Work in progress: not to be quoted without authors' permission

Paper for discussion at Health Economists' Study Group Meeting
University of Leeds
8-10 January 2003

A comparison of preference-based health-related quality of life instruments in haemodialysis patients: EQ-5D versus SF-6D

T Nicholson[1], K Gerard[1], M Mullee[1], R Mehta[1], P Roderick[1], A Armitage[2]

[1] Health Care Research Unit, University of Southampton, Level B (805), South Academic Block, Southampton General Hospital, Tremona Road, Southampton SO16 6YD
Email T Nicholson: apn@soton.ac.uk; Tel 02380 794775

[2] John Bright Renal Unit, Southmead Hospital, Bristol

Acknowledgements

This was part of a wider national study on the effectiveness, acceptability and accessibility of haemodialysis set in renal satellite and main units. We are grateful to the other members of the National Renal Satellite Evaluation Group, patients and hospital staff from the units that participated in the study. The project was funded by the NHS R&D Health Technology Assessment Programme. The views expressed in this paper and any errors are those of the authors alone.

1. Introduction

It is common in health services outcome research to require a number of outcome measures when assessing health-related quality of life (HRQoL). These may be needed for a number of different reasons. For example, to meet competing objectives within the same study such as assessment of economic and clinical performance; or a desire to test methodological issues surrounding the development of new instruments. Whatever the particular raison d'êtrè for HRQoL measurement, the researcher must carefully select the most appropriate instrument(s) from the many varieties that are available. In an ideal world this means using the instrument(s) expected to generate the most reliable and valid data within the resources available. In some situations selection can be quite challenging if little prior information exists on which to base the selection. This is a common problem for new areas of application or the arrival of a new instrument. Some of this uncertainty can be handled by following good practice, that more than one instrument is used within a study; at least one main one and alternative ones which are selected on the basis they could verify the performance of the main one(s) for the intended purpose.

In this paper we are interested in the use of so called 'off the peg' multi-attribute preference measures for economic evaluation. The EQ-5D is currently the mainstay generic preference instrument used for health benefit assessment in economic evaluation studies conducted throughout the UK and Europe[1]. There is much evidence to show it is a reliable and valid instrument for a number of patient groups and general populations[2]. However, recent publication of preference weights for a new measure, SF-6D[3,4] expands the researchers' available toolkit for studying preferences in UK populations. The SF-6D is based on a sub-set of items from the widely used health profile measure, SF-36[5]. Currently the SF-6D has limited evidence to demonstrate its validity and reliability but this will change as it becomes more widely used and we learn about it. The health services outcome research community is now in the position where it needs to understand the full ramifications of this new instrument. In particular, under what conditions could it be considered a more suitable alternative to EQ-5D or, indeed, complementary to it? In practice at least, there is good reason to speculate SF-6D could be an attractive alternative given the pressure on health economists to limit the amount of data collected in favour of clinical outcomes. Moreover, as the health state classification framework is broader it has a larger number of possible health states (18,000 compared to the 243 used in the EQ-5D) and hence could be regarded as more sensitive. (Appendix 1 describes the health state classification systems of EQ-5D and SF-6D and presents the 11 SF-36 questions used to derive the SF-6D classification). SF-6D uses self-assessed health status based over a longer time period than EQ-5D (one month, except for the physical functioning domain that relates to a typical day). In this sense could be regarded as a more robust measure of average utility, which could be particularly germane for longitudinal studies.

1.1. Caring for patients undergoing chronic haemodialysis

This paper applies HRQoL instruments to patients undergoing hospital-based chronic haemodialysis, a renal replacement therapy for people with end-stage renal disease (ESRD). ESRD has been shown to significantly impact patients' quality of life in a number of different health domains[6], although it is fatal if untreated. Many of these patients have other associated co-morbidities, such as heart problems and diabetes which contribute to the HRQoL burden.

In addition to haemodialysis, other treatment modalities available are peritoneal dialysis and kidney transplantation although the latter is not available to most patients due to the shortage of donor kidneys. Dialysis is invasive and only removes the toxic breakdown products and fluid balance, it does not cure other renal dysfunction. It has been shown to have a serious impact on patients' quality of life. For example, mean time-trade off utilities from one study were 0.43 for patients on hospital haemodialysis, compared with 0.84 for patients after kidney transplantation[7]. Another study found that mean EQ-5D utilities varied from 0.66 to 0.81 depending on the treatment modality, the lowest values being for haemodialysis[8].

Typically patients on chronic haemodialysis attend a renal unit three times per week. Haemodialysis itself takes approximately 4 hours to complete during which time patients are immobilised. Furthermore, additional time above and beyond dialysis is required to monitor the patient, connect them to the machine, wait for NHS transport[a] and make the journey to and from the unit. All of this

[a] As haemodialysis patients are entitled to free NHS transport they are at the mercy of the local NHS transportation system, this can mean very long waits for transport or less than direct journeys.

then adds up to the patient experiencing great difficulties performing their usual roles and activities of daily living, to say nothing of the impact it has on them emotionally.

To date the impact of the disease and its treatment has been measured using a variety of HRQoL instruments. There is the usual gamut of generic HRQoL measures available (including EQ-5D) and (at least one) validated disease specific measure, the Kidney Disease Quality of Life Instrument (KDQoL™)[9]. The KDQoL™ includes the SF-36 as a generic core and is supplemented with multi-item scales targeted at particular concerns of ESRD patients such as symptoms commonly experienced in kidney disease, the effects on daily life, burden of kidney disease, cognitive function, work status, sexual function, quality of social interaction, and sleep.

In 1996 the National Renal Purchasing Guidelines[10] were distributed to health authorities. These served as a guide to commissioning effective renal care and included the recommendation to develop renal satellite units (RSUs). RSUs were seen to be an alternative to main renal units (MRU). MRUs provide full renal services that could help expand the capacity of haemodialysis services and improve their geographical accessibility. RSUs are 'attached' to MRUs, but in contrast provide a chronic maintenance haemodialysis service, largely run by nurses, and often in populations at some distance from the main unit. However, there was an expectation that RSUs would be as clinically effective and safe and quite similar with respect to HRQoL; the key differences anticipated were patient satisfaction, patient preference and costs although the direction and quantity of these potential changes were difficult to predict. The Guidelines hastened the development of RSUs and the decentralising of renal services. Between 1993 and 1998 the number of haemodialysis stations (i.e. available dialysis machines) within main units increased by 37% whilst there was a 300% increase within RSUs[11, 12]. Our main study was commissioned to evaluate the impact of RSU care *in toto* (see below for more details).

2. Aim and Objectives

The reported study was an off-shoot of the main study and aimed at addressing the question 'which, if either, of the two preference-based measures of health benefit ($EQ\text{-}5D_{Utility}$ and $SF\text{-}6D_{Utility}$) performed better in a group of renal patients on haemodialysis?'. Two objectives were used to address this aim:

1. to separately assess performance of the individual measures; and
2. to compare their performance.

3. Methods

3.1. Main haemodialysis study and data

The data reported in the paper were part of the wider study commissioned to investigate the effectiveness, costs, acceptability and accessibility of RSUs. A concern was whether RSUs were as effective and safe as MRUs in meeting the Renal Association standards[13] for chronic haemodialysis patients. To address this, the study was designed as a cross sectional comparison of 736 patients from a stratified sample of 24 renal units located throughout England and Wales (or 12 pairs of main / satellite units)[14]. In the present analysis, since we are interested in the performance of the two preference measures, we analyse the data for the whole patient group (not by setting) and concentrated on a sub-set of the main study variables.

There were two main data collections used to extract patient-level information. One of these was a patient completed questionnaire (referred hereafter as the 'PQ'), the other a clinical questionnaire completed by members of the research team (referred hereafter as the 'CQ') and used to extract clinical information from medical notes and the renal unit's computer systems. The PQ was handed out to patients during routine haemodialysis sessions. It included three HRQoL instruments: EQ-5D, SF-36, and KDQoL™ and health service use and socio-demographic questions (e.g. age, gender, ethnicity, living arrangements).

The EQ-5D was selected as the main economic outcome measure. Patients were also asked to rate HRQoL using the EQ-5D visual analogue scale ($EQ\text{-}5D_{VAS}$) under two situations; when not on dialysis and undergoing dialysis. The rationale for this was that undergoing haemodialysis is not pleasant and

could be seen as placing significant additional restrictions (physical and mental) on the patient's HRQoL.

The clinical study relied on the SF-36 as its main HRQoL measure. The SF-36 comprises eight multi-item scales that can be combined into the physical and mental component weighted scores (standardised on the US population) - high scores indicate good quality of life[15].

The CQ collected information on a number of indicators pertaining to the process of renal care and haemodialysis such as site of vascular access, frequency and duration of dialysis, duration of renal replacement therapy (RRT), renal-related medication (including erythropoietin (EPO), biochemical and haematological measures (URR[b] and haemoglobin[c]), co-morbidity and functional performance. A number of co-morbidity scores were obtained but here we only report the Khan Index. It classifies patients into one of three risk groups based on co-morbid disease and age at onset of RRT (i.e. high, medium, low risk)[16]. Functional performance was nurse-assessed using the Karnofsky Performance Score that allowed the patient to be grouped into one of 11 ordered levels of functioning[17].

EQ-5D and SF-6D utilities (referred hereafter as EQ-$5D_{Utility}$ and SF-$6D_{Utility}$ respectively) were derived from respective patient-level health classification information and applying the appropriate published algorithms, obtained in both cases from UK population samples[d] (EQ-$5D_{Utility}$ from 3,395 representative members of the general population and SF-$6D_{Utility}$ from 611 members of the general population[18, 4]).

It is important to point out differences between the two utility measures with respect to data requirements for mapping health status to pre-scored utility weights. For EQ-$5D_{Utility}$ if health status information is missing across one or more of the five domains, an EQ-$5D_{Utility}$ cannot be derived unless imputed values are used. (For the purposes of this study we have not imputed missing values.) This means there are potentially two sources of missing data – health state classification data missing completely or partially – and only one source of complete utility data. In the case of the SF-$6D_{Utility}$, the picture is more complicated. Utilities can be attributed to complete or some partially completed health state classifications (derived from 11 of the SF-36 questions – see Appendix 1). Appendix 2 shows the impact of this diagrammatically. Furthermore, overall response to SF-6D could be much lower than for EQ-5D as the 'SF-6D questions' are embedded within the SF-36, a much longer instrument than EQ-5D and thus more susceptible to patient fatigue.

3.2. A framework to assess instrument performance

Brazier and Deverill recommended adaptation of standard psychometric criteria for the assessment of preference-based measures of HRQoL[19]. They produced a comprehensive checklist for judging preference-based measures, laying claim that this covers a larger range of characteristics than is usually discussed in the health economics literature. Importantly they argue since preference-based measures examine people's preferences for (or value of) different health states or changes in them, not the health status per se, such measures cannot be evaluated adequately by psychometric criteria. Hence valuations of a change in health status will vary according to a person's circumstances and so should not be expected to perfectly correlate with the size of the health state change. The term checklist was deliberately used to recognise that at the time of publication (and this remains the case today) there is no consensus in the literature on what constitutes the key criteria and how they are to be applied, but rather the authors' regarded their contribution as providing a comprehensive guide that should be considered by researchers involved in using such outcome measures on a case by case basis. We have approached their checklist in this spirit.

We set out to apply the checklist in a systematic way to separately assess EQ-$5D_{Utility}$ and SF-$6D_{Utility}$ in our group of haemodialysis patients. We added an additional set of criteria in order to compare the relative merits of these measures. Table 1 summarises the checklist and relates each item to the potential for assessment within our study. Clearly, the design of the main study shaped the breadth

[b] URR (urea reduction ratio) is a measure for assessing the quality of dialysis.

[c] Anaemia, as assessed by haemoglobin concentration, is an important problem in the management of dialysis patients. Anaemia causes breathlessness and fatigue and hence potentially could have a marked impact on patients' HRQoL.

[d] EQ-5D utilities were derived from the time trade-off method, SF-6D from standard gamble.

and depth of the assessment carried out. We did not assess dimensions of reliability as this would have required separate data collection that was beyond the scope of the commissioned study. Furthermore, some aspects of validity were based on opportunistic use of data, rather than what was desirable. However, by taking this approach we maximised the use of the data available and it has allowed us to begin to provide evidence of the two measures' practicality and validity and identify where some of the key gaps in our knowledge lie.

Table 1: Checklist for judging the merits of EQ-5D$_{Utility}$ and SF-6D$_{Utility}$

Component	Potential for assessment
A. Practicality	
• Time to complete instrument	Limited assessment
• Response rate (overall)	Assessed
• Completion rate (individual items)	Assessed
B. Reliability	
• Test re-test / Implications for sample size / Inter-rater / Between places of administration	Not possible
C. Validity	
Description	
• Content validity: Coverage of all dimensions of health of interest / sensitivity	Limited assessment
• Face validity: Items relevant and appropriate for the population	Limited assessment
• Construct validity: Can the unscored classification of the instrument detect known or expected differences or changes in health?	Limited assessment
Valuation methods	
• Assumptions about preferences seem credible	Taken as appropriate
• Technique of valuation choice-based	Taken as appropriate
• Quality of data	Taken as appropriate
Empirical	
• Evidence for the empirical validity of the instrument against revealed, stated or hypothesised preferences	Limited assessment
D. Comparative performance	
• Comparing: instrument properties from Brazier and Deverill's checklist	Limited assessment
• Comparing: descriptive statistics and use of health states	Limited assessment

3.2.1. Practicality

Practicality covered three areas that we were able to comment on. First, we were only able to perform a limited assessment of the time to complete each instrument, as it was not directly recorded. Our judgement was based on anecdotal evidence from previous studies that had direct information about EQ-5D and by counting the number of items that would need to be read and answered in each case. The second aspect, response rate, provided a global indicator of the instruments' practicality. It measured the proportion of responses obtained from the total number of respondents who agreed to complete the PQ. Completion rate, on the other hand, indicated the usefulness of the data obtained; in this case whether it could be used to derive utilities. The proportion of respondents providing the necessary data to generate utilities assessed this.

3.2.2. Reliability

As mentioned earlier we did not assess reliability, although this is recognised as an important gap in our evidence and one for future research.

3.2.3. Validity

Content and face validity are closely associated properties of an instruments' descriptive validity. They are important criteria in preference-based measures as we need to be assured that the values generated are derived from accurate descriptions of health. In this case, a judgment on the appropriateness of the EQ-5D and SF-36 (not SF-6D per se) descriptors for our particular patient group was taken at the time the PQ was developed. Arguably this is a limited assessment of

descriptive validity but it seemed reasonable given that a) the measures have been successfully applied to chronically ill patient groups before now and b) the views of three experienced clinicians working in the field (2 Nephrology Consultants and a Specialist Registrar) and an experienced quality of life researcher (who had spent substantial earlier research working with this type of patient group) supported the view that the domains and items used in both instruments seemed sensible (face validity) and reasonably comprehensive (content validity)[e]. In addition, the items from the two instruments were compared with the disease specific items (i.e. KDQoL^{TM}) to determine whether there was any overlap.

The notion of construct validity that we were able to embrace was limited to testing for known or expected similarities in health (not health status change) using the unscored items of particular health domains (convergent validity). We tested all five dimensions from EQ-5D that were expected to correlate with items from SF-36 (four of which were also part of SF-6D). EQ-5D$_{usual\ activities}$ and EQ-5D$_{pain/discomfort}$ domains were each tested against three items; EQ-5D$_{mobility}$ was tested against two items; and EQ-5D$_{self\text{-}care}$ and EQ-5D$_{anxiety/depression}$ were each tested against one item. In addition, EQ-5D$_{mobility}$, was tested against a separate data item 'wheelchair use' (see Table 4). As mentioned, four items from SF-36 previously tested against EQ-5D were common to SF-6D. Two further items from SF-6D's physical functioning subscale were tested against wheelchair use.

Correlations between utilities and general health questions obtained from EQ-5D$_{VAS}$, SF-36 (sf1) and KDQoL^{TM} (kd22) were examined. Since each general health question used different response categories it was not expected to find perfect or even necessarily high correlation, but rather a more moderate expectation was of a positive association. In addition, we examined the convergent validity of EQ-5D$_{Utility}$ and SF-6D$_{Utility}$ by examining the correlation between them.

The valuation methods used in both instruments were taken as appropriate as they related to the original development of each instrument. Brazier and Deverill[19] cautioned that researchers using preference-based measures should understand the assumptions about preferences. The criteria that a choice based valuation technique should be used was met for both EQ-5D and SF-6D but a more extensive discussion of the pros and cons of the valuation methods is beyond the scope of this paper.

A key differentiating property of a preference-based measure is evidence of empirical validity. At this juncture the driver is to 'assess whether or not a measure could generate values which reflect people's preferences' (Brazier and Deverill[19]). There are three levels of evidence that are admissible: revealed preference data, stated preference data and hypothesised preferences. It is the latter that we were able to draw on in this study. We examined whether EQ-5D and SF-6D utilities were associated with a number of hypothesised preference relationships. The following lists the hypotheses we considered:

- Negatively associated:
    - Age (since increasing age yields lower utility in the UK population norms for EQ-5D[18])
    - Duration of renal replacement therapy (RRT, assuming that longer duration on RRT would yield lower utilities due to impact of the treatment regime, although these may be the fitter patients)
    - Dialysis frequency (twice weekly affecting HRQoL less than thrice weekly due to impact of the treatment regime, although the dialysis adequacy may be poorer since thrice weekly is the recommended frequency)
    - Disability (having disability - blindness, wheelchair use or other yields lower utility)
    - Co-morbidity (measured using Khan Index)
- Positively associated:
    - URR (using the Renal Association standards as a cut-off i.e. = 65% represents less adequate dialysis)
    - Haemoglobin (using the Renal Association standards as a cut-off i.e. = 10 g/dl indicating anaemia)
    - Karnofsky performance score (higher scores indicate greater functional ability)
    - Age of leaving full-time education (higher levels of education have been associated with better HRQoL from the UK population norms[18])
    - Car ownership as a proxy for wealth

[e] Clearly the KDQoL^{TM} was expected to be the most comprehensive.

3.2.4. Comparative performance

There were two aspects to the comparative exercise. The performance of the two measures was compared and contrasted according to the Brazier and Deverill checklist[19], thus deriving an overall view on their overall performance. However, to avoid repetition, this was included under each section of the checklist. Secondly, we examined the relative sensitivity of EQ-5D and SF-6D. Ideally we would want to compare the sensitivity of the two instruments with respect to health differences and changes. But as this was not possible we have chosen to begin the process through a description of the instruments with respect to central tendency, variation and use of the available health states.

3.3. Analysis

As $EQ\text{-}5D_{Utility}$ and $SF\text{-}6D_{Utility}$ data were skewed (see Appendix 3) appropriate non-parametric analyses were performed in SPSS v11. These analyses depended on the nature of the data. The Chi square test of association was used on categorical data (with 2-4 categories). For continuous data (or categorical data with = 5 categories) compared with categorical data, the Mann-Whitney test was used where there were only 2 categories; and Jonckheere-terpstra test for ordered categorical data with = 3 categories. Spearman's correlations were used on continuous and ordered categorical data. As there are no absolute definitions of 'correlation', the 5 categories used to interpret Kappa scores were applied (i.e. < 0.2 = poor; 0.21 - 0.4 = fair; 0.41 - 0.6 = moderate; 0.61 - 0.8 = good; 0.81 - 1.0 = very good[20]. Statistical tests for agreement could not be used as the items did not use identical scales.

We undertook multiple statistical tests and therefore those found to be statistically significant needed to be interpreted with some caution. A Bonferroni correction was not used as this may be overly conservative in what was an exploratory study.

4. Results

4.1. Response and background characteristics

The main study used responses from 736 patients on chronic haemodialysis. Of these, 626 (85%) agreed to complete the PQ and form the relevant sample for the present analysis. Table 2 presents some key characteristics of the sample.

Table 2: Sample characteristics

Characteristic		N	Mean	Std Dev	%
Demographic/Economic/Social:					
Age		625	59.2	17.0	
Gender –	Male	392			62.6
Ethnicity –	Caucasian	510			85.3
	Other	88			14.7
Living arrangements – alone		144			24.1
Disability (e.g. blind or wheelchair bound)		141			23.0
Education (age left full-time education):					
0 - 15 years old		254			44.0
16 - 18 years old		221			38.3
>18 years old		102			17.7
Employment (in last 30 days):					
Employed (full or part-time) or student		109			18.7
Unemployed		47			8.1
Retired		264			45.4
Looking after home/family		25			4.3
Permanently sick or disabled		123			21.1
Other		14			2.4
Car ownership	0-1 car	470			79
	= 2 cars	125			21
Renal health and health service use:					
Duration of RRT (years)		601	5.2	5.7	
Dialysis frequency –	Thrice weekly	563			97.2
	Twice weekly	16			2.8
Total dialysis time (hours per week)		572	11.2	1.7	
Hospitalised in previous 12 months		251			42.5
Quality of dialysis –	URR (%)	591	69.2	7.6	
Correction of anaemia –	Haemoglobin (g/dl)	613	11.3	1.6	
Co-morbidity (Khan Index) –	Low	207			35.5
	Medium	203			34.8
	High	173			29.7
Functional assessment (Karnofsky Score)					
100 Fully functional		24			3.9
90 Minor problems		257			41.3
80 Some problems		161			25.9
Remaining levels combined (i.e. 0 to 70)		180			28.9

The mean age was 59 years, patients were predominately Caucasian and more than half were male (63%). Nearly a quarter of the patients lived alone and had some disability (including blindness or wheelchair use). A large number of patients (44%) had left full-time education by 16 years old. As expected, the majority were retired or permanently sick / disabled (67%), however there was still 19% that were employed or students. Car ownership was used as a proxy for wealth: only a fifth of households owned two or more cars. (This could also be taken as a reflection of fitter patients, since there is less likely to be a second car in the household if the patient cannot drive due to disability or illness.) The mean duration of time spent on renal replacement therapy (i.e. all treatments for renal failure, not just haemodialysis) was 5.2 years. However there was a wide variation in this with a median of 3.0 years.

The majority of patients underwent haemodialysis three times per week and so the mean total weekly dialysis time (11.2 hours) was close to that expected (i.e. 12 hours). Forty-three percent of patients were hospitalised at least once in the preceding 12 months. The mean URR, a measure of dialysis adequacy, was above the minimum standard recommended by the Renal Association (65%). Likewise, mean haemoglobin was above the recommended minimum threshold (= 10g/dl). As expected within this patient group, there were high levels of co-morbidity (only 36% patients were categorised as 'low' risk). Approximately 45% of patients were fully functional or only had minor problems as assessed by the Karnofsky performance score.

4.2. Practicality

Previous experience suggests EQ-5D takes about 5 minutes to complete and we had no evidence to suggest it was different in this application. EQ-5D comprises a total of 15 items to be read and answered. Based on experience with SF-36, the SF-6D is straightforward to complete although much longer than EQ-5D. It required 47 items to be read and answered. On the basis of the number of items alone, it is likely that SF-6D will take approximately three times longer to complete, closer to 15 minutes.

Overall response rates for EQ-5D and SF-6D were both 85% (i.e. study patients who agreed to do the PQ). Completion rates to derive utilities (see Appendix 2) were 93.1% for EQ-5D and 79.2% for SF-6D. $EQ\text{-}5D_{Utility}$ data was available for 87 (13.9%) more respondents. It is likely that the completion rate was lower for $SF\text{-}6D_{Utility}$ due to the additional questions required. However, these figures conceal large differences in the amount of data missing for the reasons described at the end of Section 3.1.

We believe that some of the 15 people who did not complete any EQ-5D questions may have inadvertently turned over the page without seeing the questions. This was a very unlikely scenario for SF-6D as the questions are spread over several pages. In addition, whilst the EQ-5D is available for use free to public sector employees, charges have recently been introduced for SF-36 and hence SF-6D. Table 3 shows the comparison of responders and non-responders to $EQ\text{-}5D_{Utility}$ and $SF\text{-}6D_{Utility}$.

Table 3: Responders compared with non-responders for EQ-5D and SF-6D

		$EQ\text{-}5D_{Utility}$ availability			$SF\text{-}6D_{Utility}$ availability		
		N	Median age in years (IQR)	P value [a]	N	Median age in years (IQR)	P value [a]
Utility available | Age	Yes	582	61.5 (28.0)	0.005	495	60 (28.0)	<0.001
	No	43	70 (16.0)		130	70 (18.3)	

		$EQ\text{-}5D_{Utility}$ present					$SF\text{-}6D_{Utility}$ present				
		Yes		No		P value [b]	Yes		No		P value [b]
		N	%	N	%		N	%	N	%	
Gender	Male	370	63.5	22	51.2	0.108	319	64.3	73	56.2	0.087
	Female	213	36.5	21	48.8		177	35.7	57	43.8	
Ethnicity	Caucasian	486	85.9	24	75.0	0.091	410	84.9	100	87.0	0.573
	Other	80	14.1	8	25.0		73	15.1	15	13.0	
Age (years) left full-time education	0 – 15	237	43.2	17	60.7	0.132	189	40.2	65	60.7	<0.001
	16 – 18	215	39.2	6	21.4		197	41.9	24	22.4	
	>18	97	17.7	5	17.9		84	17.9	18	16.8	
Car owner-ship	0-1	443	78.4	27	90.0	0.129	379	78.1	91	82.7	0.287
	= 2 cars	122	21.6	3	10.0		106	21.9	19	17.3	
Khan Index	Low	201	36.8	6	16.2	0.025	181	39.0	26	21.8	0.002
	Medium	184	33.7	19	51.4		151	32.5	52	43.7	
	High	161	29.5	12	32.4		132	28.4	41	34.5	

[a] Mann-Whitney Test
[b] Chi square

The patterns of non-response appeared very similar for $EQ\text{-}5D_{Utility}$ and $SF\text{-}6D_{Utility}$. $SF\text{-}6D_{Utility}$ had 87 more incomplete cases which may explain why the age of leaving full-time education achieved statistical significance here but not $EQ\text{-}5D_{Utility}$. Non-response to both instruments was significantly associated with greater age and co-morbidity.

4.3. Validity

As discussed earlier, EQ-5D appeared (with the limited assessment made) to have content and face validity. SF-6D per se was not assessed (only SF-36), although given the larger number of descriptors, it is unlikely to be worse than EQ-5D. There was no overlap between EQ-5D or SF-6D and the condition specific dimensions of KDQoL™. However, it was unclear whether the KDQoL™ dimensions were important sources of utility for haemodialysis patients. The findings from the tests of convergent validity are presented in the Table below.

Table 4: Tests of convergent validity

Instrument	Hypothesised association (item from SF-36 unless stated otherwise)	N	P value
EQ-5D domain			
Mobility [d]	Wheelchair use (not from SF-36)	552	<0.001 [a]
	Health limits climbing 1 flight of stairs (sf7)	570	<0.001 [a]
	Health limits walking 100 yards (sf11)	563	<0.001 [a]
Self-care [e]	SF-6D item: Health limits bathing or dressing self (sf12)	579	<0.001 [a]
Usual activities [f]	SF-6D item: Were limited in the kind of work or other activities (due to physical problems) (sf15)	505	<0.001 [a]
	Physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups (sf20)	580	<0.001 [b]
	SF-36 Role-physical subscale	507	<0.001 [b]
Pain/discomfort [f]	SF-36 Pain subscale (comprises 2 items below)	599	<0.001 [b]
	SF-6D item: Bodily pain (sf21)	*596*	*<0.001* [b]
	SF-6D item: Pain interfered with normal work (sf22)	*582*	*<0.001* [b]
Anxiety/depression [f]	SF-36 Mental health subscale	570	<0.001 [b]
EQ-5D$_{Utility}$	EQ-5D$_{VAS}$ (Correlation coefficient 0.541)	557	<0.001 [c]
	sf1 (SF36) (Correlation coefficient 0.508)	573	<0.001 [c]
	kd22 (KDQoL^{TM}) (Correlation coefficient 0.479)	570	<0.001 [c]
SF-6D domain			
Physical functioning:			
• Health limits vigorous activities (sf3)	Wheelchair use (not from SF-36)	524	0.090 [a]
• Health limits moderate activities (sf4)		532	<0.001 [a]
SF-6D$_{Utility}$	EQ-5D$_{VAS}$ (Correlation coefficient 0.597)	481	<0.001 [c]
	SF1 (SF36) (Correlation coefficient 0.549)	481	<0.001 [c]
	KD22 (KDQoL^{TM}) (Correlation coefficient 0.532)	481	<0.001 [c]

[a] Chi square
[b] Jonckheere-Terpstra Test
[c] Spearman's rho
[d] Confined to bed combined with some problems due to low number of responses to the former
[e] Unable to wash/dress self combined with some problems
[f] EQ-5D questions relate to a day, SF-36 questions relate to past 4 weeks

The EQ-5D mobility question was significantly associated with wheelchair use. For the two SF-36 questions that comprise the SF-6D physical functioning scale, one was significantly associated with wheelchair use and whilst the other was not, although responses were in the direction expected.

All other items tested showed a significant association, although the tests could not give the strength of the association. These comparisons were for EQ-5D domains compared with SF-36 questions (some of which are also included in the SF-6D) or SF-36 subscales. The items available were not always ideal as the timeframe over which questions were asked varied between a day and the past 4 weeks, but they provided the best choices for the data available. The general health questions from EQ-5D$_{VAS}$, SF-36 and KDQoL^{TM} all showed moderate correlation with EQ-5D$_{Utility}$ or SF-6D$_{Utility}$.

In addition, in terms of direct convergent validity between EQ-5D$_{Utility}$ and SF-6D$_{Utility}$, the correlation was good (N = 477, Spearman correlation coefficient 0.705, P <0.001).

4.4. Empirical validity

Tables 5 and 6 show the analyses for the variables selected.

Table 5: Correlation between EQ-5D or SF-6D and variables

	EQ-5D Utility		SF-6D Utility	
	N	Correlation [a] (P value)	N	Correlation [a] (P value)
Age (years)	582	-0.031 (0.450)	495	-0.091 (0.043)
Duration of renal replacement therapy (RRT, years)	564	-0.023 (0.581)	478	0.026 (0.575)
Karnofsky performance score	580	0.317 (<0.001)	495	0.286 (<0.001)
[a] Correlation Coefficient = Spearman's rho				

Table 6: Association between EQ-5D or SF-6D and variables

		EQ-5D Utility			SF-6D Utility		
		N	Median (IQR)	P value	N	Median (IQR)	P value
Dialysis frequency per week	2 x	13	0.743 (0.442)	0.253 [a]	11	0.702 (0.246)	0.044 [a]
	3 x	525	0.689 (0.309)		450	0.606 (0.183)	
Disability - blind, wheelchair use or other	Yes	132	0.516 (0.502)	<0.001 [a]	105	0.577 (0.136)	<0.001 [a]
	No	439	0.691 (0.298)		338	0.626 (0.183)	
Co-morbidity: Khan index	Low	201	0.691 (0.217)	0.004 [b]	181	0.626 (0.175)	<0.001 [b]
	Medium	184	0.648 (0.276)		151	0.604 (0.187)	
	High	161	0.639 (0.467)		132	0.597 (0.156)	
URR	>65% [c]	417	0.689 (0.280)	0.454 [a]	362	0.607 (0.190)	0.499 [a]
	= 65%	135	0.620 (0.448)		108	0.617 (0.161)	
Haemoglobin	>10 g/dl [c]	446	0.689 (0.280)	0.950 [a]	383	0.612 (0.182)	0.774 [a]
	=10 g/dl	126	0.689 (0.280)		104	0.603 (0.179)	
Age when left full-time education	0-15 years	237	0.620 (0.406)	0.050 [b]	189	0.602 (0.196)	0.070 [b]
	16-18 years	215	0.691 (0.298)		197	0.621 (0.170)	
	> 18 years	97	0.689 (0.280)		84	0.610 (0.168)	
Car ownership	0-1	443	0.656 (0.413)	0.005 [a]	379	0.602 (0.164)	<0.001 [a]
	= 2	122	0.691 (0.234)		106	0.661 (0.215)	
[a] Mann-Whitney Test [b] Jonckheere-Terpstra Test [c] i.e. meets Renal Association standards							

From Tables 5 and 6 it can be seen that SF-6D$_{Utility}$ results followed the same pattern as EQ-5D$_{Utility}$, although there was some variation in those that achieved statistical significance (i.e. dialysis frequency and age left full-time education). It should be noted that the relationships between age of leaving full time education and either EQ-5D$_{Utility}$ or SF-6D$_{Utility}$ were not straightforward (linear). There was some evidence in the raw data for age of leaving full time education that patients had not interpreted the question as intended (i.e. very young and very old ages).

Correlation was poor between both EQ-5D$_{Utility}$ and SF-6D$_{Utility}$ and age and duration on RRT, and only fair for Karnofsky performance score. There were no significant associations between EQ-5D$_{Utility}$ or SF-6D$_{Utility}$ and URR or haemoglobin. The hypothesised associations between EQ-5D$_{Utility}$ or SF-6D$_{Utility}$ and other variables were statistically significant for disability, co-morbidity and car ownership.

There were some differences in the associations between EQ-5D and SF-6D utilities and other variables, although none were statistically significant. The correlation between EQ-5D$_{Utility}$ and duration on renal replacement therapy was in the hypothesised direction although poor, however it was in the opposite direction for SF-6D$_{Utility}$. In fact it is unlikely that the relationship is straightforward anyway, as mentioned earlier, as those on RRT the longest are probably the fitter patients. Likewise, SF-6D$_{Utility}$ operated in the opposite direction to EQ-5D$_{Utility}$, for URR.

4.5. Comparative performance

As mentioned previously, the comparison of the two measures according to the Brazier and Deverill checklist[19] has been covered in the previous sections (4.2 to 4.4). The descriptive statistics for the two instruments are shown in Table 7.

Table 7: Descriptive statistics for EQ-5D and SF-6D

	$EQ\text{-}5D_{Utility}$	$SF\text{-}6D_{Utility}$
N	583	496
Mean	0.600	0.637
SD	0.292	0.139
Median	0.689	0.610
IQR	0.280	0.179
Minimum	-0.349	0.296
Maximum	1	1

These data show that the mean utilities were similar, although there was slightly more variation in the median utilities. However, the distributions were different; $EQ\text{-}5D_{Utility}$ was negatively skewed and $SF\text{-}6D_{Utility}$ positively skewed. It is interesting that the distribution of $SF\text{-}6D_{Utility}$ was narrower than $EQ\text{-}5D_{Utility}$ as shown by the standard deviation and inter-quartile range. Whilst the maximum scores were both 1, the lowest score was negative for EQ-5D, although not the lowest possible score (-0.594). The scoring system for SF-6D does not allow for negative scores and in this sample the lowest value was 0.296, the lowest possible score.

Use of the health states was examined for people with a complete EQ-5D or SF-6D health state. Sixty-eight of the 243 EQ-5D health states (28%) were used to describe health on a typical day when not dialysing. Overall 67 (11.5%) people used the best health state (i.e. 11111). Eight EQ-5D health states accounted for 52.3% of health states used and 20 health states accounted for 79.2%. The worst health state was 23333, although the worst possible health state is 33333. In addition, 28 people did not provide complete information for a health state classification and 15 respondents had no data.

For SF-6D, 430 (2.4%) of the 18,000 SF-6D health states were used to describe health. Overall 3 (0.6%) people used the best health state (i.e. 111111). 182 health states accounted for 50% of health states used. The worst health state was 645655, which is the worst possible health state in SF-6D. 128 people did not answer sufficient questions to derive a complete health state classification. In addition two people had no health state classification. (One person had not answered sufficient questions and the other had not answered any questions.) (See Appendix 2)

The patients who used the best health state (111111) in SF-6D also used the best health state in EQ-5D (11111). However, the patients who used the worst heath state in SF-6D (645655) did not use the worst health state in EQ-5D, but rather 22333. Conversely, the patient with the 'worst' health state in EQ-5D (23333) had an incomplete SF-6D health state (5-2655). Tables 8 and 9 show the self-reported levels derived for EQ-5D and SF-6D.

Table 8: Self-reported EQ-5D levels

Level	**1 - No problems**		**2 - Some problems**		**3 - Extreme problems**	
Dimension	N	%	N	%	N	%
$EQ\text{-}5D_{Mobility}$	220	36.1%	387	63.4%	3	0.5%
$EQ\text{-}5D_{Self\text{-}care}$	430	71.4%	158	26.2%	14	2.3%
$EQ\text{-}5D_{Usual\ activities}$	157	26.3%	347	58.2%	92	15.4%
$EQ\text{-}5D_{Pain/discomfort}$	194	32.0%	358	59.0%	55	9.1%
$EQ\text{-}5D_{Anxiety/depression}$	323	53.5%	251	41.6%	30	5.0%

Table 9: Self-reported levels derived for SF-6D

Level	1		2		3		4		5		6	
Dimension	N	%	N	%	N	%	N	%	N	%	N	%
SF-6D$_{Physical}$	21	3.5	87	14.5	160	26.7	167	27.8	86	14.3	79	13.2
SF-6D$_{Role\ limit}$	129	23.8	155	28.5	47	8.7	212	39.0				
SF-6D$_{Social}$	136	22.4	85	14.0	212	34.9	109	18.0	65	10.7		
SF-6D$_{Bodily\ pain}$	129	21.6	64	10.7	116	19.5	120	20.1	124	20.8	43	7.2
SF-6D$_{Mental\ health}$	127	21.2	149	24.8	199	33.2	84	14.0	41	6.8		
SF-6D$_{Vitality}$	17	2.9	80	13.8	158	27.2	131	22.6	194	33.4		

EQ-5D responses are relatively easy to summarise. The majority of patients used the middle category (i.e. some problems) for mobility, usual activities and pain, and had no problems with self-care and were not anxious / depressed. However, SF-6D responses are less easy to summarise briefly due to the increased number of response categories and less clear cut response for the majority. This is unsurprising given the availability of more health states. The use of a greater number of health states in SF-6D could be interpreted as demonstrating more sensitivity than EQ-5D.

As mentioned at the end of Section 4.3, the correlation was good between EQ-5D$_{Utility}$ and SF-6D$_{Utility}$,.

5. Discussion

The aim of this paper was to use the Brazier and Deverill checklist[19] as a comprehensive framework to systematically evaluate the two preference-based measures of health benefit (EQ-5D$_{Utility}$ and SF-6D$_{Utility}$). The question addressed was which, if either, performed better in a group of renal patients on haemodialysis. We were able to present data which could be used for a limited assessment of the practicality and validity of EQ-5D and SF-6D. The main aspect we did not asses was reliability. For those aspects we were able to measure, the pros and cons of each instrument can be summarised as follows.

In terms of practicality, the EQ-5D appeared quicker to complete and had a 14% higher completion rate than SF-6D. Health economists might expect that a pre-scored multi-attribute health classification system would follow clearly defined decision rules in handling missing data. This is the case for EQ-5D - all questions must be completed to obtain a complete health state and hence utility. The situation is much more complicated for SF-6D. As shown in Appendix 2, incomplete answers to the 11 SF-36 questions results in three possible combinations of completeness of health state and utility data, rather than the one for EQ-5D.

Overall non-response patterns for EQ-5D$_{Utility}$ and SF-6D$_{Utility}$ were similar and significantly associated with greater age and co-morbidity. This is obviously of some concern if results were to be used to generalise to the general haemodialysis population. Another issue regarding practicality is the public sector employees wishing to use the SF-6D will now be charged (for SF-36 questions), whilst the EQ-5D is free.

Content and face validity appeared to be acceptable for EQ-5D and were unlikely to fair any worse for SF-6D. Whilst neither instruments' items overlapped with the condition specific dimensions of KDQoL^{TM}, it was unclear whether these dimensions were important ones in the rating of patients' utilities. This should be the subject of further research.

In terms of convergent validity, the associations were in the direction expected and most were statistically significant, although it is accepted that due to the multiple statistical tests, some may have achieved significance purely by chance.

Our tests of empirical validity found that apart from URR, EQ-5D$_{Utility}$ and SF-6D$_{Utility}$, results followed the same patterns of association although there was some variation in those that achieved statistical significance. Interestingly the variables that showed an association with the utilities were those reflecting more fixed attributes or social factors (i.e. Karnofsky performance score, disability, co-morbidity and car ownership), rather than treatment-related aspects (URR and haemoglobin). This could be interpreted two ways. It could be that neither measure is good at detecting treatment-related effects, or alternatively, that the treatment-related outcomes do not reflect issues valued highly (i.e.

preferences). However, the latter interpretation requires caution as the utility weights are those derived from the general population, not patients per se.

Our comparison of the performance of EQ-5D and SF-6D found that the mean utility values were very similar in this patient group, although there was slightly greater variation in their medians. However, the distribution was narrower for SF-$6D_{Utility}$, than EQ-$5D_{Utility}$. This could be interpreted as SF-$6D_{Utility}$ being less sensitive than EQ-$5D_{Utility}$, but may simply reflect a scaling issue in that EQ-$5D_{Utility}$ can range from -0.594 to 1, whereas SF-$6D_{Utility}$, can only range from 0.296 to 1.

We also assessed the instruments' sensitivity in relation to the number of health states used. Here the results were as expected with only 68 EQ-5D health states used in comparison to the 430 for SF-6D. Given the large number of SF-6D health states used, direct comparison with EQ-5D health states would be difficult without using somewhat arbitrary decision rules and so in this paper we only examined the extreme health states. This revealed some interesting differences. Only 3 patients used SF-6D's best health state whilst 67 patients had used EQ-5D's best health state. At the opposite end of the spectrum, the patients who used the worst health in SF-6D did not use the worst health state in EQ-5D. It is clear from looking at the items included, the two instruments examine slightly different aspects of HRQoL. Thus, generalising as to which to use in a health care setting may not be helpful since aspects may vary in importance for certain patient groups.

There was a good correlation between EQ-$5D_{Utility}$ and SF-$6D_{Utility}$,, although as the instruments are derived by different valuations methods (i.e. time trade-off and standard gamble for EQ-$5D_{Utility}$ SF-$6D_{Utility}$ respectively) one would not expect a perfect correlation.

Within the current (main) study we were interested in comparisons across settings. However, for a more comprehensive view we would need to look at follow-up care and whether either instrument detected treatment effects.

As this is work in progress, we intend to refine our analyses further, but in the meantime would appreciate some feedback particular with respect to the following. Firstly, does the SF-$6D_{Utility}$ relate to one month or one day, since the questions are predominately asked in relation to the past 4 weeks (8 of the 11)? If it is 4 weeks it means patients could provide a better reflection of their utilities in longitudinal studies. Secondly, is it desirable that someone who does not answer all the 11 questions may still obtain a SF-$6D_{Utility}$, whereas this is not possible for EQ-$5DD_{Utility}$ and are utilities derived from incomplete data reliable? Thirdly, how should the scaling issues about SF-$6D_{Utility}$ be addressed given that the lowest possible utility score is 0.296 and not zero (i.e. equivalent to death) or negative (worse than being dead as in EQ-$5D_{Utility}$)?

The following are more general issues that we feel are important to reflect on. Would the instruments would work equally well at determining the HRQoL differences associated with treatments? Which statistical methods should be used to assess validity (association / correlation / ? other) since correlation was not deemed appropriate for the available data and the tests of association used do not give the strength of the association? As we stated at the outset, we recognise this is the beginning of a process to evaluate the two instruments and it is clearly too early to judge the full ramifications of SF-6D.

References

1. http://www.euroqol.org/index.htm . 2002. Accessed 3-12-2002.
2. Brooks R, The EuroQol Group. EuroQol: the current state of play. *Health Policy* 1996;**37**:53-72.
3. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology* 1998;**51**:1115-28.
4. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* 2002;**21**:271-92.
5. Ware JE Jr., Snow KK, Kosinski M, Gandek B. SF-36 Health Survey. Manual and interpretation guide. Boston, Massachusetts: The Health Institute, New England Medical Center, 1993.
6. Wight JP, Edwards L, Brazier J, Walters S, Payne JN, Brown CB. The SF36 as an outcome measure of services for end stage renal failure. *Quality in Health Care* 1998;**7**:209-21.
7. Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A *et al*. Measurement of quality of life in end-stage renal disease: the time trade -off approach. *Clin Invest Med* 1987;**10**:14-20.
8. De Wit GA, Ramsteijn PG, De Charro FT. Economic evaluation of end stage renal disease treatment. *Health Policy* 1998;**44**:215-32.
9. Hays RD, Kallich JD, Mapes DL, Coons SJ, Carter WB. Development of the kidney disease quality of life (KDQOL) instrument. *Qual Life Res* 1994;**3**:329-38.
10. Department of Health. Renal Purchasing Guidelines. 1996. London, NHS Executive.
11. Feest TG, Rajamahesh J, Taylor H, Roderick P. The 1999 UK renal survey - adult patient numbers, renal unit facilities and processes of care. In Ansell D, Feest TG, eds. *UK Renal Registry Report 2000*, Bristol: UK Renal Registry, 2000.
12. Ansell D and Feest T. The Third Annual Report. The UK Renal Registry. 2000. Bristol, UK Renal Registry.
13. Renal Association and Royal College of Physicians. Treatment of adult patients with renal failure. Recommended standards and audit measures. 2nd Edition. 1997. London, Royal College of Physicians.
14. Roderick P, Nicholson T, Mehta R, Gerard K, Mullee M, Drey N *et al*. An evaluation of the costs and effectiveness of and quality of care of renal replacement therapy provision in renal satellite units in England and Wales. *Health Technol.Assess.* (forthcoming)
15. Ware JE Jr, Kosinski M, and Keller S.D. SF-36 physical and mental component summary measures: a users manual. 1994. New England Medical Center, Boston., The Health Institute. Tertiary SF-36 physical and mental component summary measures: a users manual.
16. Khan IH, Catto GR, Edward N, Fleming LW, Henderson IS, MacLeod AM. Influence of coexisting disease on survival on renal-replacement therapy. *Lancet* 1993;**341**:415-8.
17. Karnofsky D, Burcherval J. The Clinical Evaluation of Chemotherapeutic Agents in Cancer. In Macleod C, ed. New York: Columbia University Press, 1949.
18. Kind P, Hardman G, and Macran, S. UK population norms for EQ-5D. Discussion paper 172. 1999. York, Centre for Health Economics, University of York.
19. Brazier J, Deverill MA. Checklist for Judging Preference-Based Measures of Health Related Quality of Life: Learning from Psychometrics. *Health Economics* 1999;**8**:41-51.
20. Altman DG. *Practical Statistics for Medical Research*, p 404. London: Chapman & Hall, 1991.

Appendix 1: EQ-5D and SF-6D classification systems

EQ-5D Classification

1 Mobility
1 I have no problems in walking about
2 I have some problems in walking about
3 I am confined to bed

2 Self-Care
1 I have no problems with self-care
2 I have some problems washing or dressing myself
3 I am unable to wash or dress myself

3 Usual Activities (e.g. work, study, housework, family or leisure activities)
1 I have no problems with performing my usual activities
2 I have some problems with performing my usual activities
3 I am unable to perform my usual activities

4 Pain/Discomfort
1 I have no pain or discomfort
2 I have moderate pain or discomfort
3 I have extreme pain or discomfort

5 Anxiety/Depression
1 I am not anxious or depressed
2 I am moderately anxious or depressed
3 I am extremely anxious or depressed

SF-6D Classification

1 Physical functioning
1 Your health does not limit you in vigorous activities
2 Your health limits you a little in vigorous activities
3 Your health limits you a little in moderate activities
4 Your health limits you a lot in moderate activities
5 Your health limits you a little in bathing and dressing
6 Your health limits you a lot in bathing and dressing

2 Role limitation
1 You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems
2 You are limited in the kind of work or other activities as a result of your physical health
3 You accomplish less than you would like as a result of emotional problems
4 You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problem

3 Social functioning
1 Your health limits your social activities none of the time
2 Your health limits your social activities a little of the time
3 Your health limits your social activities some of the time
4 Your health limits your social activities most of the time
5 Your health limits your social activities all of the time

4 Pain
1 You have no pain
2 You have pain but it does not interfere with your normal work (both outside the home and housework)
3 You have pain that interferes with your normal work (both outside the home and housework) a little bit
4 You have pain that interferes with your normal work (both outside the home and housework) moderately
5 You have pain that interferes with your normal work (both outside the home and housework) quite a bit
6 You have pain that interferes with your normal work (both outside the home and housework) extremely

5 Mental health
1 You feel tense or downhearted and low none of the time
2 You feel tense or downhearted and low a little of the time
3 You feel tense or downhearted and low some of the time
4 You feel tense or downhearted and low most of the time
5 You feel tense or downhearted and low all of the time

6 Vitality

1 You have a lot of energy all of the time
2 You have a lot of energy most of the time
3 You have a lot of energy some of the time
4 You have a lot of energy a little of the time
5 You have a lot of energy none of the time

Questions from SF-36 used to derive the SF-6D

1. Physical functioning (during a typical day):

		Yes, limited a lot	Yes, limited a little	No, not limited at all
SF3	*Vigorous activities*, such as running, lifting heavy objects, participating in strenuous sports	☐	☐	☐
SF4	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐	☐	☐
SF12	Bathing or dressing yourself	☐	☐	☐

2. Role limitation

During the **past 4 weeks**, how much of the time have you had any of the following problems with your work or other regular activities **as a result of your physical health**? ✓a box

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF15	Were limited in the kind of work or other activities	☐	☐	☐	☐	☐

During the **past 4 weeks**, how much of the time have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)? ✓a box

		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF18	Accomplished less than you would like	☐	☐	☐	☐	☐

3. Social functioning

SF32 During the **past 4 weeks**, how much of the time has your **physical health or emotional problems** interfered with your **social activities** (like visiting friends, relatives, etc)? ✓a box

All of the time	Most of the time	Some of the time	A little of the time	None of the time
☐	☐	☐	☐	☐

4. Pain

SF21 How much bodily pain have you had during the **past 4 weeks**? ✓a box

None	Very mild	Mild	Moderate	Severe	Very severe
☐	☐	☐	☐	☐	☐

SF22 During the **past 4 weeks**, how much did *pain* interfere with your normal work (including both work outside the home and housework)? ✓a box

Not at all	A little bit	Moderately	Quite a bit	Extremely
☐	☐	☐	☐	☐

5. Mental health

How much of the time during the **past 4 weeks**… ✓a box on each line

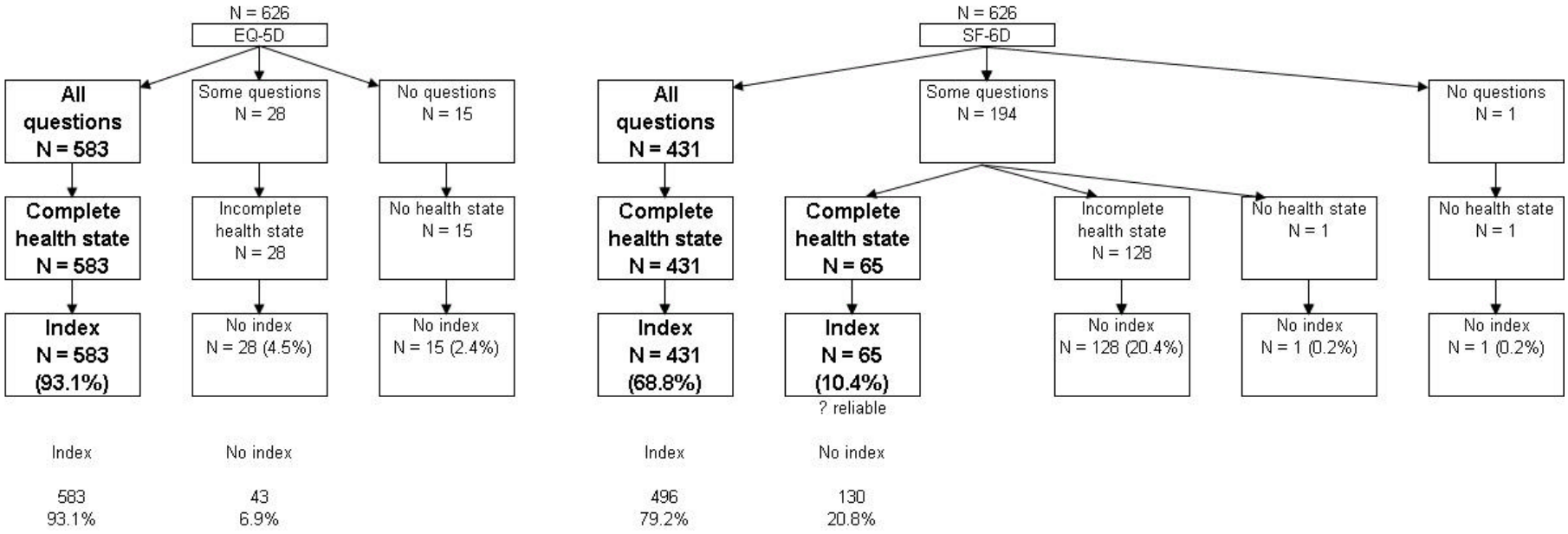
		All of the time	Most of the time	Some of the time	A little of the time	None of the time
SF24	Have you been very nervous?	☐	☐	☐	☐	☐
SF28	Have you felt downhearted and depressed?	☐	☐	☐	☐	☐

6. Vitality

How much of the time during the **past 4 weeks**… ✓a box on each line

SF27	Did you have a lot of energy?	☐	☐	☐	☐	☐

Appendix 2: EQ-5D and SF-6D completion rates



Appendix 3: Distributions of EQ-5D$_{Utility}$, SF-6D$_{Utility}$ and EQ-5D$_{VAS}$

Figure A1: Frequency distribution of EQ-5D$_{Utility}$

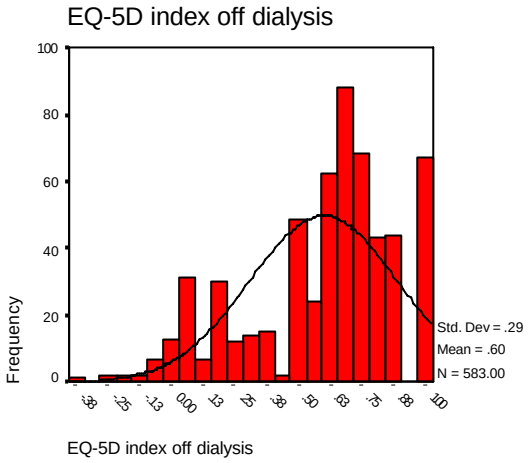


Figure A2: Frequency distribution of SF-6D$_{Utility}$

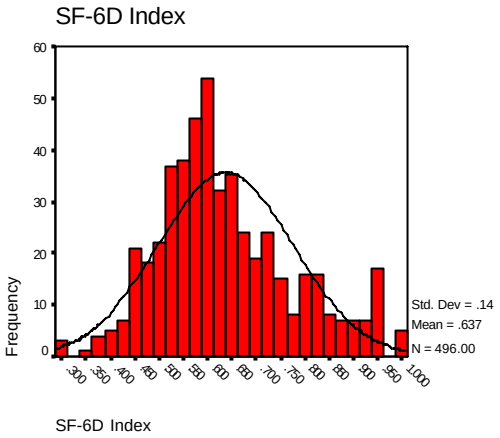
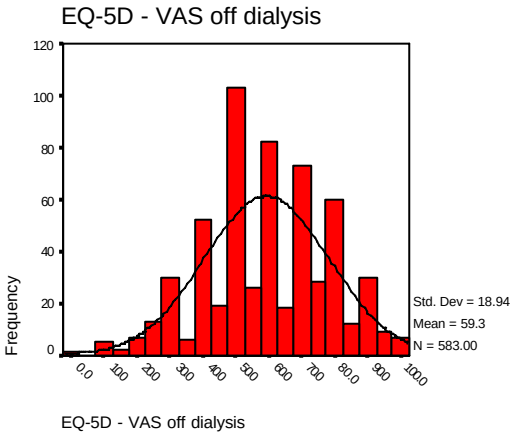


Figure A3: Frequency distribution of EQ-5D$_{VAS}$



For information

Figure A4: Graph of EQ-5D$_{VAS}$ on and off dialysis