Health Policy and Practice, University of East Anglia, Norwich, UK*
[c] *School of Community Health Sciences, University of Nottingham, Nottingham, UK*
[d] *School of Chemical Sciences and Pharmacy, University of East Anglia, Norwich, UK*
[e] *Academic Rheumatology, University of Nottingham, Nottingham, UK*

SUMMARY

We sought to compare the performance of the EQ-5D and SF-6D with regard to the criteria of practicality, convergent validity, and construct validity, the level of agreement between the two measures was also assessed. Responses from 1865 individuals aged $\geq$ 45 years in one general practice were analysed. Of these, 93.1% completed the EQ-5D, compared with 86.4% for the SF-6D, where individuals who were older, female, of a lower occupational skill level, from an area of lower deprivation, or used prescribed medication were significantly less likely to complete the SF-6D. The performance of both measures was comparable with regard to both convergent and construct validities, as both the EQ-5D and SF-6D scores were closely related to scores on the EuroQol visual analogue scale (VAS) ($p<0.001$) and able to discriminate between people who did and did not take: (i) analgesics and (ii) other prescribed medication. Despite EQ-5D and SF-6D scores being highly correlated ($p<0.001$), individuals who were healthier (according to the VAS) had higher mean scores on the EQ-5D ($p<0.001$), whereas less healthy individuals had higher mean scores on the SF-6D (individuals with knee pain, osteoarthritis, back pain, rheumatoid arthritis, and hip pain had significantly lower mean scores on the EQ-5D, $p<0.001$). Copyright © 2007 John Wiley & Sons, Ltd.

Received 22 November 2005; Revised 24 July 2007; Accepted 2 August 2007

KEY WORDS: EQ-5D; SF-6D; validity; outcome measurement and valuation

INTRODUCTION

There are now a number of utility measures that can be used to estimate and compare the benefits of different health care interventions, including the EuroQol EQ-5D (EQ-5D) (Brooks, 1996) and SF-6D (Brazier *et al.*, 2002) (which is derived from responses to the SF-36 questionnaire, Ware and Sherbourne, 1992). Measures of utility attach a score to a health state description, where 0 corresponds to death and 1 to full health (Drummond *et al.*, 2005), but as different utility measures are based on different health descriptions (Brazier *et al.*, 2004) and different valuation methods (Tsuchiya *et al.*, 2006), they can assign different utility scores to the same individual. A number of studies have thereby been undertaken to compare EQ-5D and SF-6D scores for patients with a particular clinical condition (Conner-Spady and Suarez-Almazor, 2003; Longworth and Bryan, 2003; Barton *et al.*, 2004; Brazier *et al.*, 2004; Gerard *et al.*, 2004; Szende *et al.*, 2004; Fisk *et al.*, 2005; Marra *et al.*, 2005; McDonough *et al.*, 2005; Petrou and Hockley, 2005; Pickard *et al.*, 2005; Stavem *et al.*, 2005; Michaels *et al.*, 2006),

*Correspondence to: Health Economics Group, School of Medicine, Health Policy and Practice, University of East Anglia, Norwich NR4 7TJ, UK. E-mail: g.barton@uea.ac.uk.

where a common finding is that there are small, but important, differences between the utility estimates of the two measures (Bryan and Longworth, 2005).

In contrast to many of these previous studies, within this paper, we seek to compare the relative performance of the EQ-5D and SF-6D with regard to practicality, convergent validity, and construct validity. These are criteria which influence the decision of which outcome measure to use (Fitzpatrick *et al*., 1998; Brazier and Deverill, 1999; Brazier *et al*., 1999), and such an examination will also allow one to estimate the degree of confidence we should place on scores derived from the EQ-5D and SF-6D (Streiner and Norman, 2003). By comparing the results of the EQ-5D and SF-6D across individuals with a wide range of socio-demographic factors, and different clinical conditions, we also seek to identify when, and by what extent, EQ-5D and SF-6D scores are likely to differ. Throughout the paper we often conduct analyses that others have used to compare different measures of utility (Lubetkin and Gold, 2003; Brazier *et al*., 2004; Gerard *et al*., 2004; Bryan and Longworth, 2005; Lamers *et al*., 2006), where such an approach is justified by the argument that, in this area, the burden of evidence arises from a series of converging experiments, not a single experiment (Streiner and Norman, 2003).

Gerard *et al*. (2004) assessed the criterion of practicality for the EQ-5D and SF-6D (assessed in terms of completion rates) in a group of patients undergoing hospital haemodialysis for end-stage renal failure. It was found that the completion rate was lower for the SF-6D, and that increasing age, co-morbidity, and blindness were associated with a decline in the completion rate of the SF-6D, which was argued to pose a threat to the extent to which results based on the SF-6D can be generalised (Gerard *et al*., 2004). By seeking information on a wide range of socio-demographic factors (age, gender, ethnicity, smoking status, body mass index (BMI), occupation, and level of deprivation) and clinical conditions (back pain, knee pain, hip pain, heart disease, stroke, diabetes, cancer, asthma, osteoarthritis, and rheumatoid arthritis), for all individuals aged $\geq$ 45 years, who were registered in one general practice, we assess whether these results apply in different population groups, and whether other individual characteristics are also associated with a reduced completion rate for either the EQ-5D or SF-6D. Indeed, assessing whether outcome measures are acceptable has been argued to be essential, particularly as it has been far less frequently examined than issues such as reliability and validity (Fitzpatrick *et al*., 1998). Moreover, such results may also contribute to the current debate about whether to expand the number of responses within each of the five dimensions of the EQ-5D from 3 levels to 5 levels (Lamers, 2006), where the increased descriptive ability may result in a reduced response rate.

Marra *et al*. (2005) assessed both the convergent and construct validities (Streiner and Norman, 2003) of the EQ-5D and SF-6D. They found that both measures had moderate to strong correlations with many continuous measures of rheumatoid arthritis severity, and that patients in different dichotomous rheumatoid arthritis severity groups had different estimated utility scores, thereby concluding that both measures perform well with regard to these criteria (Marra *et al*., 2005). In this paper we use similar methods to assess the convergent and construct validities of the EQ-5D and SF-6D with regard to indicators of health status.

Brazier *et al*. (2004) focused on the assessment of the level of agreement between the EQ-5D and SF-6D across seven different patient groups, by estimating the intraclass correlation coefficient (ICC) and the relationship between the two measures using different ordinary least squares (OLS) regressions. They concluded that there were significant differences in agreement across patient groups and that there was a need to examine the implications for estimates of the impact of health care interventions (Brazier *et al*., 2004). Again we seek to ascertain whether their findings also apply to other groups, and to address the latter point we assess whether the differences between the EQ-5D and SF-6D scores can be considered to constitute a minimally important difference (MID) (Jaeschke *et al*., 1989). We considered a change of 0.03 to be a MID as this is equivalent to the mean change in the SF-6D that was associated with a reported change in general health in nine reviewed studies (Walters and Brazier, 2003) and approximately equivalent to a previous estimate of the MID for the EQ-5D (Sullivan *et al*., 2005),

DOI: 10.1002/hec

although we acknowledge that different estimates of the MID have been made for both the EQ-5D (Marra *et al*., 2005; Sullivan *et al*., 2005; Walters and Brazier, 2005) and SF-6D (Walters and Brazier, 2005).

Bryan and Longworth (2005) also compared the level of agreement between the EQ-5D and SF-6D and, informed by the Bland–Altman plots (Bland and Altman, 1986) for a group of liver transplant patients, they argued that scores on the SF-6D were consistently higher for very poor states (with a mean value below 0.4), but that scores on the EQ-5D tended to be higher for healthier states. Acknowledging that the different valuation techniques are not the only reason as to why the scores of the EQ-5D (which is based on the time trade-off (TTO) technique, Torrance, 1986) and SF-6D (which is based on the standard gamble (SG) technique, Dolan *et al*., 1996) may differ (Brazier *et al*., 2004; Tsuchiya *et al*., 2006), it may be that the different valuation techniques provide a partial explanation for such an occurrence. Namely that as TTO scores have been shown to be higher for milder states (Dolan *et al*., 1996), and SG scores have been shown to be higher for more severe states (Dolan *et al*., 1996), EQ-5D scores may also tend to be higher for milder states, with SF-6D scores tending to be higher for more severe states, with a 'cross-over' of TTO (EQ-5D) and SG (SF-6D) values occurring at a certain utility value (Tsuchiya *et al*., 2006). Thus, in addition to seeking to confirm that, for other patient groups, healthier individuals tend to have higher scores on the EQ-5D and less healthy individuals higher scores on the SF-6D, we also assess whether these differences are significant and could be considered to constitute a MID, and seek to ascertain where the cross-over point of scores on the EQ-5D and SF-6D might be.

The structure of the paper is as follows. We first outline the methods used to obtain information from individuals and then provide further information about the EQ-5D and SF-6D. The remainder of the paper is then set out in accordance with our outline of the previous literature, although it should be borne in mind that there is no agreed set of terminology within the literature (Fitzpatrick *et al*., 1998). We assess the practicality of the EQ-5D and SF-6D, look at the convergent and construct validities of both measures (with regard to indicators of health status), and then compare the level of agreement between the two measures (where we assess whether scores on the EQ-5D are significantly higher for healthier states and significantly lower for less healthy states and estimate the relationship between EQ-5D and SF-6D scores).

METHODS

Participants and procedures

As part of the recruitment for a study designed to assess the effectiveness and cost-effectiveness of different lifestyle interventions for knee pain (LIKP) all individuals aged $\geq$ 45 years, and registered in one UK general practice, were sent an ascertainment questionnaire, with the exception of those that were deemed (by their general practitioner) to be unable to complete information requested in a questionnaire. The general practice co-ordinated sending out the ascertainment questionnaires, and individuals were asked to complete the questionnaire and return it to the LIKP study team at the University of Nottingham.

For the purposes of this paper, for each respondent, including those who did not have knee pain, the data in the ascertainment questionnaire were used to collate information on the following variables, each of which are described in subsequent sections: (i) seven socio-demographic factors (age, gender, ethnicity, smoking status, body mass index (BMI), occupational skill level and the level of deprivation in the area), (ii) 10 clinical conditions (back pain, hip pain, knee pain, heart disease, stroke, asthma, cancer, diabetes, rheumatoid arthritis, and osteoarthritis), (iii) the two measures of utility (the EQ-5D and SF-6D), and (iv) three indicators of health status (their score on the EuroQol visual analogue scale (VAS), Brooks, 1996, and whether they used analgesics or any other prescribed medication).

DOI: 10.1002/hec

Socio-demographic factors. In order to compare the performance of the EQ-5D and SF-6D, across different socio-demographic groups individuals were categorised into one of four age groups: (i) 45 to <55 years, (ii) 55 to <65 years, (iii) 65 to <75 years, and (iv) ≥ 75 years. For ethnicity, due to the small number of individuals within non-white categories, individuals were re-coded as either white or non-white. Similarly, individuals were categorised according to whether they had, or had not, ever smoked regularly for a period of at least three months. In line with WHO recommendations (WHO, 2001), four BMI groups were created: (i) underweight ($<18.5\,kg/m^2$), (ii) normal (18.5 to $<25\,kg/m^2$), (iii) overweight (≥ 25 to $<30\,kg/m^2$), and (iv) obese ($\geq 30\,kg/m^2$). The occupational skill level of each individual was estimated by assigning the standard occupational classification (SOC2000) skill level (Office for National Statistics, 2000) to their current job title, or previous job title if their current job title was not reported (e.g. for those individuals who had retired). The SOC2000 categorises each job title into one of four skill levels (4 = highest and 1 = lowest), dependent upon the length of time deemed necessary for a person to become fully competent in the performance of tasks associated with such a job title (Office for National Statistics, 2000). We further categorised individuals who reported they had never worked to the lowest occupational skill level and those who reported they were either a housewife or househusband to the 2nd lowest occupational skill level (the same as a housekeeper). Finally, each individual's postcode was used to calculate the estimated level of deprivation in the area in which they live, according to the index of multiple deprivation (IMD) (Office of the Deputy Prime Minister, 2004), where a higher IMD score denotes that the area is estimated to be more deprived. The IMD score is a measure of multiple deprivation at the small area level, and it combines information on deprivation from seven domains: (i) income, (ii) employment, (iii) health and disability, (iv) education, skills and training, (v) barriers to housing and services, (vi) living environment, and (vii) crime (Office of the Deputy Prime Minister, 2004). However, it should be noted that the IMD score had a higher level of missing data than that for other variables as the relevant question appeared on the final page of the ascertainment questionnaire, and individuals were only requested to provide their address if they were willing to participate further in the study. For this reason the IMD score was divided into a binary category: (i) IMD <10 and (ii) IMD ≥ 10.

Clinical conditions. Individuals were deemed to have the conditions of back pain, hip pain, and knee pain if they reported that they had these conditions on most days of the last month. Similarly, they were deemed to have heart disease, stroke, asthma, cancer, diabetes, rheumatoid arthritis, and osteoarthritis if they reported that they had been diagnosed with them by their general practitioner.

Utility measures. Responses to the EuroQol (Brooks, 1996) and SF-36 questionnaires (Ware and Sherbourne, 1992) were used to calculate two measures of utility: the EQ-5D and SF-6D. The EuroQol questionnaire was developed by the EuroQol group (Brooks, 1996) and has two components – five-dimension questions (referred to as the EQ-5D) and a VAS/thermometer. In the previous section, the respondent is asked the level of problems they have (no problems, some/moderate problems, and severe/extreme problems) with regard to mobility, self-care, usual activities, pain, and anxiety/depression. Responses to these five dimensions can be converted into one of 243 different EQ-5D health states, which range between the best state of no problems on all five dimensions (11111) and the worst state of severe/extreme problems on all five dimensions (33333). The best state is often termed full health and the worst state as the pits (Brazier *et al*., 2002). Using any one of a number of regressions that are available (Dolan *et al*., 1995; Badia *et al*., 2001; Dolan and Roberts, 2002; Tsuchiya *et al*., 2002; Shaw *et al*., 2005), a (preference based) utility score can be assigned to each of the 243 EQ-5D health states. In the UK, the most commonly used regression (often referred to as the York A1 tariff) was based on the preferences elicited from a survey of 3395 residents using the TTO (Dolan *et al*., 1995). We thereby used this regression to assign a utility score to each of the potential 243 EQ-5D health states, where utility scores can range between -0.594 (the pits) and 1 (full health).

DOI: 10.1002/hec

Responses to 11 of the questions on the SF-36 questionnaire (Ware and Sherbourne, 1992) can be used to estimate a score on the SF-6D (Brazier *et al*., 2002). The SF-6D is composed of six dimensions (physical functioning, role limitations, social functioning, pain, mental health, and vitality), each of which have between four and six levels. In order to assign a utility score to each of the 18 000 potential SF-6D health states preferences were elicited from 611 UK residents using a modified version of the SG technique. Various regressions were developed in an attempt to assign a utility score to each of the SF-6D states (Brazier *et al*., 2002). We use the consistent (Badia *et al*., 1999) version of the SF-6D algorithm (Brazier *et al*., 2004) on which utility scores range between 0.296 for the pits (645655) and 1 for full health (111111).

Indicators of health status. The first indicator of health status we used was the second component of the EuroQol questionnaire (Brooks, 1996) – the VAS. In the VAS individuals are asked to indicate how good or bad their health state is (on the day they complete the questionnaire), on a scale where 0 corresponds to worst imaginable health state and 100 to best imaginable health state. In order to make comparisons between individuals of different health status, responses to the VAS were categorised into four levels, each of which contained an approximately equivalent number of respondents. The second indicator of health status was developed from a question concerning whether the individual currently took pain-killers (bought over the counter or prescribed by their general practitioner), where the response options were never, occasionally or regularly. The latter two responses were combined in order to give a binary variable denoting whether the individual currently took analgesics (either occasionally or regularly) or whether they did not. Finally, the third indicator was developed from the response (yes/no) to the question of whether the individual took any other prescribed medication.

Analysis

Descriptive statistics. Within the LIKP study researchers at the University did not have access to patient records at the general practice, as such information could only be gained on those individuals who returned the ascertainment questionnaire. Thus, it was not possible to assess whether responding individuals were representative of all individuals registered at the general practice. However, we do summarise the characteristics of respondents, and using data from the 2001 census (as described at: www.statistics.gov.uk/census2001), we make limited comparisons between the characteristics of respondents and those of people in England who are within the same age category.

Practicality. In line with Gerard *et al*. (2004) we assess practicality in terms of completion rates, and ultimately whether sufficient information was provided in order to calculate a utility score. Thus, for the EQ-5D, only if all five questions were completed was the EQ-5D deemed to have been completed, whereas, for the SF-6D, for some individuals, it was possible to calculate a utility score, even if some of the 11 questions were not answered (see Brazier *et al*., 2002; Gerard *et al*., 2004) for further information). For each of the EQ-5D and SF-6D, where it should be noted that the SF-36 appeared before the EQ-5D in the ascertainment questionnaire, we compared the characteristics of those who completed these measures to the characteristics of those who did not. We thereby assessed whether individuals with certain socio-demographic factors, certain clinical conditions or those with a certain level on the indicators of health status were more or less likely to complete the EQ-5D or SF-6D, differences were examined with the χ^2 test, and in these, and all subsequent, analyses a p-value <0.05 was deemed to be significant. Additionally, we also analysed the completion rates for the individual dimensions of the utility measures for the groups who were significantly less likely to complete either the EQ-5D or SF-6D.

DOI: 10.1002/hec

Convergent and construct validities. According to Streiner and Norman (2003), the level of convergent validity can be estimated by how closely a measure is related to other measures of the same construct. Similarly, construct validity can be defined in terms of whether a measure can discriminate between two groups, one which has a certain trait, and the other which does not (Streiner and Norman, 2003). In line with others (Macran *et al.*, 2007) we sought to assess both of these forms of validity with regard to indicators of health status (i.e. the EuroQol VAS and the use of medication). The VAS was used to assess convergent validity, where the rank correlation (Spearman's rho) between the VAS and each of the EQ-5D and SF-6D was estimated. Similarly, to assess construct validity, we assessed whether the EQ-5D and the SF-6D could discriminate between (i) individuals who reported that they currently took analgesics, and those who reported they did not and (ii) individuals who reported that they took other prescribed medication, and those who reported they did not. For both these comparisons we considered whether the difference between those who took analgesics/other medication and those who did not constituted a MID, the *t*-test was also conducted in order to assess whether these differences were significant.

Level of agreement. The level of agreement between the EQ-5D and SF-6D was estimated (for groups of individuals with different socio-demographic factors and different clinical conditions) by calculating both the Pearson's correlation coefficient and the ICC (Shrout and Fleiss, 1979). The ICC was based on a two-way random effects model, with absolute agreement, and we considered that a correlation ≥ 0.5 would denote that the two measures were strongly correlated (Guyatt *et al.*, 1987). Additionally, we also calculated the Bland–Altman plots (Bland and Altman, 1986) for the EQ-5D and SF-6D by plotting the average value of the EQ-5D and SF-6D scores (x-axis) against the difference between the EQ-5D and SF-6D score (y-axis), where as the difference was calculated by subtracting the SF-6D score from the EQ-5D, a score below (above) zero would denote that a particular individual had a utility score that was higher (lower) according to the SF-6D.

To identify when, and by what extent, EQ-5D and SF-6D scores are likely to differ we also estimated (for individuals who completed both measures) the mean difference between the EQ-5D and SF-6D scores (EQ-5D minus SF-6D) for individuals with different socio-demographic factors and different clinical conditions. For each of these groups we considered whether the mean difference between the two measures constituted a MID, the *t*-test was also conducted in order to identify significant differences.

In order to estimate whether less healthy individuals had significantly higher scores on the SF-6D and healthier individuals significantly higher scores on the EQ-5D, we compared the difference between the two measures for groups with different VAS scores. Additionally, in an attempt to assess where the cross-over between the EQ-5D and SF-6D scores occurred, we used the OLS regression technique to estimate the relationship between the SF-6D and the EQ-5D: $SF\text{-}6D = \alpha + \beta.EQ\text{-}5D$. The responses from all individuals were used in this regression, and we sought to identify the cross-over point, above which scores on the EQ-5D would be predicted to be higher than scores on the SF-6D, and below which scores on the SF-6D would be predicted to be higher than scores on the EQ-5D. In an attempt to corroborate the results of this regression we then compared the predicted results with other mean EQ-5D and mean SF-6D scores which had previously been reported in the literature. The studies for which results are summarised were located in a MEDLINE search (conducted October 1st 2006) using the key words EQ-5D and SF-6D, and were supplemented by other papers which the authors were aware of. Results are only presented for papers which had a sample size > 20 and reported both the mean EQ-5D and mean SF-6D scores (if results were presented at different time points (e.g. before and after intervention) only baseline scores are reported). If multiple papers were based on the same data set results are only reported for one of these papers.

DOI: 10.1002/hec

RESULTS

Descriptive statistics

There were 6765 individuals registered with the general practice on July 1st 2004, 3122 were aged ≥ 45 years, and 2770 were sent an ascertainment questionnaire (352 individuals were deemed to be unable to complete the ascertainment questionnaire). Ascertainment questionnaires were returned to the LIKP study team by 1865 individuals (67% of those who were sent one). It should, however, be noted that not all sections were completed by all individuals; thus, we note the number of individuals for whom data were available in each analysis. The characteristics of responding individuals, in terms of socio-demographic factors, clinical conditions, and indicators of health status, are summarised in Tables I, II and III, respectively. Of the respondents aged ≥ 45 years in the 2001 census, 18.9% were aged ≥ 75 years, 21.2% were aged ≥ 65 to <75, 26.6% were aged ≥ 55 to <65, and 33.3% were aged ≥ 45 to <55 years, and 53.6% were female. Thus, the responding individuals in this study were slightly older, and a greater proportion were female than in the general population.

Practicality

All five of the EQ-5D questions were completed by 1737 individuals (93.1%), but a utility score could only be calculated for 1612 individuals (86.4%) on the SF-6D ($p<0.001$). The completion rate for the EQ-5D did not vary according to socio-demographic factors (Table I) or the level of indicators of health status (Table III). However, the EQ-5D completion rate was found to be significantly higher for individuals who reported they had cancer, diabetes, heart disease, or osteoarthritis (Table II). Conversely, the SF-6D completion rate was found to be significantly lower for individuals who were older, female, had a lower occupational skill level, were from an area with higher deprivation (Table I), or who took prescribed medication (Table III), but significantly higher for individuals with asthma (Table II). Indeed, older individuals were found to be significantly less likely to complete five of the six dimensions of the SF-6D, females and those with a lower occupational skill level were significantly less

Table I. EQ-5D and SF-6D completion rates for individuals in different socio-demographic groups

		N	EQ-5D (%)	SF-6D (%)
Age (years)	> 75	392	95.4	82.4
	≥ 65 to <75	435	95.9	87.1
	≥ 55 to <65	511	97.8	93.0
	≥ 45 to <55	430	97.2	95.3‡
Gender	Female	985	96.5	87.3
	Male	800	96.8	92.6†
Ethnicity	White	1726	97.0	90.2
	Non-white	42	97.6	88.1
Smoked regularly	Yes	829	97.5	90.8
	No	954	96.8	89.5
BMI	Obese	277	96.4	89.5
	Overweight	625	97.0	90.2
	Underweight	24	91.7	87.5
	Normal	804	97.3	90.3
Occupational	Lowest (1)	131	96.2	83.2
Skill level	Skill 2	634	97.3	86.4
	Skill 3	410	96.8	92.7
	Highest (4)	567	98.2	95.2‡
IMD	High deprivation	678	97.2	90.3
	Low deprivation	581	97.2	94.0*

*$p<0.05$, †$p<0.01$, and ‡$p<0.001$ denote significant differences between groups.

DOI: 10.1002/hec

Table II. EQ-5D and SF-6D completion rates for individuals with and without different clinical conditions

		N	EQ-5D (%)	SF-6D (%)
Osteoarthritis	No	1645	92.5	86.0
	Yes	220	98.2†	89.5
Heart disease	No	1662	92.7	86.2
	Yes	203	97.0*	88.7
Diabetes	No	1754	92.8	86.2
	Yes	111	98.2*	90.1
Cancer	No	1765	92.9	86.5
	Yes	100	98.0*	86.0
Asthma	No	1740	92.9	85.9
	Yes	125	96.0	94.4†
Rheumatoid arthritis	No	1770	92.9	86.3
	Yes	95	97.9	88.4
Stroke	No	1803	93.1	86.4
	Yes	62	95.2	87.1
Knee pain	No	1351	97.3	90.5
	Yes	424	96.7	89.9
Hip pain	No	1562	97.4	90.8
	Yes	189	95.8	88.9
Back pain	No	1428	97.6	90.8
	Yes	325	96.6	89.5

*$p<0.05$, †$p<0.01$, and ‡$p<0.001$ denote significant differences.

likely to complete the physical functioning and role limitation dimensions, and those taking the prescribed medication were also significantly less likely to complete the role limitation dimension (Table IV).

Further examination of the level of missing data for individual questions of the SF-36 revealed that, of the three questions used to develop a score on the physical functioning dimension (questions 3a, 3b, and 3j), question 3a had a relatively high level of missing data (190 individuals failed to answer this question, compared with 125 and 114 for questions 3b and 3j). Similarly, the level of missing data was also relatively high for the two questions used to score the role limitation dimension (4c: 174 missing, 5b: 154), compared with other questions used to develop SF-6D scores (7: 95 missing, 8: 97, 9b: 115, 9e: 112, 9f: 114, and 10: 108), and also to questions within the EQ-5D (mobility: 96 missing, self-care: 95, usual activities: 97, pain: 100, and anxiety/depression: 99).

Convergent and construct validities

In terms of convergent validity, the rank correlation between the VAS and the EQ-5D was estimated to be 0.607 ($N = 1684, p<0.001$), compared with a correlation of 0.650 between the SF-6D and the VAS ($N = 1578, p<0.001$). With regard to construct validity, those individuals who reported they took analgesics had a lower mean utility score than those who did not, according to both the EQ-5D and SF-6D (Table III), where the difference in utility was both significant and considered to be a MID. The mean difference in utility between those who took other prescribed medication, and those who did not, was also considered to be a MID and significant (Table III).

Level of agreement

The EQ-5D and SF-6D scores, for the 1590 individuals who completed both these measures, were strongly correlated, with a Pearson's correlation coefficient of 0.717 ($p<0.001$) and an ICC of 0.644 ($p<0.001$). However, the level of correlation did vary according to different socio-demographic factors (Table V) and clinical conditions (Table VI). The Bland–Altman plot (Figure 1) also showed that there

DOI: 10.1002/hec

Table III. EQ-5D and SF-6D completion rates and mean scores for individuals at different levels on indicators of health status, mean differences (MD) are also calculated, both between groups (for the same measure) and within groups (EQ-5D score minus SF-6D score)

			Completion rate						
		N	EQ-5D (%)	SF-6D (%)	EQ-5D mean score	MD between groups (EQ-5D)	SF-6D mean score	MD between groups (SF-6D)	Within groups: MD (N)
VAS	0 to <65	378	97.4	89.7	0.537		0.610		−0.066‡ (334)
	≥ 65 to <80	410	98.3	92.7	0.768		0.742		0.024† (376)
	≥ 80 to <90	486	97.7	92.2	0.847		0.803		0.046‡ (440)
	≥ 90 to ≤ 100	443	98.9	92.8	0.920		0.879		0.042‡ (408)
Taking analgesics	Yes	1344	96.9	90.0	0.747	−0.134‡	0.750	−0.080‡	
	No	409	98.0	92.2	0.881		0.829		
Other prescribed medication	Yes	1090	96.9	88.7	0.730	−0.127‡	0.734	−0.081‡	
	No	679	97.5	92.9†	0.858		0.815		

*$p<0.05$, †$p<0.01$, and ‡$p<0.001$ denote significant differences.

DOI: 10.1002/hec

Table IV. Completion rates for the six dimensions of the SF-6D for groups of individuals which had significantly different SF-6D completion rates

		N	PF (%)	RL (%)	SF (%)	P(%)	MH (%)	V (%)
Age (years)	≥ 75	392	92.9	89.3	95.9	97.2	95.9	95.2
	≥ 65 to <75	435	93.3	94.7	97.9	98.4	97.9	97.2
	≥ 55 to <65	511	96.9	96.5	98.4	99.2	99.0	98.8
	≥ 45 to <55	430	97.7‡	97.4‡	98.8*	98.4	98.8†	98.8‡
Gender	Female	985	94.2	93.2	98.1	98.5	98.0	97.5
	Male	800	96.5*	96.5†	97.4	98.1	97.9	97.6
Occupational	Lowest (1)	131	94.7	90.8	96.9	97.7	97.7	96.9
Skill Level	Skill 2	634	93.1	94.0	97.9	98.7	98.3	97.8
	Skill 3	410	98.3	95.9	98.8	99.0	98.5	97.8
	Highest (4)	567	97.2‡	98.1‡	99.3	99.1	99.5	99.5
IMD	High deprivation	678	95.6	95.4	98.2	98.5	98.4	97.8
	Low deprivation	581	96.7	96.9	99.0	99.0	99.0	98.8
Other prescribed medication	Yes	1090	95.8	93.9	98.3	98.5	98.3	97.7
	No	679	95.7	97.2†	98.7	99.1	98.8	98.7

PF: physical functioning; RL: role limitations; SF: social functioning; P: pain; MH: mental health; V: vitality. *$p<0.05$, †$p<0.01$, and ‡$p<0.001$ denote significant differences.

Table V. Levels of correlation between the EQ-5D and SF-6D according to the intraclass correlation coefficient (ICC) and Pearson's correlation coefficient (Pearson) for individuals in different socio-demographic groups, mean VAS scores and mean differences (EQ-5D minus SF-6D) are also noted

		N	ICC	Pearson	VAS mean	Mean difference
Age (years)	≥ 75	318	0.643‡	0.740‡	69.97	−0.001
	≥ 65 to <75	371	0.637‡	0.704‡	75.72	0.001
	≥ 55 to <65	470	0.650‡	0.730‡	77.29	0.017*
	≥ 45 to <55	406	0.580‡	0.640‡	78.25	0.043‡
Gender	Female	846	0.637‡	0.717‡	76.35	0.006*
	Male	733	0.643‡	0.710‡	75.68	0.006
Ethnicity	Non-white	1535	0.642‡	0.716‡	76.26	0.016‡
	White	37	0.701‡	0.770‡	67.50	−0.007
Smoked regularly	Yes	745	0.649‡	0.724‡	75.15	0.007
	No	840	0.638‡	0.710‡	76.80	0.231‡
BMI	Obese	246	0.627‡	0.746‡	68.52	−0.013
	Overweight	554	0.649‡	0.720‡	75.48	0.014*
	Underweight	20	0.609‡	0.628‡	81.91	0.039
	Normal	720	0.627‡	0.689‡	78.94	0.027‡
Occupational	Lowest (1)	144	0.621‡	0.725‡	69.47	−0.003
Skill level	Skill 2	504	0.629‡	0.707‡	75.70	0.001
	Skill 3	375	0.671‡	0.741‡	74.30	0.004
	Highest (4)	534	0.639‡	0.701‡	78.96	0.043‡
IMD	High deprivation	678	0.649‡	0.716‡	75.54	0.013*
	Low deprivation	581	0.654‡	0.729‡	76.37	0.015*

*$p<0.05$, †$p<0.01$, and ‡$p<0.001$.

was a systematic variation in the EQ-5D and SF-6D scores, with less healthy individuals (mean score <0.5) tending to have higher scores on the SF-6D, and healthier individuals (mean score >0.8) tending to have higher scores on the EQ-5D. Figure 1 presents the Bland–Altman plot for the 1590 individuals who completed both measures, plots for individuals with different socio-demographic factors or clinical conditions also showed a similar trend (analyses not shown, but available from authors).

For the 1590 individuals who completed both the EQ-5D and SF-6D, the mean EQ-5D score (0.782, 95% confidence interval 0.771–0.794) was significantly ($p<0.001$) higher than the mean SF-6D score

DOI: 10.1002/hec

Table VI. Levels of correlation between the EQ-5D and SF-6D according to the intraclass correlation coefficient (ICC) and Pearson's correlation coefficient (Pearson) for individuals with different clinical conditions, mean VAS scores and mean differences (EQ-5D minus SF-6D) are also noted

	N	ICC	Pearson	VAS mean	Mean difference
Asthma	116	0.622‡	0.741‡	70.79	0.018
Cancer	85	0.640‡	0.734‡	70.33	0.013
Heart disease	178	0.641‡	0.777‡	65.51	−0.028
Diabetes	99	0.701‡	0.802‡	65.22	−0.045*
Knee pain	376	0.567‡	0.690‡	67.39	−0.054‡
Stroke	51	0.628‡	0.769‡	66.15	−0.059*
Osteoarthritis	194	0.570‡	0.694‡	65.92	−0.062‡
Rheumatoid arthritis	84	0.582‡	0.702‡	62.34	−0.062†
Back pain	285	0.546‡	0.698‡	64.91	−0.075‡
Hip pain	163	0.507‡	0.693‡	63.23	−0.101‡

*$p<0.05$, †$p<0.01$, and ‡$p<0.001$.

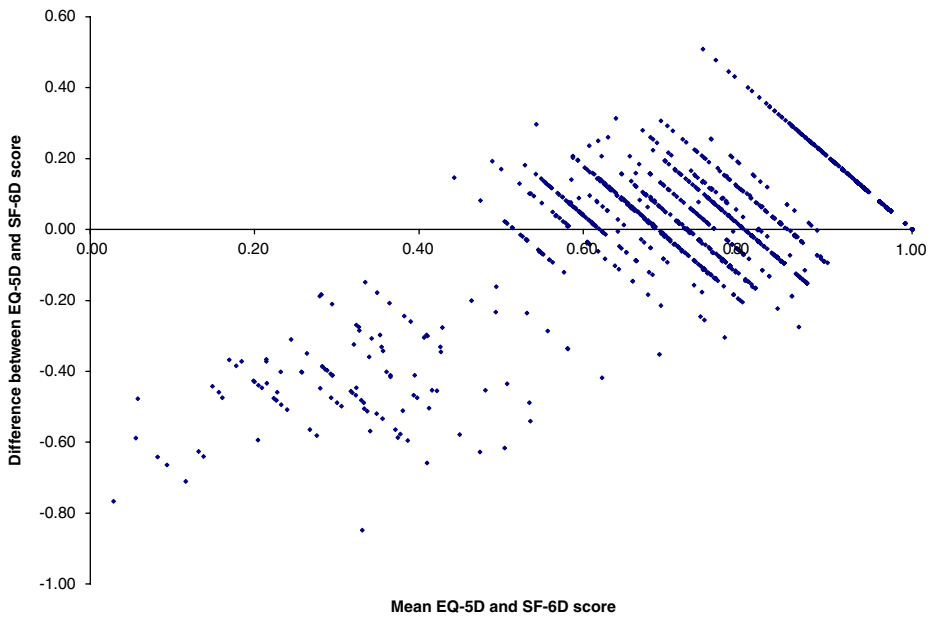

Figure 1. Bland–Altman plot for all 1590 individuals who completed both the EQ-5D and SF-6D

(0.767, 95% confidence interval 0.759–0.774) (mean difference = 0.016, 95% confidence interval 0.008–0.024). Indeed 34.4% of these individuals were estimated to be in full health according to the EQ-5D, compared with just 4.3% for the SF-6D. The mean difference between the mean EQ-5D and mean SF-6D scores (EQ-5D minus SF-6D) for individuals with different socio-demographic factors and different clinical conditions are estimated in Tables V and VI, respectively. With regard to socio-demographic factors, for individuals who were aged ≥ 45 to <55 years there was a MID between the mean EQ-5D score (0.832) and the mean SF-6D score (0.790), the same was also true for individuals in the highest occupational skill level (mean EQ-5D score = 0.839, mean SF-6D score = 0.796). In contrast to these two groups of individuals (where the mean EQ-5D score was higher than the mean SF-6D score), for individuals with clinical conditions a MID occurred only when the mean EQ-5D score was lower than

DOI: 10.1002/hec

the mean SF-6D score, this was the case for seven of the 10 clinical conditions – diabetes, knee pain, stroke, osteoarthritis, rheumatoid arthritis, back pain, and hip pain.

In order to assess whether less healthy individuals had significantly higher scores on the SF-6D and healthier individuals significantly higher scores on the EQ-5D, VAS scores were categorised into four levels: (i) 0 to <65, (ii) ≥ 65 to <80, (iii) ≥ 80 to <90, and (iv) ≥ 90 to 100. When the EQ-5D and SF-6D scores were compared within these groups it can be seen that, for less healthy individuals (VAS$<$ 65), the mean EQ-5D score was significantly lower than the mean SF-6D score (mean difference $= -0.066$), whereas healthier individuals had significantly higher mean scores according to the EQ-5D (Table III). There was a similar tendency across the socio-demographic factors and clinical conditions as many groups of individuals with low VAS scores (e.g. rheumatoid arthritis, back pain, or hip pain) had a significantly higher mean score on the SF-6D (Table VI), and many groups with high VAS scores (age ≥ 45 to <55 years, highest occupational skill level) had significantly higher mean scores on the EQ-5D (Table V). For each of these referred to groups of individuals the difference between the mean score on the EQ-5D and the mean score on the SF-6D could also be considered to constitute a MID.

In an attempt to identify where the cross-over between EQ-5D and SF-6D scores occurred we also used regression analysis, where it was estimated that $SF\text{-}6D = 0.414 + 0.451.EQ\text{-}5D$ (Adjusted $R^2 = 0.514$). This regression predicts that individuals with an EQ-5D score <0.754 would score higher on the SF-6D than on the EQ-5D. In Table VII the results are summarised for other studies which compared the EQ-5D and SF-6D and found the mean SF-6D score to be higher than the mean EQ-5D score, consistent with the predictions of our regression the mean EQ-5D score in each of these 17 patient groups was <0.754. In Table VIII results are summarised for papers which found the mean score to be higher on the EQ-5D. With three exceptions, the results of these papers were again in line with the predictions of our regression as they had a mean EQ-5D score > 0.754. Two of the exceptions had mean EQ-5D scores (0.729 and 0.72) which were close to the estimated cross-over value of 0.754, and actually fell within the lower confidence interval for this value (in the above regression the 95% confidence interval surrounding the intercept was 0.396–0.431, and that surrounding the estimate of 0.451 was 0.429–0.473). In the third exception there was no significant difference between the mean score on the EQ-5D and the mean score on the SF-6D (van Stel and Buskens, 2006).

Table VII. Summary results for papers which estimated the mean score to be higher on the SF-6D

	Mean score	
Population group (*N*)	EQ-5D	SF-6D
Irritable bowel syndrome (296) (Brazier *et al.*, 2004)	0.662	0.666
Rheumatoid arthritis (309) (van den Hout *et al.*, 2005)	0.648	0.676
Lower back pain (263) (Brazier *et al.*, 2004)	0.636	0.658
Stroke patients (98) (Pickard *et al.*, 2005)	0.62	0.68
Chronically ill haemodialysis patients (496) (Gerard *et al.*, 2004)	0.616	0.639
Healthy women > 75 years (291) (Brazier *et al.*, 2004)	0.614	0.662
Leg ulcer (430) (Brazier *et al.*, 2004)	0.552	0.647
Chronic obstructive pulmonary disease (230) (Brazier *et al.*, 2004)	0.542	0.572
Chronic low back pain (238) (Thomas *et al.*, 2005)	0.533	0.600
Asthma (poor control) (46) (Szende *et al.*, 2004)	0.52	0.63
Mental health patients (355) (Lamers *et al.*, 2006)	0.513	0.680
Liver transplant patients (183) (Longworth and Bryan, 2003)	0.517	0.606
Rheumatology patients (161) (Conner-Spady and Suarez-Almazor, 2003)	0.49	0.62
Very severe rheumatoid arthritis (self-report) (27) (Marra *et al.*, 2005)	0.47	0.50
Osteoporosis (404) (Brazier *et al.*, 2004)	0.442	0.521
Lumber spine disorders (3097) (McDonough *et al.*, 2005)	0.39	0.57
Very bad (self-reported) health (179) (Petrou and Hockley, 2005)	0.217	0.490

DOI: 10.1002/hec

Table VIII. Summary results for papers which estimated the mean score to be higher on the EQ-5D

Population group (*N*)	Mean score	
	EQ-5D	SF-6D
Very good (self-reported) health (5219) (Petrou and Hockley, 2005)	0.941	0.874
Asthma (good control) (36) (Szende *et al*., 2004)	0.93	0.80
General population (14736) (Petrou and Hockley, 2005)	0.845	0.799
Mild rheumatoid arthritis (self-report) (34) (Marra *et al*., 2005)	0.84	0.74
Hearing impaired adults (609) (Barton *et al*., 2004)	0.801	0.778
Patients with uncomplicated varicose veins (246) (Michaels *et al*., 2006)	0.77	0.74
HIV patients (59) (Stavem *et al*., 2005)	0.77	0.73
Menopausal women (278) (Brazier *et al*., 2004)	0.729	0.716
Age-related macular degeneration (209) (Espallargues *et al*., 2005)	0.72	0.66
Symptomatic coronary stenosis (561) (van Stel and Buskens, 2006)	0.64	0.63

DISCUSSION

We have shown that the results of the EQ-5D and SF-6D are comparable with regard to the performance of convergent and construct validities – the scores for both measures were strongly correlated with scores on the EuroQol VAS, and both measures were able to discriminate between individuals who took, and did not take, analgesics/other prescribed medication (Table III). However, with regard to practicality, an EQ-5D score could be calculated for 93.1% of respondents, compared with 86.4% for the SF-6D. Indeed, the SF-6D completion rate was found to be significantly lower for individuals who were older, female, had a lower occupational skill level, were from an area with higher deprivation (Table I), or who took prescribed medication (Table III). Against our expectations, EQ-5D completion rates were significantly higher for individuals with osteoarthritis, heart disease, diabetes, and cancer, on the SF-6D, the same was also true for individuals with cancer.

With regard to the issue of agreement between the two measures, despite the scores on the EQ-5D and SF-6D being highly correlated, the mean difference between the two measures often constituted a MID (Tables V and VI). Indeed less healthy individuals (e.g. VAS score < 65) had significantly higher mean scores on the SF-6D, and healthier individuals (e.g. VAS score ≥ 65) had significantly higher mean scores on the EQ-5D (see Tables III, V and VI). Moreover, individuals with an EQ-5D score < 0.75 were predicted to have a higher score on the SF-6D, and individuals with an EQ-5D score > 0.75 were predicted to have a lower score on the SF-6D. These predictions were generally corroborated by the mean EQ-5D and SF-6D scores reported in other papers (Tables VII and VIII).

Explanations

The two dimensions of the SF-6D which were most commonly found to have a significantly higher rate of missing data were physical functioning and role limitation (Table IV) – questions 3a (limitations of your health, with regard to vigorous activities), 4c (problems with your work or other regular daily activities as a result of your physical health: limited in the kind of work or other activities), and 5b (problems with your work or other regular daily activities as a result of emotional problems: accomplished less than you would like) were shown to have relatively high levels of missing data. One comment made a number of times next to question 3a, by respondents who did not complete it, was 'not applicable' with a few also stating 'I don't do vigorous activities'. Similarly, as questions 4c and 5b mention 'work', and the groups of individuals who had particularly high rates of missing data may not currently be in work (many of the older respondents were retired, many females reported that they were a housewife, and some classified in the low skilled occupation group were not working, but classified as such due to their previous job), it could be that these questions are also sometimes left unanswered

DOI: 10.1002/hec

because they are deemed not applicable to the individual, or they do not know whether they have limitations in these areas as they do not currently undertake such activities.

The EQ-5D was able to discriminate between individuals who did and did not use prescribed medication (on the basis of their mean utility scores), even though 25.8% of individuals who reported using medication had a utility score of 1.00 according to the EQ-5D. This implies that individuals within the same group often have quite differing levels of utility, and that the descriptive ability of the EQ-5D may be relatively inferior at the individual level (as has been argued previously, Brazier *et al.*, 2002). However, despite this the EQ-5D was still able to discriminate between groups on the basis of their mean scores and thereby performed well with regard to the criterion of construct validity. The fact that such a high proportion of individuals were classified in full health according to the EQ-5D may, however, explain the slightly lower level of rank correlation than the SF-6D, with the VAS.

The finding that healthier individuals had significantly higher mean scores on the EQ-5D and that less healthy individuals had significantly higher mean scores on the SF-6D may be partially explained by the different valuation systems used within each of the EQ-5D and SF-6D. In particular, TTO scores, which are used in the EQ-5D, have been shown to be higher for milder states, and SG scores (used in the SF-6D) have been shown to be higher for more severe states (Dolan *et al.*, 1996) (see Introduction for further discussion). A further explanation for the fact that healthier individuals tended to have higher scores on the EQ-5D is the argument that the SF-6D is most likely to be more sensitive (due to its larger descriptive system) in groups experiencing mild to moderate health problems (Brazier *et al.*, 2002).

Implications

The lower completion rates for the SF-6D would pose a particular problem if there were heterogeneity in the effect of treatment, and the treatment effect differed between those who did, and did not, complete the SF-6D. For example, if the benefits of a particular treatment were lower for older individuals, and the completion rate was also lower for older individuals, as has been shown previously (UK Cochlear Implant Study Group, 2004), then conducting a complete case analysis, where all individuals with missing data are excluded on the assumption that data were missing completely at random (Briggs *et al.*, 2003; Manca and Palmer, 2005), would result in an overestimation of the true average treatment effect and a more favourable estimate of the average level of cost-effectiveness. Thus, one implication of the finding that certain groups of individuals are significantly less likely to complete the SF-6D is that a complete case analysis should not be conducted, and imputation should be undertaken in order for a more appropriate estimate of true treatment effect, and the associated level of uncertainty, to be made.

Despite the EQ-5D having fewer dimensions, and fewer levels within those dimensions, it had comparable performance with the SF-6D with regard to convergent and construct validities and superior performance with regard to practicality. One interpretation of these results may be that it is unnecessary to expand the number of levels on the EQ-5D, and that this would only increase the rate of non-completion (with the associated problems discussed above). That said, we found no evidence that a greater number of question response categories were associated with a higher level of missing data, indeed question 7 of the SF-36 has the greatest number of response categories, but had the lowest level of missing data of all the 11 SF-6D questions.

One implication of the finding that less healthy individuals had significantly lower mean scores on the EQ-5D, and that healthier individuals had significantly higher mean scores on the EQ-5D is that the utility benefits of alleviating health conditions are generally likely to be estimated to be higher according to the EQ-5D than the SF-6D, as has been demonstrated previously (Conner-Spady and Suarez-Almazor, 2003; Longworth and Bryan, 2003; Pickard *et al.*, 2005; Thomas *et al.*, 2005; Lamers *et al.*, 2006; Michaels *et al.*, 2006; van Stel and Buskens, 2006). This, in turn, would mean that cost-

DOI: 10.1002/hec

effectiveness estimates are also likely to be more favourable according to the EQ-5D, compared with the SF-6D.

Comparisons with other studies

With regard to practicality, our results are in line with others who have shown that increasing age is associated with a decline in the completion rate of the SF-6D (Brazier *et al*., 2004; Gerard *et al*., 2004), and that missing data are higher for the physical functioning dimension of the SF-6D (Fisk *et al*., 2005). However, in contrast to our results, Gerard *et al*. (2004) also found that co-morbidity was associated with a reduced completion rate – we found that individuals with osteoarthritis, heart disease, diabetes, and cancer were more likely to complete the EQ-5D (than individuals without these conditions), and individuals with asthma were more likely to complete the SF-6D. We are unsure of why these results occurred in our study, but because Gerard *et al*. (2004) recruited haemodialysis patients and we recruited members of the general public, it may be that the individuals with clinical conditions in our study had less severe problems and saw such research to be of greater value (thereby completing more of the questionnaire) than those individuals without clinical conditions. It could also be that those individuals with clinical conditions were more keen to complete the EQ-5D and SF-6D in order to demonstrate how their condition affected them, possibly in the hope that they would be recruited in the next stage of the study.

We are not aware of any other studies which have assessed the construct and convergent validities of the EQ-5D and SF-6D with regard to indicators of health status. However, our results do concur with the study by Marra *et al*. (2005), which looked at the convergent and construct validities of the EQ-5D and SF-6D with regard to rheumatoid arthritis patients. However, in contrast, Espallargues *et al*. (2005) report that both the EQ-5D and SF-6D were unable to differentiate between patients with different severities of visual function and visual impairment.

Finally, with regard to the level of agreement, we found that stronger levels of correlation between the scores on the EQ-5D and SF-6D than others have observed (Brazier *et al*., 2004). Similarly, the agreement between these two measures was higher than that observed by Lubetkin and Gold (2003) when they assessed the level of correlation between the EQ-5D, SF-12 (Ware *et al*., 1996), and HUI3 (Feeny *et al*., 2002). That said, we still observe differences between the absolute scores on the EQ-5D and SF-6D, and our finding that less healthy individuals (i.e. with EQ-5D scores less than approximately 0.75) are predicted to have higher scores on the SF-6D generally concurs with the findings of others (see Table VII), as does the prediction that healthier individuals would have higher scores on the EQ-5D (see Table VIII). It also implies that results from different utility measures are not interchangeable, as has been argued elsewhere (Hatoum *et al*., 2004; Marra *et al*., 2004).

Study weaknesses

The data used in this paper were ascertained in a cross-sectional survey of individuals registered in one UK general practice. There are two main implications of this. The first is that we were only able to assess performance with regard to a limited criteria, and were not able to assess, for example, reliability or responsiveness (see Fitzpatrick *et al*., 1998; Streiner and Norman, 2003 for further discussion); thus, we are unable to conclude whether either of these measures is ultimately superior. The second implication is that the results of our study may not be generalisable to other groups of individuals. This arises because the data in this study were self-reported and not corroborated by the general practitioner, and because there would have been significant heterogeneity between individuals categorised as having the same condition (e.g. there are many forms and stages of cancer). Similarly, our results may not be generalisable because certain groups of individuals had a reduced completion rate for the SF-6D and we assessed convergent validity, construct validity and the level of agreement using a complete case analysis approach (Briggs *et al*., 2003; Manca and Palmer, 2005). The need for caution is indeed highlighted by

DOI: 10.1002/hec

the fact that we found that the presence of certain clinical conditions was associated with an increased completion rate, whereas Gerard *et al.* (2004) found the opposite effect. That said, the possibility of reporting spurious effects is limited by the large sample size of our study. Similarly, our findings were consistent across many different groups of individuals, e.g. completion rates were always lower for the SF-6D than the EQ-5D, and those who were less healthy (according to the VAS) consistently had lower scores on the EQ-5D.

A further limitation of this paper is that the created groups of individuals may not be as distinct as one would hope. For example, when assessing construct validity, we assessed whether the measures could discriminate between individuals who reported that they currently took analgesics/medication, and those who reported they did not. This was undertaken because of the *a priori* expectation that those who have a lower quality of life will tend to take more analgesics/medication. However, by taking analgesics/medication a person's quality of life may improve, and thus some people within this group may in fact have a similar quality of life to those who do not take analgesics/medication. The consequence of this would be to reduce the differences between the two groups, however, in spite of this the EQ-5D and SF-6D were still able to discriminate (on the basis of mean scores) between those who took analgesics/medication and those who did not. In a similar way, the appropriateness of categorising those who reported themselves to be a 'housewife/househusband' as a 'housekeeper' may be questioned, and this may reduce the differences between individuals who were classified in different groups of occupational skill level. In spite of this, we still found that the completion rate for the SF-6D was lower for those in lower occupational skill groups. Thus, the fact that differences were still detected in the presence of such potential mis-categorisations provides further support for both the construct validity of the two measures, and the argument that the completion rate of the SF-6D differs according to certain socio-demographic factors.

CONCLUSION

We have shown that the completion rates for the SF-6D are significantly lower than for the EQ-5D, and that individuals who are older, female, of lower occupational skill level, who live in a more deprived area, or use prescribed medication are less likely to complete the SF-6D. Particularly, for treatments were there is heterogeneity of effect, this may reduce the generalisability of results based on the SF-6D. Both the EQ-5D and SF-6D performed adequately in terms of convergent validity and construct validity, with regard to indicators of health status. We did, however, also find that the EQ-5D and SF-6D gave quite different utility estimates, particularly for groups of individuals at either end of the health spectrum – less healthy individuals had significantly higher mean scores on the SF-6D, and healthier individuals had significantly higher mean scores on the EQ-5D.

DECLARATION

We confirm that none of the authors have any conflicts of interest with regard to this article. Ethical approval for this study was granted by the Nottingham Research Ethics Committee.

ACKNOWLEDGEMENTS

We thank all patients who completed the Lifestyle Interventions for Knee Pain (LIKP) study questionnaire. The LIKP study was funded by the UK Arthritis Research Campaign (ARC) (grant number 13550). PhD funding for Garry Barton was provided by the UK Economic & Social Research Council (ESRC) (PTA-037-2004-00051). A previous version of this paper was presented at the 22nd EuroQol plenary meeting in Oslo, September 2005.

DOI: 10.1002/hec

REFERENCES

Badia X, Roset M, Herdman M. 1999. Inconsistent responses in three preference-elicitation methods for health states. *Social Science & Medicine* **49**: 943–950.

Badia X, Roset M, Herdman M, Kind P. 2001. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Medical Decision Making* **21**: 7–16.

Barton GR, Bankart J, Davis AC, Summerfield AQ. 2004. Comparing utility scores before and after hearing-aid provision: results according to the EQ-5D, HUI3 and SF-6D. *Applied Health Economics and Health Policy* **3**: 103–105.

Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measures. *Lancet* **8476**: 307–310.

Brazier JE, Deverill M. 1999. A checklist for judging the preference-based measures of health related quality of life: learning from psychometrics. *Health Economics* **8**: 41–51.

Brazier JE, Deverill M, Green C, Harper R, Booth A. 1999. A review of the use of health status measures in economic evaluation. *Health Technology Assessment* **3**: 1–164.

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

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

Briggs AH, Clark T, Wolstenholme J, Clarke P. 2003. Missing...presumed at random: cost-analysis of incomplete data. *Health Economics* **12**: 377–392.

Brooks R. 1996. EuroQol: the current state of play. *Health Policy* **37**: 53–72.

Bryan S, Longworth L. 2005. Measuring health-related utility: Why the disparity between EQ-5D and SF-6D? *European Journal of Health Economics* **6**: 253–260.

Conner-Spady B, Suarez-Almazor ME. 2003. Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Medical Care* **41**: 791–801.

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

Dolan P, Gudex C, Kind P, Williams A. 1996. Valuing health states: a comparison of methods. *Journal of Health Economics* **15**: 209–231.

Dolan P, Roberts J. 2002. Modelling valuations for EQ-5D health states: an alternative model using differences in valuations. *Medical Care* **40**: 442–446.

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. 2005. *Methods for the Economic Evaluation of Health Care Programmes* (3rd edn). Oxford University Press: New York.

Espallargues M, Czoski-Murray CJ, Bansback NJ, Carlton J, Lewis GM, Hughes LA, Brand CS, Brazier JE. 2005. The impact of age-related macular degeneration on health status utility values. *Investigative Ophthalmology & Visual Science* **46**: 4016–4023.

Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, Denton M, Boyle M. 2002. Multi-attribute and single attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care* **40**: 113–128.

Fisk JD, Brown MG, Sketris IS, Metz LM, Murray TJ, Stadnyk KJ. 2005. A comparison of health utility measures for the evaluation of multiple sclerosis treatments. *Journal of Neurology, Neurosurgery and Psychiatry* **76**: 58–63.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. 1998. Criteria for assessing patient based outcome measures for use in clinical trials. *Health Technology Assessment* **14**: 1–74.

Gerard K, Nicholson T, Mullee M, Mehta R, Roderick P. 2004. EQ-5D versus SF-6D in an older, chronically ill patient group. *Applied Health Economics and Health Policy* **3**: 91–102.

Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. 1987. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* **42**: 773–778.

Hatoum HT, Brazier JE, Akhras KS. 2004. Comparison of the HUI3 with the SF-36 preference based SF-6D in a clinical trial setting. *Value in Health* **7**: 602–609.

Jaeschke R, Singer J, Guyatt GH. 1989. Measurement of health status: ascertaining the minimal clinical important difference. *Controlled Clinical Trials* **10**: 407–415.

Lamers L. 2006. Adjustment of existing EQ-5D TTO values for use of an EQ-5D five level descriptive system. *The European Journal of Health Economics* **7**(S1): S57.

Lamers LM, Bouwmans CAM, van Straten A, Donker MCH, Hakkaart L. 2006. Comparison of EQ-5D and SF-6D utilities in mental health patients. *Health Economics* **15**: 1229–1236.

Longworth L, Bryan S. 2003. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. *Health Economics* **12**: 1061–1077.

Lubetkin EI, Gold MR. 2003. Areas of decrement in health-related quality of life (HRQL): Comparing the SF-12, EQ-5D, and HUI 3. *Quality of Life Research* **12**: 1059–1067.

DOI: 10.1002/hec

Macran S, Wileman S, Barton GR, Rusell IT. 2007. The development of a new measure of quality of life in the management of gastro-oesophageal reflux disease: the REFLUX questionnaire. *Quality of Life Research* **16**: 331–343.
Manca A, Palmer S. 2005. Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials. *Applied Health Economics and Health Policy* **4**: 65–75.
Marra CA, Esdaile JM, Guh D, Kopec JA, Brazier JE, Koehler BE, Chalmers A, Anis AH. 2004. A comparison of four indirect methods of assessing utility values in rheumatoid arthritis. *Medical Care* **42**: 1125–1131.
Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, Esdaile JM, Anis AH. 2005. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Social Science & Medicine* **60**: 1571–1582.
McDonough CM, Grove MR, Tosteson TD, Lurie JD, Hilibrand AS, Tosteson AN. 2005. Comparison of EQ-5D, HUI, and SF-36-derived societal health state values among spine patient outcomes research trial (SPORT) participants. *Quality of Life Research* **14**: 1321–1332.
Michaels JA, Brazier JE, Campbell WB, MacIntyre JB, Palfreyman SJ, Ratcliffe J. 2006. Randomized clinical trial comparing surgery with conservative treatment for uncomplicated varicose veins. *British Journal of Surgery* **93**: 175–181.
Office for National Statistics. 2000. *Standard Occupational Classification 2000* (*Volume 2*): The Coding Index
Office of the Deputy Prime Minister. 2004. *The English Indices of Deprivation 2004*. ODPM Publications: Wetherby, UK.
Petrou S, Hockley C. 2005. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. *Health Economics* **14**: 1169–1189.
Pickard AS, Johnson JA, Feeny DH. 2005. Responsiveness of generic health-related quality of life measures in stroke. *Quality of Life Research* **14**: 207–219.
Shaw JW, Johnson JA, Coons SJ. 2005. US valuation of the EQ-5D health states: development and testing of the D1 valuation model. *Medical Care* **43**: 203–220.
Shrout PE, Fleiss JL. 1979. Intraclass correlations: uses in assessing rater reliability. *Personality and Social Psychology Bulletin* **18**(2): 420–428.
Stavem K, Froland SS, Hellum KB. 2005. Comparison of preference-based utilities of the 15D, EQ-5D and SF-6D in patients with HIV/AIDS. *Quality of Life Research* **14**: 971–980.
Streiner DL, Norman GR. 2003. *Health Measurement Scales: A Practical Guide to Their Development and Use* (3rd edn). Oxford University Press: New York.
Sullivan PW, Lawence WF, Ghushchyan V. 2005. A national catalog of preference-based scores for chronic conditions in the United States. *Medical Care* **43**: 736–749.
Szende A, Svensson K, Stahl E, Meszaros A, Berta GY. 2004. Psychometric and utility-based measures of health status of asthmatic patients with different disease control level. *Pharmacoeconomics* **22**: 537–547.
Thomas KJ, MacPherson H, Ratcliffe J, Thorpe L, Brazier J, Campbell M, Fitter M, Roman M, Walters S, Nicholl JP. 2005. Longer term clinical and economic benefits of offering acupuncture care to patients with chronic low back pain. *Health Technology Assessment* **9**: 1–109.
Torrance GW. 1986. Measurement of health state utilities for economic appraisal. *Journal of Health Economics* **5**: 1–30.
Tsuchiya A, Brazier J, Roberts J. 2006. Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets. *Journal of Health Economics* **25**: 334–346.
Tsuchiya A, Ikedab S, Ikegamib N, Nishimurac S, Sakaid I, Fukudae T, Hamashimaf C, Hisashigeg A, Tamura M. 2002. Estimating an EQ-5D population value set: the case of Japan. *Health Economics* **11**: 341–353.
UK Cochlear Implant Study Group. 2004. Criteria of candidacy for unilateral cochlear implantation in postlingually deafened adults II: cost-effectiveness analysis. *Ear and Hearing* **25**: 336–360.
van den Hout WB, de Jong Z, Munneke M, Hazes JM, Breedveld FC, Vliet Vlieland TP. 2005. Cost-utility and cost-effectiveness analyses of a long-term, high-intensity exercise program compared with conventional physical therapy in patients with rheumatoid arthritis. *Arthritis and Rheumatism* **53**: 39–47.
van Stel HF, Buskens E. 2006. Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease. *Health and Quality of Life Outcomes* **4**: 20.
Walters SJ, Brazier JE. 2003. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes* **1**: 4.
Walters SJ, Brazier JE. 2005. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research* **14**: 1523–1532.
Ware JE, Kosinski M, Keller SD. 1996. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical Care* **34**: 220–233.
Ware JE, Sherbourne C. 1992. The MOS 36 item short-form health survey: conceptual framework and item selection. *Medical Care* **30**: 473–483.
WHO. 2001. Obesity: preventing and managing the global epidemic. *Report of a WHO Consultation on Obesity*. WHO: Geneva.

DOI: 10.1002/hec