A panel data comparison of two commonly-used health-related quality of life instruments

Zachary Gerring*[1]

and

Jaikishan Desai[1]

Abstract

Economic evaluations of health interventions require measuring their impact and this is usually undertaken with generic health-related quality of life measurement instruments like the EQ-5D and SF-6D. In this paper we examine the agreement between EQ-5D and the SF-6D using panel data on a set of 1176 New Zealand patients referred for elective surgery between 2003 and 2006. We analyse the relationship between the two instruments amongst all patients who completed at least two interviews over the 18-month study period, and separately analyse the sensitivity of the two instruments in measuring the impact of elective surgery (amongst those who had surgery). Preliminary results suggest there is poor agreement between the EQ-5D and SF-6D indices longitudinally. Furthermore, the two instruments are different in capturing the immediate change in health-related quality of life after surgery.

[1] Health Services Research Centre, Victoria University of Wellington, New Zealand

*Corresponding author. Contact details: Health Services Research Centre, Old Government Buildings, 15 Lambton Quay, Wellington, New Zealand. Ph. 04 463 6570

This paper presents preliminary findings of an analysis of two commonly-used quality of life instruments. For a revised copy, please contact corresponding author.

I. Introduction

In New Zealand and overseas, economics-based evaluations of health care programmes are increasingly being used in health care resource allocation and policy-making. For example, the Pharmaceutical Management Agency (PHARMAC) performs economic evaluations to quantify the added 'benefit' of funding new medicines (Pharmaceutical Purchasing Agency 2009). Such evaluations depend on preference-based health-related quality of life (HRQoL) instruments to estimate health state utilities that can be used to generate quality adjusted life years (QALYs) in cost-utility analysis (CUA). There are currently a plethora of preference-based instruments available to generate health utilities, two of which include the EuroQoL EQ-5D and the recently developed SF-6D (Brazier, Roberts et al. 2002). These instruments differ with respect to their descriptive systems, preference valuation method, and source of preferences (Brazier, Roberts et al. 2004). As a result, the instruments can be expected produce different health utilities, and therefore offer competing estimates of HRQoL. To determine the strength of these instruments, and maximise measurement performance for a given population under study, head-to-head comparisons can be made against scientific review criteria (Brazier and Deverill 1999). In doing so, the appropriateness of each instrument for use in economic evaluations can be established.

Reliability and responsiveness represent two important criteria for assessing the longitudinal performance HRQoL instruments. Reliability refers to the degree in which a health measure produces the same results in a stable population, and can be assessed by measuring the consistency of the instrument over time (test-retest) or the concordance between two instruments (Brazier and Deverill 1999). By contrast, responsiveness refers to the ability of the instrument to detect clinical change in health, and is commonly assessed in longitudinal studies where change is expected to occur (Fayers and Machin 2007).

Research comparing the EQ-5D and SF-6D for the same set of patients is limited (Brazier, et al. 2004), and to our knowledge there are no studies that have compared the instruments with panel data in patients referred for elective surgery. We aim to examine the concordance between the EQ-5D and SF-6D using panel data on a set of New Zealand patients referred for elective surgery between 2004 and 2006. We begin by presenting a brief description of the two instruments, before conducting a comparative analysis.

II. Review of the literature

EQ-5D and the SF-36

The EQ-5D, developed by the 'EuroQol group', is a self assessed health status classification instrument designed to measure health-related quality of life (Brooks and De Charro 1996). The instrument has five domains: mobility, self-care; usual activities; pain/discomfort; and anxiety/depression, each of which has a three-point response scale (1 = no problems, 2 = some problems, 3 = major problems). A five-digit sequence of numbers relating to the responses within each domain is produced, ranging from 11111 (no problems across all domains) to 33333 (major problems across all domains), which is used to define the health state (Devlin, Hansen et al. 2003). With the addition of 'unconscious' and 'dead' states, a total of 245 unique health states can be defined (Drummond, Sculpher et al. 2005). EQ-5D value weights for the New Zealand general population can be obtained using Devlin's sub-sample tariff (Devlin, Hansen et al. 2003). Devlin and colleagues used the visual analogue scale[1] (VAS) technique to derive index weights for a set of 13 EQ-5D health states from a random sample of New Zealand adults.

The Short-Form – 36 (SF-36) health instrument is a commonly used generic measure of health, consisting of 36 items grouped into 8 domains: physical functioning (10 items), physical health and daily roles (4 items); bodily pain (2 items); general health (5 items); vitality (4 items); social functioning (2 items); emotional health and daily roles (4 items); and mental health (5 items; Brazier, Jones et al. 1993). Within each item, the patient can make one of a set amount of responses. For example, in the emotional health and daily roles dimension, the patient can make one of five responses: 'all of the time', 'most of the time', 'some of the time', 'a little of the time' or 'none of the time'. An additional item, the self-reported health transition, measures perceived change in health, but is not scored (Brazier, Harper et al. 1992). The SF-36, however, does not reflect population preferences, and therefore cannot be used to derive utility scores. In light of this, Brazier and colleagues (2002) developed an algorithm that generates a compressed version of the SF-36 – the SF-6D – with preference weights that can be used to generate utilities. For the purposes of the present study, we applied Brazier's SF-6D utility algorithm (Brazier, Roberts et al. 2002) to all patient responses. Brazier's algorithm reduces the number of domains in the SF-36 from eight to six by dropping the general health domain and amalgamating 'role limitation due to physical problems' and 'role limitation due to emotional problems.' The SF-6D has six dimensions (physical functioning,

[1] VAS values are not 'choice-based', and the extent to which VAS data represent health state preferences and utilities is controversial (see, for example Parkin and Devlin 2006). However, the VAS remains the simplest and most popular method to measuring preference for health states.

role limitations, social functioning, pain, mental health and vitality), each with four to six levels, giving the instrument the capacity to describe 18,000 unique health states. The scoring model for the SF-6D was derived from a random sample of 611 members of the UK national population using the standard gamble (SG) technique (Brazier and Roberts 2004).

Comparison of quality of life measures

Independently, both the EQ-5D and SF-6D have shown good test-retest reliability and responsiveness to changes in quality of life across a wide range of populations and clinical conditions (see, for example: Brazier, Harper et al. 1992; Brazier, Walters et al. 1996; Scott, Tobias et al. 1999; Bosch and Hunin 2000; Johnson and Pickard 2000; Marra, Woolcott et al. 2005; Horowitz, Abadi-Korek et al. 2010; Kontodimopoulos, Pappa et al. 2010). However, comparative analyses have revealed that the instruments produce different index scores for a given population or disease group, and agreement between instruments is generally found to be poor. Differences between EQ-5D and the SF-6D are not easily defined, and are often concealed by small mean differences between the two measures. Thus, comparative analyses have placed particular emphasis on examining the distribution of index scores.

Petrou *et al.* (2005) examined the relationship between EQ-5D and SF-6D utility scores using a general population sample derived from the 1996 Health Survey for England. The mean utility score derived from the EQ-5D (0.845) was above that of the SF-6D (0.799), yielding a mean difference of 0.046. However, the distribution of responses varied considerably, and the EQ-5D scores ranged from -0.308 to 1 while the SF-6D scores ranged from 0.296 to 1. Further, there was a wide range in SF-6D values for EQ-5D responses defined as 1.0 (perfect health), with considerable levels of impairment in physical functioning, pain, mental health, and vitality detected by the SF-6D but not the EQ-5D. Similar findings have been reported across a variety of patient groups (Brazier, Roberts et al. 2004; Van Stel and Buskens 2006; Zhao, Yue et al. 2010).

Brazier and colleagues (2004) measured the convergent validity[2] between similar dimensions of the EQ-5D and SF-6D across seven patient groups (lower back pain, chronic obstructive pulmonary disease, irritable bowel syndrome, leg ulcer, menopausal woman, osteoporosis, and healthy older women). Evidence of convergent validity was found between corresponding dimensions of the EQ-5D and SF-6D, including: physical functioning (EQ-5D) and mobility (SF-6D); usual activities role

[2] Convergent validity is an estimation of agreement between instruments measuring the same concept (Brazier and Deverill 1999)

limitation (EQ-5D) and social functioning (SF-6D); Pain/discomfort (EQ-5D) and pain (SF-6D); and anxiety/depression (EQ-5D) and mental health (SF-6D). The range of responses for EQ-5D and SF-6D were -0.4 to 1.0 and 0.3 to 1.0, respectively. The distribution of responses across all dimensions of each instrument ceiling effects in the EQ-5D and floor[3] effects in the SF-6D. Importantly, relationship between EQ-5D and SF-6D varied according to clinical condition. Overall, the SF-6D was observed to generate larger index values than EQ-5D, although the mean difference was small.

Longworth and Bryan (2003) compared the EQ-5D and SF-6D in patients eligible for liver transplant over a 12 month period in England and Wales. The EQ-5D was found to be responsive to transplant, but not the SF-6D. There was evidence of a substantial floor effect in the SF-6D, however the instrument was also more responsive to changes at the higher end of the scale. The authors concluded that high variation exists between the two measures, which they attributed to the narrow scoring range of the SF-6D.

The aetiology of disparities between the EQ-5D and SF-6D index values remains controversial. The descriptive systems of the instruments (Bryan and Longworth 2005; Grieve, Grishchenko et al. 2009), the scoring algorithms (Søgaard, Christensen et al. 2009), and the method of health state valuation used to derive utility scores (Bryan and Longworth 2005) are commonly cited reason for disagreement between the instruments. For example, compared to the SF-6D, the EQ-5D has fewer domains (five to six for the SF-6D), each of which has fewer levels (three versus four to six). Further, as Gieve *et al*.(2009) identified, the EQ-5D does not have an equivalent domain for 'vitality', and that the 'usual activities' domain in the EQ-5D does not fully represent 'social functioning' in the SF-6D. As a result, it is likely that the EQ-5D does not capture the impact of a disease or disability on vitality and social functioning to the same extent as the SF-6D, which partially explains why EQ-5D utility values are consistently found to be lower than their SF-6D counterparts.

The method of health state valuation used to derive utility scores has been found to produce different utilities for the same health states (Gudex, Dolan et al. 1996; Bryan and Longworth 2005; Tsuchiya, Brazier et al. 2006). The SG technique, used to derive the SF-6D weights used in the present paper, generates weights that generally exceed those derived using the VAS technique

[3] A ceiling effect occurs when a patient records the highest or near highest score, where as a floor effects occurs when a patient presents a baseline score at or near the lowest defined score (Drummond et al., 2005). As such, the patient is limited in their ability to demonstrate future change in health status, even if a change is clinically evident. Evidence suggests that the SF-6D suffers from the floor effect, whereas, at the other end of the scale, the EQ-5D is susceptible to the ceiling effect. The existence of a floor or ceiling effect has been identified as a potential cause of unresponsiveness in quality of life instruments, and likely contributes to differences in the sensitivity between the EQ-5D and SF-6D.

(Drummond, Sculpher et al. 2005). Further, Gudex et al. (Gudex, Dolan et al. 1996) reported that social class, education level, home ownership, and experience of illness had an impact on the health state valuations elicited using the VAS technique.

III. Data

The Pathways to Care and Outcomes for Elective Surgery (Pathways) study was a prospective cohort study of 1603 patients who were tracked from initial referral for elective surgery for a period of 18 months with regular data collection every 3 months. Funded by the funded by the Health Research Council, the study sought to: a) to identify and describe a cohort of patients considered for referral to elective surgery from primary care across six DHB localities, b) identify patients who satisfy the criteria for publicly-funded elective surgery, but do not receive surgery, c) to identify inequities in access to surgery that are modifiable by policy intervention, and d) to develop recommendations to maximise the efficiency and equity of the referral process (Dowell, Morgan et al. 2007).

The Pathways study was set in six district health board (DHB) localities, selected for their relative volumes of elective surgery discharges and demographic compositions. All general practitioners working within each locality were invited to participate in the study, representing a total of 333 general practices, and 828 General Practitioners (GPs). A total of 175 general practices participated in the study, representing 345 GPs (a 42% participation rate). Over a period of 3 or 5 weeks, depending on the response rate, GPs were asked to recruit all patients they considered eligible for referral for elective surgery. Patients were excluded if they were younger than 18 years old; if they were deemed inappropriate for inclusion by their GP for clinical reasons; if they were unable to consent to participation; if the referral was for screening for a disease; or if the referral was for emergency reasons. A recruitment questionnaire was completed by the GP for each patient, which detailed provisional diagnosis, the specialty to which the patient was referred, reasons for referral, need for surgery, urgency of specialist consultation, and whether the referral was made to the public or private sector. GPs were also asked to provide patients with study information, and seek verbal permission to send the study team the completed referral questionnaire and patient contact details to the study team.

Patients consenting to their contact details being sent to the study team were contacted by telephone, informed of the study design and objectives, and asked to participate in the longitudinal phase of the study. Participants were asked to complete a baseline interview by telephone, which included questions about the participants' medical history, referral details (health insurance status, reasons for referral, referral destination), socio-demographic characteristics, and health-related quality of life (using the EQ-5D and SF-36 questionnaires). A total of 1176 patients completed the baseline interview. Participants were then asked to complete a series of identical follow-up interviews by telephone every 3 months for a maximum of 18 months, or until intervention was received and 6 months post-intervention. The follow-up interviews collected information on accident and emergency department admissions, all hospital admissions (public/private, inpatient and outpatient), prescription medication, tests and investigations, care for dependants and volunteer work, paid work, leisure activities, and relevant direct and indirect costs for the previous three months. Further, the EQ-5D and SF-36 questionnaires were completed at each follow-up interview[4]. As a result, we have data on patients' quality of life using both measures from multiple interviews, pre- and post-intervention.

IV. Methods

The data-set contains multiple measures of quality of life, as measured by the EuroQoL EQ-5D and SF-36, pre- and post surgery. To estimate EQ-5D index scores for each patient, we applied Devlin's sub-sample algorithm (Devlin, Hansen et al. 2003), which obtained domain weights using methods described above. We obtained the SF-6D responses and indices from the SF-36 using Brazier's algorithm (Brazier, Roberts et al. 2002). To calculate the limits of agreement (Bland and Altman 1986) , both sets of index scores were transformed onto a 0-1 scale.

The primary aim of this paper is to assess the consistency of the EQ-5D and SF-6D longitudinally, and examine their sensitivity in detecting changes in health-related quality of life resulting from a clinically important event. We do this first by examining the consistency of each index in repeated interviews 3 months apart. For this, we selected a sub-sample of patients with at least two measurements for each index pre- and post surgery, excluding measurements splintered by a treatment event; this was done to ensure that a major treatment event (for example, elective surgery) did not influence results.

Analysis of the consistency sub-group was performed in two stages. Stage one involved an assessment data collected at each measurement period. Descriptive statistics (mean standard deviation [SD], 25th percentile, median, 75th percentile, inter-quartile range [iqr], and minimum and maximum values) were computed for each interview interval, and compared across specialty referrals. The paired t-test was used was used to measure within-subject difference between the

[4] Addition details about the Pathways to care and elective surgery study, its methodology and response rates, are reported elsewhere: Dowell et al. (2007).

two sets of index score. The strength of association between the EQ-5D and SF-6D was assessed using Pearson's product moment correlation. To measure agreement between the EQ-5D and SF-6D index values, we first computed the concordance correlation coefficient (Lin et al., 1989) for each measurement period and constructed Bland-Altman plots for the two groups of index scores (Bland and Altman 1986). For comparative purposes, we then computed the intra-class correlation coefficient (ICC) using linear mixed effects regression with factors of subject and HRQoL instrument. To assess the ceiling effect of each domain of the EQ-5D, the distribution of responses within each SF-6D domain was computed where the EQ-5D index score equalled 1.0 (perfect health) and the SF-6D index score was less than 1.0. Stage two of the analysis introduced the longitudinal element of responses. For the consistency sub-group, we estimated a patient fixed effects model and a three-level variance components model.

In order to estimate the sensitivity of each index in measuring the change in health-related quality of life resulting from elective surgery we selected a sub-sample of patients with at least one pre-operative interview and one post-operative interview. A patient fixed effects model was used to assess sensitivity by each index separately, and a variance components model is used to compare the two indices.

V. Results

Of the total 1603 patients recruited for the study 1176 patients completed the baseline interview and 1119 had complete data on all analysis variables.[5] A total of 668 patients qualified for the consistency sub-group, and 438 patients qualified for the sensitivity sub-group. The consistency and sensitivity analyses were based only on patients who completed both the questionnaires.

The demographic characteristics of those who completed the baseline interview, and the consistency and sensitivity sub-samples are reported in table 1. In the baseline (complete) sample, 57 percent of patients were female and 73 percent described themselves as New Zealand European, followed by Maori (8 percent), Pacific (2 percent), and Asian (2 percent). The mean age of patients was 55 years; 31 percent were aged between 18 and 44 years, 35 percent between 54 and 64 years, and 34 percent aged 65 years or older. Sixteen percent had a university degree or qualification and

[5] Together with the follow-up surveys 4,976 completed questionnaires were available for analysis, and of these 170 (3.4%) had missing or incomplete EQ-5D data, and 842 (16.9%) had missing or incomplete SF-6D data.

60 percent of patient referrals were made to the public sector. The demographic characteristics of the consistency and sensitivity sub-samples are largely similar to the full sample.

Table 1: Demographic and background characteristics of patients referred for elective surgery

Variable	**All (n=1119)**	**Consistency (n=668)**	**Sensitivity (n=438)**
Gender (female %)	57.0	58.0	58.6
Age: Mean age in years (SD)	55 (17.0)	56 (16.8)	55 (16.3)
Age group (%)			
18-44	31.0	27.9	31.2
45-64	35.0	36.5	34.4
65+	34.0	35.6	34.4
Ethnicity			
NZ European	73.0	72.1	73.5
Maori	8.1	6.4	7.5
Pacific*	2.1	2.5	1.5
Asian	2.2	2.2	1.1
Other	14.9	13.6	13.3
Education			
University degree or qualification	15.6	15.7	13.7
Referral			
Public	59.72	57.69	58.9
Private	39.66	41.24	40.1

Notes: *Samoan, Tongan, Niuean, or Cook Islands origin.
Not all percentages add to 100 as respondents for whom a response was not recorded are not displayed.

Patient referral by medical specialty is displayed in Table 2. The largest group of respondents were referred for a general surgery/vascular specialist assessment (34 percent), followed by orthopaedics (27 percent), gynaecology (11 percent), and plastics (9 percent). Within patients who received treatment, 39 percent were referred to general/vascular surgery, 22 percent to orthopaedics, and 13 percent to gynaecology. Compared to national referral information, general surgery and orthopaedics specialties were overrepresented, where as Ear, Nose and Throat (ENT), and ophthalmology were underrepresented (Raymont, Morgan et al. 2002). A more complete breakdown of the demographic and clinical characteristics of the study population is reported elsewhere (Raymont, Morgan et al. 2002).

Table 2: Patient referrals by speciality, compared with national statistics and patients who received treatment.

Specialty	**Completed M0* (%)**	**National (%)****	**Treatment received**
General surgery/vascular	34.2	22.0	39.4
Orthopaedics	27.4	19.0	22.3

Gynaecology	10.6	14.0	13.2
Plastics	9.2	5.0	10.2
Urology	5.5	5.0	5.2
ENT	6.5	15.0	4.5
Ophthalmology	3.9	18.0	2.9
Cardiothoracic	1.3	0.0	1.3
Neurosurgery	1.3	1.0	1.1

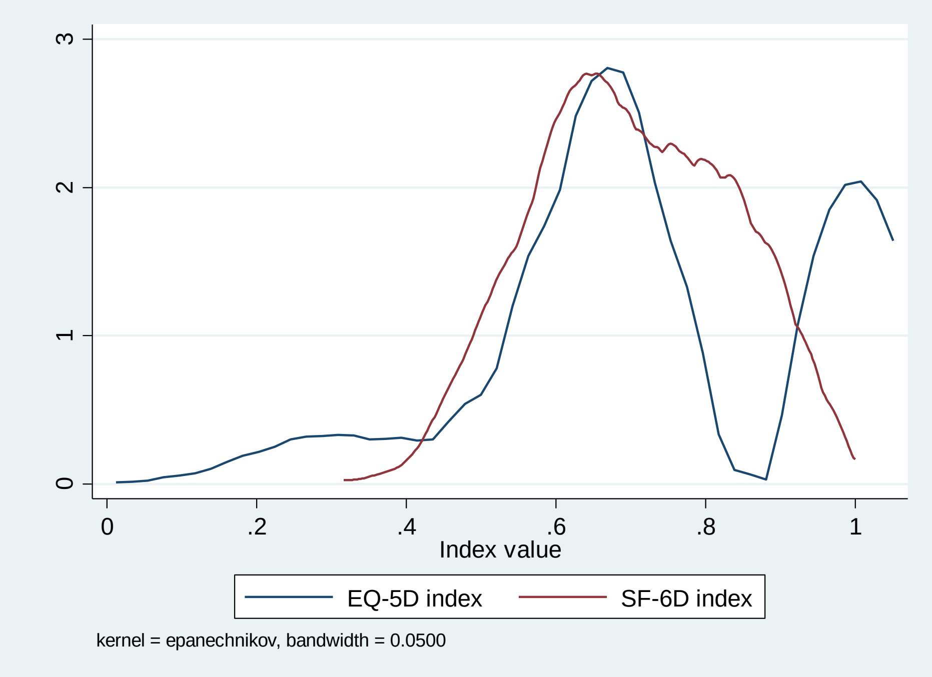
N=1119
*M0 = baseline interview
**National figures provided by New Zealand Health Information Service (NZHIS)

Descriptive statistics for the two health-related quality of life indices for all patients and the consistency and sensitivity sub-groups are presented in Table 3. The mean index values for the two sub-groups were similar to those of the full cohort. A small but significant difference was observed between the EQ-5D indices for the full sample and the sensitivity sample. The mean difference between index scores for all patients was 0.033, and exceeded the accepted range for evaluative purposes, frequently cited as 0.03 (Drummond 2001). While the mean index values for the EQ-5D and SF-6D were similar, there was substantial disagreement in the distribution of responses. Figure 1 shows that the while the two indices are quite consistent in the middle part of the 0-1 scale, there is considerable disagreement in the two tails of the distribution. The index values for the EQ-5D ranged from 0.007 to 1, with 37 percent of patients in health states defined as perfect. In contrast, the SF-6D index values ranged from 0.301 to 1, with only 1.1 percent of patients in health states equal to perfect health. The EQ-5D responses show a bimodal distribution, due to a high density of responses at 1.0, with a negative skew. The SF-6D plot has a narrower scoring range than the EQ-5D and a positive skew.

Table 3: Descriptive statistics for the EQ-5D and SF-6D index scores, baseline to 18 months (all patients, consistency sample, and sensitivity sample).

	All patients		**Consistency**		**Sensitivity**	
	EQ-5D	**SF-6D**	**EQ-5D**	**SF-6D**	**EQ-5D**	**SF-6D**
Number	4806	4134	3028	2632	2178	1974
Mean	0.763	0.730	0.755	0.728	0.753	0.731
Standard deviation	0.216	0.135	0.214	0.134	0.218	0.136
25th percentile	0.627	0.627	0.627	0.624	0.627	0.624
Median	0.716	0.733	0.716	0.729	0.716	0.734
75th percentile	1.000	0.852	1.000	0.845	1.000	0.852
Inter-quartile range	0.373	0.225	0.373	0.221	0.373	0.228
Minimum	0.007	0.301	0.007	0.322	0.007	0.301
Maximum	1.000	1.000	1.000	1.000	1.000	1.000

Figure 1: Kernel density plot of EQ-5D and SF-6D index scores at baseline interview



Compared to the SF-6D, the EQ-5D has a much wider dispersion of responses, both pre- and post-surgery. Further, there are a substantial number of responses for the EQ-5D that extend below the lowest SF-6D index value, indicating a floor effect in the SF-6D. By contrast, the EQ-5D shows a high density of responses at 1 which is not mirrored by the SF-6D distribution. This suggests that the EQ-5D is less capable of detecting health states near perfect health. To investigate this further, we examined the distribution of SF-6D responses where the EQ-5D index value equals 1, and the SF-6D is a score of less than 1. The SF-6D identified high levels of impairment in physical functioning, pain, mental health and vitality that was not identified by the EQ-5D. For example, amongst patients who did not report any problems with mobility defined by the EQ-5D, 40 percent of patients reported some limitation in the physical functioning domain of the SF-6D. Similarly, for patients who reported no problems with pain using the EQ-5D, 28 percent identified some issues with pain in daily life defined by the SF-6D. Hence, a large proportion of patients who report full health in EQ-5D report some problems in SF-6D.

Table 6: Distribution of data points (%) where the EQ-5D index score equalled 1 and the SF-6D was less than 1 for each SF-6D domain, consistency sample.

Level	Physical functioning	Role limitation	Social functioning	Pain	Mental health	Vitality
1	376 (43)	604 (72)	697 (79)	397 (45)	540 (61)	30 (7)
2	354 (40)	182 (18)	93 (11)	249 (28)	224 (25)	475 (52)
3	107 (12)	31 (4)	68 (8)	165 (19)	98 (11)	240 (27)
4	29 (3)	62 (6)	19 (2)	43 (5)	16 (2)	101 (10)
5	13 (1)	-	2 (>1)	21 (2)	1 (>1)	33 (3)
6	0	-	-	4 (>1)	-	-

Descriptive statistics for EQ-5D and SF-6D indices by surgical speciality are presented in table 4. The mean EQ-5D index values exceeded the mean SF-6D value across all patient groups with the exception of the orthopaedic sub-group. Patients referred for an orthopaedic specialist assessment had the highest level of impairment, with mean index scores of 0.624 and 0.662 for the EQ-5D and SF-6D, respectively. Patients referred for a plastic and reconstructive surgical specialist assessment reported the least impairment, with mean scores of 0.814 and 0.769 for the EQ-5D and SF-6D, respectively. The largest mean difference between index values was observed in patients referred for an ophthalmological consultation (0.09), and the smallest mean difference is observed in patients referred for an orthopaedic consultation (0.038).

Table 4: Descriptive statistics for the EQ-5D and SF-6D index scores, baseline interview, by referral specialty.

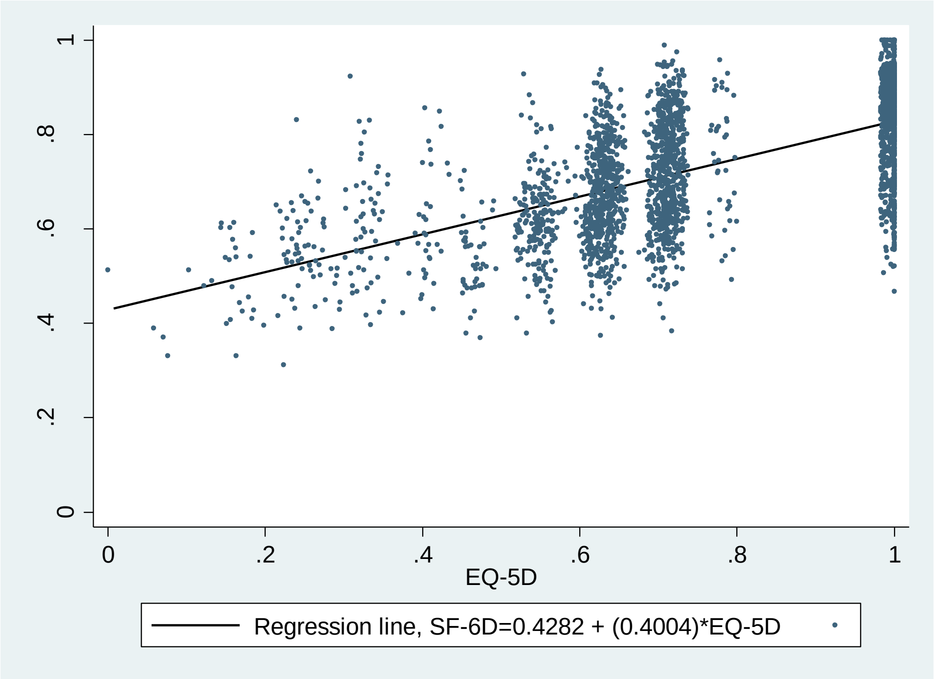
Specialty	Index	N	mean	sd	p50	min	max	ICC
General/VS	EQ-5D	366	0.749	0.2	0.716	0.133	1	0.394
	SF-6D	314	0.716	0.131	0.725	0.401	1	
Orthopaedics	EQ-5D	293	0.624	0.203	0.627	0.062	1	0.443
	SF-6D	262	0.662	0.13	0.652	0.316	1	
Gynaecology	EQ-5D	111	0.773	0.211	0.716	0.168	1	0.407
	SF-6D	102	0.708	0.124	0.696	0.506	1	
Plastics	EQ-5D	94	0.814	0.215	1.000	0.239	1	0.687
	SF-6D	70	0.769	0.136	0.799	0.476	1	
ENT	EQ-5D	71	0.786	0.182	0.716	0.39	1	0.404
	SF-6D	62	0.725	0.112	0.707	0.489	0.958	
Urology	EQ-5D	57	0.777	0.215	0.716	0.23	1	0.476
	SF-6D	49	0.719	0.138	0.7	0.452	0.929	
Ophthalmology	EQ-5D	43	0.805	0.204	0.721	0.253	1	0.565
	SF-6D	34	0.715	0.137	0.707	0.398	1	
Other*	EQ-5D	28	0.707	0.258	0.716	0.168	1	0.623
	SF-6D	15	0.671	0.11	0.639	0.485	0.852	
Total	EQ-5D	1063	0.728	0.132	0.707	0.062	1	0.522
	SF-6D	908	0.704	0.215	0.696	0.316	1	

*Other: includes cardiothoracic and neurosurgical referrals; VS denotes vascular surgery
Note: Only baseline data are reported here, therefore the descriptive statistics are different from those reported in table 1.

i. Consistency of EQ-5D and SF-6D

Descriptive statistics for the two indices, along with various measures of agreement are presented in Table 5. Overall, the mean value for EQ-5D index (0.755) exceeded the mean for the SF-6D index (0.728), and the difference was statistically significant. A scatter plot of the consistency sub-population is displayed in Figure 2. A positive association can be observed, with an R^2 value of 0.4043. There are substantial deviations from the 45 degree line of perfect agreement, which is particularly evident at the lower end of the scale, where a greater number of EQ-5D responses were reported. Pearson's product moment correlation (0.633; 95% C.I. 0.612 - 0.658; P-value<0.001), reflects this disagreement. Similarly, using Spearman's correlation coefficient, only moderate correlation was found across the whole consistency sample (0.645; 95% C.I: 0.620 - 0.665; P<0.001).

Figure 2: Bivariate scatter plot of EQ-5D and SF-6D index scores with linear regression line, consistency sample.



The intra-class correlation (ICC) revealed poor agreement between the instruments across all interview periods.[6] At baseline, the ICC for agreement was 0.522 and over the interview periods ranged from 0.522 to 0.627. The agreement between the instruments was further examined using the concordance correlation coefficient (Lin 1989). This statistic measures the extent of deviation from the 45 degree line of perfect agreement, and combines it with Pearson's correlation (r) to produce a statistic (ranging from 0 to 1) that measures both accuracy and precision (Lin and Torbeck 1998). A concordance correlation coefficient of 0.455 was found at baseline, and was consistently lower than its corresponding intra-class correlation coefficient across all interview periods.

Data were checked to assess suitability for measuring the Bland-Altman limits of agreement statistic,[7] and then the limits of agreement were calculated using the transformed responses at baseline (Figure 3). The graph plots the difference between the EQ-5D and SF-6D scores (on the y axis) against the mean of the two transformed scores (on the x-axis). A strong correlation would result in the data points being evenly distributed across the line of perfect average agreement, y=0 (Bland and Altman 1986). The distribution of data points shows substantial lack of agreement between the two indices, particularly at the lower end of the index scale. The mean difference between the transformed measures was 0.147, and the limits of agreement ranged from -0.215 to 0.510. Thus, for 95 percent of individuals in the consistency sample at baseline, the SF-6D index would be between 21.5 percent less and 51.0 percent greater than the EQ-5D index value. This indicates that there is poor agreement between the instruments.

Table 5: Descriptive statistics for the EQ-5D and SF-6D index scores at each measurement interval, consistency sample.

Period	Tool	N	mean	sd	t-stat	r	ICC	CCC	SE	95% C.I.
M0	EQ-5D	595	0.725	0.210	3.162*	0.572	0.522	0.445	0.021	0.473-0.554
	SF-6D	515	0.699	0.130						
M3	EQ-5D	616	0.756	0.218	6.686*	0.632	0.582	0.494	0.02	0.519-0.597
	SF-6D	533	0.725	0.134						
M6	EQ-5D	624	0.754	0.219	4.753*	0.654	0.586	0.541	0.02	0.53-0.607
	SF-6D	536	0.733	0.136						
M9	EQ-5D	470	0.773	0.212	4.246*	0.668	0.601	0.557	0.022	0.543-0.631
	SF-6D	420	0.746	0.130						
M12	EQ-5D	334	0.769	0.213	2.759*	0.657	0.619	0.55	0.027	0.557-0.662
	SF-6D	294	0.741	0.136						
M15	EQ-5D	243	0.774	0.206	2.133**	0.663	0.627	0.525	0.032	0.549-0.674
	SF-6D	201	0.736	0.130						
M18	EQ-5D	180	0.757	0.207	1.517†	0.599	0.572	0.501	0.042	0.443-0.606

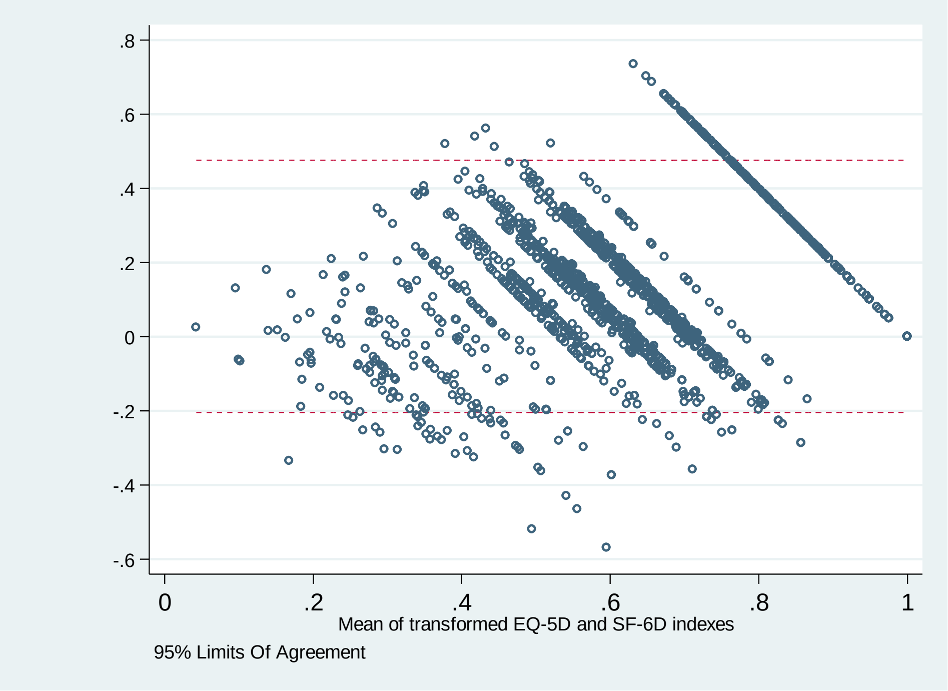
[6] The ICC equals 1 if there is perfect agreement, and 0 if there is no agreement.

[7] These require constant mean and standard deviation of differences between the instruments, and normal distribution between the differences.

SF-6D	160	0.740	0.128

Notes: sd denotes standard deviation of the mean; t-stat denotes t-statistic; ICC: intra-class correlation coefficient; r: Pearson's product moment correlation coefficient; CCC: concordance correlation coefficient; SE: standard error of the concordance correlation coefficient
*P<0.001, **P<0.01, † insignificant

Figure 3: Bland-Altman plot for consistency patients at baseline



Next we turn to the results of regression models aimed at examining the consistency between the two indices with longitudinal data. Even though the EQ-5D and SF-6D indices are designed to be on the same 0-1 scale, differences in the descriptive health states in the two questionnaires lead to actual index values having a more restricted range with the EQ-5D having a ceiling effect and the SF-6D a floor effect (Drummond, Sculpher et al. 2005); in our sample the EQ-5D has a minimum value of 0.0072 and the SF-6D has a minimum of 0.301. In order to compare the two it is necessary to re-scale the actual values of the two indices to a 0-1 range and we do this prior to estimating the regression models.[8]

We first examine the relationship between the two indices cross-sectionally, i.e. at each interview. A patient fixed-effects model is estimated at each interview round, with the two methods making up the clusters at each interview; the fixed-effects specification controls for all observed and unobserved patient-level factors, and allows us to focus attention on the difference between the

[8] The index value for each patient (for each method) is rescaled by subtracting the minimum of the scale (in the sample) and dividing by 1 minus the minimum.

two indices. Table 7 shows that after controlling for all observed and unobserved factors the SF-6D has a consistently lower health utility value, which is in the 0.13 to 0.16 range.

The availability of multiple measurements for both indices permits a more efficient estimation of the difference between the two indices, and in the next step we combine (stack) data from all interviews and account for method and patient clustering with a three-level variance components model. Results of this model are presented in Table 8, and it shows that the mean difference between the two indices is of the order of 0.14 and the coefficient is estimated much more precisely than the cross-sectional specification. Estimates of the variance components are also very significant indicating a better specified model.

Table 7: Results of interview-specific patient fixed-effects regressions for examining consistency between EQ-5D and SF-6D indices.

Interview	Dummy: SF-6D	S.E.	Constant	S.E	n	F-statistic	Adjusted R^2
Baseline	-0.146	0.007	0.721	0.005	1320	379.01	0.597
3 months	-0.152	0.007	0.764	0.005	1240	454.59	0.665
6 months	-0.137	0.007	0.753	0.005	1246	362.88	0.665
9 months	-0.131	0.008	0.766	0.006	936	281.21	0.683
12 months	-0.125	0.009	0.756	0.006	676	189.98	0.700
15 months	-0.140	0.010	0.765	0.007	460	181.73	0.709
18 months	-0.126	0.013	0.754	0.009	372	99.99	0.642

Note: S.E. denotes standard error of dummy or coefficient

Table 8: Results of a **t**hree-level variance component model for examining the consistency between EQ-5D and SF-6D.

Three-level model	Coefficient	S.E.	P-value
SF-6D	-0.140	0.004	0.000
Constant	0.757	0.006	0.000
Random-effects parameters			
Patient	0.143	0.004	
Method	0.042	0.004	
Residual	0.141	0.001	
ICC	0.486	-	

Note: S.E. denotes standard error of coefficient

In conclusion it is very clear from these data that there is poor agreement between the EQ-5D and SF-36 indices. Previous studies have demonstrated this with cross-sectional data, and we are able to confirm this finding with longitudinal repeated measurements.

ii. Sensitivity of EQ-5D and SF-6D indices for measuring change in health-related quality of life

Turning to the sensitivity of the two indices for measuring change in the quality of life that can be expected with a well-defined health (elective surgery) we first examine the pre and post-surgery distribution of each index in Figures 4 and 5. The kernel density estimates show quite clearly that there is a positive shift in both distributions which we explore further with regression models.

Figure 4: Kernel density plot of EQ-5D index scores pre- and post intervention

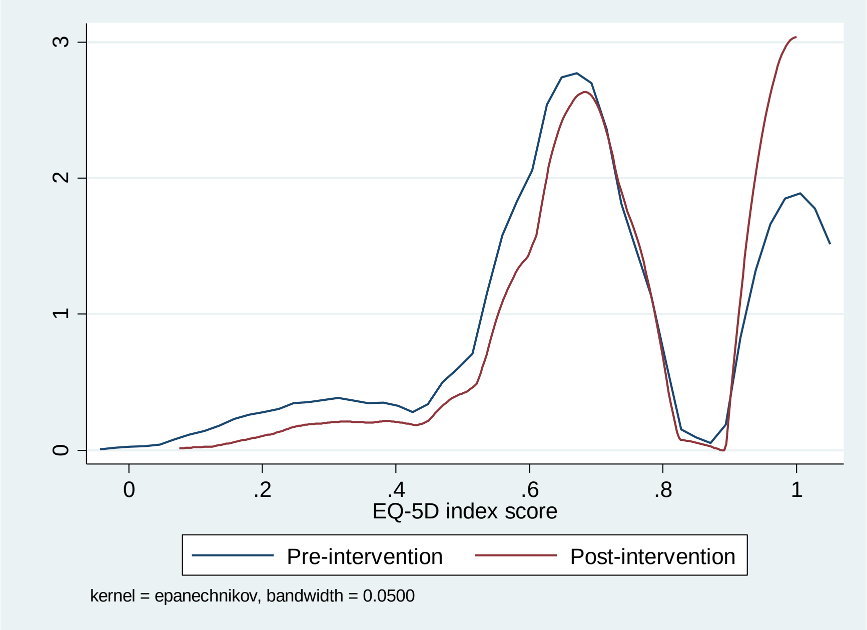
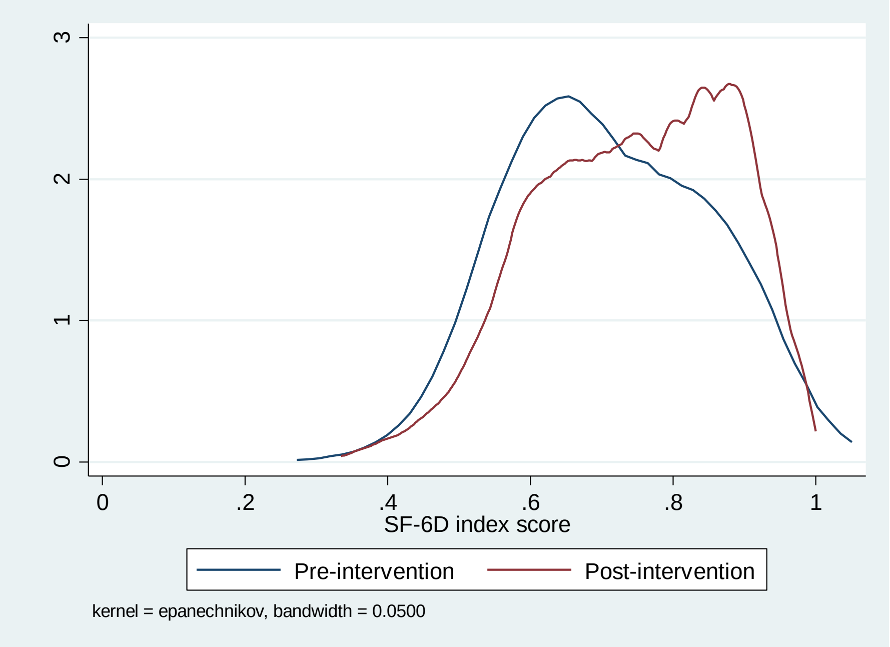


Figure 5: Kernel density plot of SF-6D index scores pre- and post intervention



We first estimate a patient fixed-effects model for each method separately (Table 9); the fixed effects specification controls for unchanging observed and unobserved patient characteristics.[9] The model includes dummy variables for all post-surgery interviews, and so their coefficients represent the change in mean health-related quality of life due to the intervention. The pre-surgery measurements make up the base group, and as a result the coefficient for the constant term measures mean pre-surgery quality of life. The results in Table 9 show a clear pattern of improvement in post-surgery quality of life, but there is an important difference between the two indices. While EQ-5D shows monotonic increases at every 3 months, improvement in the SF-6D index is evident only 3 months after the surgery, though thereafter the two indices follow a similar time path (Figure 6). It is not clear why the SF-6D is less – immediately – responsive to a significant health event, especially since it is EQ-5D which displays a greater lumping of values at the perfect health end of the scale's range; we intend to examine this more carefully in future research.

Table 9: Results of interview-specific patient fixed-effects regressions for examining sensitivity of EQ-5D and SF-6D indices in measuring change in health-related quality of life due to elective surgery.

	EQ-5D	**S.E**	**SF-6D**	**S.E.**
Constant	0.716*	0.005	0.582*	0.004
Post-intervention interview 1	0.026**	0.009	0.005†	0.008
Post-intervention interview 2	0.080*	0.010	0.096*	0.008
Post-intervention interview 3	0.097*	0.013	0.111*	0.012
Post-intervention interview 4	0.114*	0.019	0.112*	0.018
n	1465		1689	
F-statistic	28.37 (1, 1463)		51.46 (4, 1683)	
Adjusted R^2	0.597		0.580	

*P<0.0001; **P=0.005; †p=0.538

Next we estimate a combined three-level variance components model with data on both methods. A dummy variable is included for the SF-6D index to measure difference between the two methods. A series of dummy variables are included for the post-operative measurements to measure change in health-related quality of life relative to pre-surgery measurements, and four SF-6D x interview interaction terms are included to assess interview-specific differences between the two indices. Results of this regression are presented in Table 10 and these confirm the pattern observed in the method-specific regression (Table 9). The coefficient for the SF-6D is negative and significant and indicates that the mean SF-6D value is 0.13 lower than the mean EQ-5D index. As in Table 9, the

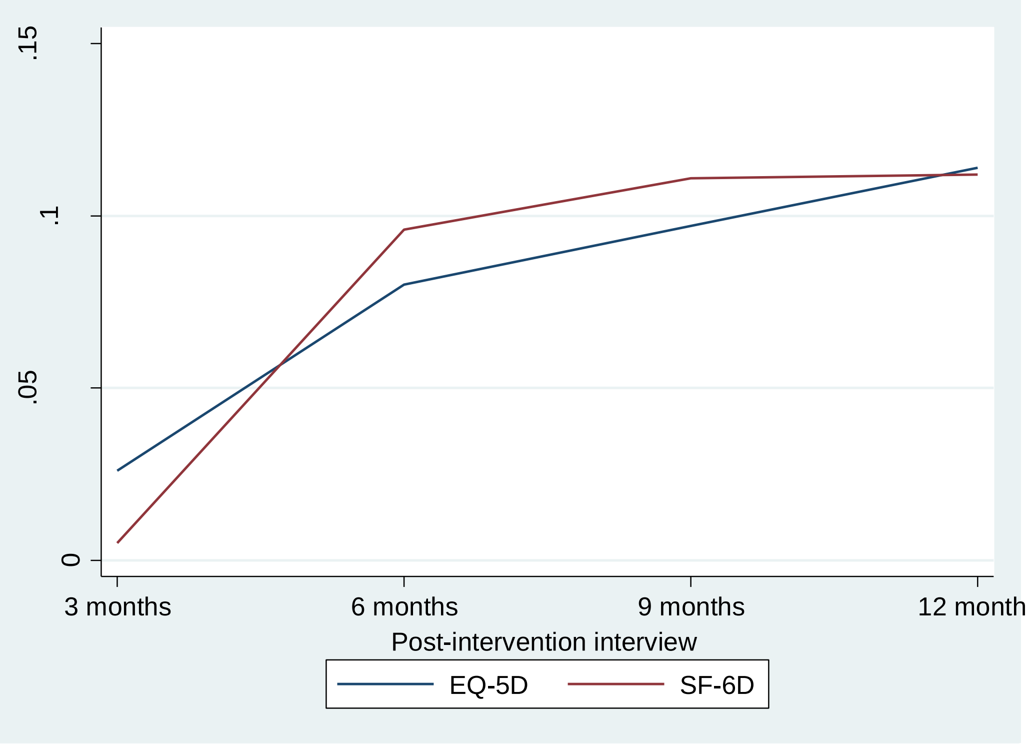
[9] While this is useful for isolating the change in health-related quality of life, it is not clear whether it is appropriate in the context of an important health event which might have a substantial influence on the time trend of excluded patient characteristics; we leave this issue for later research noting that a simple solution is to run a random effects model and determine model choice on the basis of the Hausman test.

coefficients for the interval dummy variables are significant and display a monotonically increasing pattern. The coefficient for the interaction term for method and the first post-intervention measurement is significant indicating that the two methods are indeed different in capturing the immediate change in quality of life after surgery; none of the other interaction coefficients are significant.

Table 10: Results of a three-level variance component model for examining the sensitivity of EQ-5D and SF-6D indices in measuring change in health-related quality of life due to elective surgery.

Three-level models	Coefficient	S.E.	P-value	95% C.I.
SF-6D	-0.130	0.007	0.000	-0.144 – -0.117
Post-intervention interview 1	0.030	0.008	0.000	0.014 – 0.046
Post-intervention interview 2	0.085	0.009	0.000	0.068 – 0.102
Post-intervention interview 3	0.102	0.012	0.000	0.079 – 0.126
Post-intervention interview 4	0.115	0.017	0.000	0.081 – 0.149
SF-6D*Post-intervention interview 1	-0.027	0.012	0.021	-0.050 – -0.004
SF-6D*Post-intervention interview 2	0.009	0.013	0.494	-0.016 – 0.033
SF-6D*Post-intervention interview 3	0.001	0.017	0.934	-0.033 – 0.035
SF-6D*Post-intervention interview 4	-0.009	0.026	0.722	-0.059 – 0.041
Constant	0.720	0.008	0.000	0.704 – 0.737

Figure 6: Sensitivity of the EQ-5D and SF-6D to treatment event



Concluding remarks

In this paper, we assessed the concordance of EQ-5D and SF-6D in patients referred for elective surgery and their responsiveness to clinical intervention. Our results, while preliminary, suggest that quality of life scores have different results depending on the type of instrument used. This study presents the following key findings: 1) The EQ-5D generated larger mean values than the SF-6D across all interview periods and the intra-class correlation coefficient between them was 0.486, indicating poor agreement; 2) there was substantial disagreement in the distributions of the index scores, particularly in the tails of the distribution; 3) the EQ-5D exhibited a substantial ceiling effect, and there was evidence of a floor effect in the SF-6D; and 4) the EQ-5D was more responsive than the SF-6D in detecting change in health status immediately following clinical intervention (within 3 months).

While substantial differences in mean health status are observed between referral specialties (in particular, orthopaedic referrals), the overall health status of patients referred for elective surgery in the present study was relatively good (mean EQ-5D = 0.763, SF-6D = 0.730). Further, the changes in health related quality of life were small, and occurred at the upper end of the scoring range. In this context, the SF-6D appears to be better suited in detecting and tracking changes in the health-related quality of life. However, it is important not to lose sight of the observation that the EQ-5D is more responsive to change in health status following clinical intervention.

The principal findings suggest that there is little agreement between the indices derived from the instruments, and they differ in their responsiveness to clinical intervention. Previous studies have demonstrates this with cross-sectional data, and we are able to confirm this finding with longitudinal repeated measures.

References

Bland, J. M. and D. G. Altman (1986). "Statistical methods for assessing agreement between two methods of clinical measurement." The Lancet **1**: 307-310.

Bosch, J. L. and M. G. M. Hunin (2000). "Comparison of the Health Utilities Index Mark 3 (HUI3) and the EuroQol EQ-5D in patients treated for intermittent claudicatio." Quality of Life Research **9**: 591-600.

Brazier, J. and M. Deverill (1999). "A checklist for judging preference-based measures of health related quality of life: Learning from psychometrics." Health Economics **8**(1): 41-51.

Brazier, J., R. Harper, et al. (1992). "Validating the SF-36 health survey questionnaire: new outcome measure for primary care." British Medical Journal **305**: 160-164.

Brazier, J., N. M. B. Jones, et al. (1993). "Testing the validity of the Euroqol and comparing it with the SF-36 health survey questionnaire." Quality of Life Research **2**(3): 169-180.

Brazier, J., J. Roberts, et al. (2004). "A comparison of the EQ-5D and SF-6D across seven patient groups." Health Economics **13**(9): 873-884.

Brazier, J. E. and J. Roberts (2004). "The estimation of a preference-based measure of health from the SF-12." Medical Care **42**(9): 851-859.

Brazier, J. E., J. Roberts, et al. (2002). "The estimation of a preference-based measure of health from the SF-36." Journal of Health Economics **21**: 271-292.

Brazier, J. E., S. J. Walters, et al. (1996). "Using the SF-36 and the Euroqol on an elderly population." Quality of Life Research **5**: 195-204.

Brooks, R. and F. De Charro (1996). "EuroQol: The current state of play." Health Policy **37**(1): 53-72.

Bryan, S. and L. Longworth (2005). "Measuring health-related utility: why the disparity between EQ-5D and SF-6D?" Eur J Health Econom **50**: 253-260.

Devlin, N. J., P. Hansen, et al. (2003). "Logical inconsistencies in survey respondents' health state valuations - A methodological challenge for estimating social tariffs." Health Economics **12**(7): 529-544.

Dowell, A., S. Morgan, et al. (2007). Access to elective surgery for patients referred under ACC. Wellington, Department of Primary Health Care and General Practice, Wellington School of Medicine and Health Sciences, University of Otago.

Drummond, M. F. (2001). "Introducing economic and quality of life measurements into clinical studies." Ann Med **33**: 344-349.

Drummond, M. F., M. J. Sculpher, et al. (2005). Methods for the economic evaluation of health care programmes. New York, Oxford University Press.

Fayers, P. and D. Machin (2007). Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes. Chichester, Wiley.

Grieve, R., M. Grishchenko, et al. (2009). "SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility." European Journal of Health Economics **10**: 15-23.

Gudex, C., P. Dolan, et al. (1996). "Health state valautions from the general public using the visual analogue scale." Quality of Life Research **5**(532-531).

Horowitz, E., I. Abadi-Korek, et al. (2010). "EQ-5D as a generic measure of health-related quality of life in Israel: reliability, validity and responsiveness." Isr Med Assoc J **12**(12): 715-720.

Johnson, J. A. and A. S. Pickard (2000). "Comparison of the EQ-5D and SF-12 Health Surveys in a General Population Survey in Alberta, Canada." Medical Care **38**(1): 115-121.

Kontodimopoulos, N., E. Pappa, et al. (2010). "Comparing the sensitivity of EQ-5D, SF-6D and 15D utilities to the specific effect of diabetic complications." Eur J Health Econ **Dec 5 [Epub ahead of print]**.

Lin, L. (1989). "A concordance correlation coefficient to evaluate reproducibility." Biometrics **45**: 255-268.

Lin, L. and L. D. Torbeck (1998). "Coefficient of accuracy and concordance correlation coefficient: new statistics for methods comparison." PDA Journal of Pharmaceutical Science and Technology **52**(2): 55-59.

Longworth, L. and S. Bryan (2003). "An emperical comparison of EQ-5D and SF-6D in liver transplant patients." Health Economics **12**: 1061-1067.

Marra, C. A., J. C. Woolcott, et al. (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**(7): 1571-1582.

Parkin, D. and N. Devlin (2006). "Is there a case for using visual analogue scale valuations in cost-utility analysis?" Health Economics **15**: 653-664

Petrou, S. and C. Hockley (2005). "An investigation into the emperical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population " Health Economics **14**: 1169-1189.

Pharmaceutical Purchasing Agency (2009). Cost-utility analysis (CUA) explained. Wellington, PHARMAC.

Raymont, A., S. Morgan, et al. (2002). "New Zealand general practitioners' non-urgent referrals to surgeons: who and why?" NZMJ **121**(1275): 57-34.

Scott, K. M., M. I. Tobias, et al. (1999). "SF-36 health survey reliability, validity and norms for New Zealand." Australian and New Zealand Journal of Public Health **23**(4): 401-406.

Søgaard, R., F. B. Christensen, et al. (2009). "Interchangeability of the EQ-5D and the SF-6D in Long-Lasting Low Back Pain." Value in Health **12**(4): 606-612.

Tsuchiya, A., J. Brazier, et al. (2006). "Comparison of valuation methods used to generate the EQ-5D and the SF-6D value sets." Journal of Health Economics **25**(2): 334-346.

Van Stel, H. F. and E. Buskens (2006). "Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease." Health and Quality of Life Outcomes **4**.

Zhao, F. L., M. Yue, et al. (2010). "Validation and comparison of euroqol and short form 6D in chronic prostatitis patients." Value in Health **13**(5): 649-656.