BJPsych The British Journal of Psychiatry (2014)
205, 236–243. doi: 10.1192/bjp.bp.112.122283

Using generic preference-based measures in mental health: psychometric validity of the EQ-5D and SF-6D

Brendan Mulhern, Clara Mukuria, Michael Barkham, Martin Knapp, Sarah Byford, Djøra Soeteman and John Brazier

Background

Generic preference-based measures (EuroQoL-5D (EQ-5D) and SF-6D) are used in the economic evaluation of mental health interventions. However, there are inconsistent findings regarding their psychometric properties.

Aims

To investigate the psychometric properties of the EQ-5D and SF-6D in different mental health conditions, using seven existing data-sets.

Method

The construct validity and responsiveness of the measures were assessed in comparison with condition-specific indicators.

Results

Evidence for construct validity and responsiveness in common mental health and personality disorders was found (correlations 0.22–0.64; effect sizes 0.37–1.24; standardised response means 0.45–1.31). There was some evidence for validity in schizophrenia (correlations 0.05–0.43), but responsiveness was unclear.

Conclusions

EQ-5D and SF-6D can be used in the economic evaluation of interventions for common mental health problems with some confidence. In schizophrenia, a preference-based measure focused on the impact of mental health should be considered.

Declaration of interest

J.B. developed the SF-6D. M.B. was part of the CORE-OM development group.

Cost utility analysis is used to assess the cost-effectiveness of interventions across mental health conditions, and is employed by reimbursement agencies such as the National Institute for Health and Care Excellence (NICE) to inform the allocation of healthcare resources.[1] Cost utility analysis uses quality-adjusted life years (QALYs) as the outcome measure. The QALY combines values for the quantity and health-related quality of life (HRQoL) into a single score. This allows for comparisons across treatments and disorders in terms of the cost per QALY gained from an intervention.

To derive a value for HRQoL, or utility, generic preference-based patient-reported outcome measures (PROMs) of health such as the EuroQol-5D (EQ-5D)[2,3] or the SF-6D[4,5] can be used. These measures include a health state descriptive system and a utility scale that is derived from the preferences of the general population for health states described by the measure. The scale is anchored on the 1–0 full health–dead scale (where negative states are valued as worse than dead). Generic preference-based PROMs can be used in clinical trials alongside condition-specific PROMs to assess both the comparative and cost-effectiveness of interventions.

As the use of cost utility analysis has increased, there has been interest in establishing the psychometric validity of preference-based PROMs for use in different mental health conditions. It has been found that the EQ-5D and SF-6D demonstrate construct validity and responsiveness for depression, but the results for anxiety disorders are less convincing.[6–9] Research in schizophrenia[10] and psychosis[11] populations found mixed evidence for validity. For personality disorders research indicates that the EQ-5D may be related to condition-specific indicators and be sensitive to changes in HRQoL.[12,13]

Research in this area is important as the level of validity of the measures affects the sensitivity of HRQoL measurement and the subsequent QALY values produced. This has an impact on the use of the measures in clinical practice and research, and may also influence the decision-making process in favour of the conditions where preference-based PROMs are valid. If there is evidence that the preference-based PROMs are not valid, the limitations of the measures should be taken into account in the decision-making process and in clinical contexts. Furthermore, alternative methods of measuring HRQoL can be considered.

The inconsistent findings suggest that further work to establish the validity of the measures is needed. The aim of this study is to investigate the psychometric performance of the EQ-5D and SF-6D across different mental health conditions (defined as common mental health problems, mixed common mental and personality disorders, schizophrenia, and personality disorders). Seven large data-sets were used to assess construct validity and responsiveness to change over time in comparison with validated condition-specific PROMS. The current study complements prior work[9,10] by pooling data from multiple sources and combining the evidence in an overview of the psychometric strengths and weaknesses of the measures. Three hypotheses were developed based on a series of systematic reviews examining the performance of EQ-5D and SF-6D in mental health.[14] We hypothesised that EQ-5D and SF-6D would demonstrate construct validity and responsiveness in common mental health problems and mixed diagnoses (hypothesis 1), and personality disorders (hypothesis 2). This is because the descriptive systems directly assess common mental health concepts and will therefore display a level of sensitivity and relationship with the condition-specific indicators. We also hypothesised that EQ-5D and SF-6D would demonstrate a low level of construct validity and responsiveness to schizophrenia symptoms due to limited sensitivity of the preference-based PROM descriptive systems (hypothesis 3).

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

Method

Identification of data-sets

Literature searches were conducted to identify studies that had used the EQ-5D and/or the SF-6D alongside a condition-specific measure in evaluating treatment efficacy in anxiety, depression, schizophrenia and personality disorders.[9,10,14] In total, 69 authors of relevant studies were contacted. Twelve data-sets (17% of those requested) were received and reviewed for acceptable condition-specific comparison measures or clinical indicators and a relevant condition. Seven data-sets met the criteria. Five were excluded: three as they focused on general population samples, and two as they did not include a relevant comparison measure. The seven data-sets are described in online Table DS1 and comprised: (1) cost-effectiveness of antidepressant medication (Assessing Health Economics of Antidepressants, AHEAD[15]); (2) psychological interventions for postnatal depression (PoNDER[16]); (3) Improving Access to Psychological Therapies cohort study (SDO-IAPT[17]); (4) cognitive–behavioural therapy for recurrent self-harm (Prevention of Parasuicide with Manual Assisted Cognitive behaviour Therapy, POPMACT[18,19]); (5) the Study on Cost-Effectiveness of Personality Disorder Treatment (SCEPTRE[20]); (6) the Quality of Life following Adherence Therapy for People Disabled by Schizophrenia and their Carers (QUATRO[21]); and (7) art therapy for schizophrenia (Multicenter study of Art Therapy in Schizophrenia: Systematic Evaluation, MATISSE[21]).

The first three studies comprised samples with common mental health problems ($n = 3512$), the fourth study included mixed common mental and personality disorder diagnoses leading to self-harm ($n = 480$), the fifth study included a personality disorder sample ($n = 932$), and the sixth and seventh studies included people presenting with schizophrenia ($n = 826$).

Measures

The generic preference-based PROMs were compared with a condition-specific measure in each data-set. The measure pairs are detailed in Table DS1.

Generic preference-based measures

EQ-5D. The EQ-5D[2,3] is a widely used, generic preference-based PROM that measures health status on five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) with three associated response options (no problem, some problems, extreme problems). This produces 243 possible health states. The utility score was derived from preferences for 45 states and ranges from –0.594 to 1. The EQ-5D is the preferred instrument for use in submissions to the NICE appraisal process.[1]

SF-6D. The SF-6D[4,5] is a generic preference-based PROM with six dimensions (physical functioning, role limitations, social functioning, pain, mental health and vitality), with between four and six response options that generate 18 000 health states. The health state classification system was developed from the SF-36/SF-12. The utility scale for the SF-6D was derived from preferences for 249 states and ranges from 0.296 to 1. It is accepted by a number of reimbursement agencies around the world including the Canadian Agency for Drugs and Technologies in Health[23] and the Australian Pharmaceutical Benefits Advisory Committee.[24]

Condition-specific measures

Hospital Anxiety and Depression Scale (HADS). The HADS[25] is a 14-item self-report measure that contains two 7-item subscales: depression (HADS-D) and anxiety (HADS-A). The total score for each dimension is 21 (items are scored 0–3), with high scores indicative of increased levels of anxiety and depression. A score of 8+ indicates a possible case, and a score of 11+ a probable case. The overall score (HADS-T) is also used as a measure of global functioning. The HADS has been widely used across clinical groups and research settings, and there is evidence for its psychometric validity.[26] EQ-5D was assessed alongside HADS in mild and moderate anxiety and depression samples from the AHEAD and POPMACT studies.

Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM). The CORE-OM[27–29] is a self-report measure developed in the UK for routine use in psychological services, and psychometric validity has been demonstrated.[30,31] CORE-OM comprises 34 items addressing domains of subjective well-being, symptoms (anxiety, depression, physical problems, trauma), functioning (general functioning, close relationships, social relationships) and risk (risk to self, risk to others). Items are scored on a 5-point, 0–4 scale. CORE clinical scores are computed as the mean of all completed items multiplied by 10 (range 0–40). SF-6D was assessed alongside the CORE-OM in mild and moderate anxiety and depression samples from the PoNDER and SDO-IAPT studies.

Structured Interview for DSM-IV Personality (SIDP-IV). Personality disorder diagnoses were assessed using the SIDP-IV.[32] This instrument includes the 11 formal DSM-IV-TR Axis II diagnoses (e.g. schizoid personality disorder) including personality disorder mixed, the two DSM-IV-TR appendix diagnoses (depressive and negativistic personality disorder) and, in addition, DSM-III-R self-defeating personality disorder. Items are scored on a 4-point, 0–3 scale, with scores of 2 and 3 indicating the presence of personality disorder traits. In this study, EQ-5D was assessed alongside the SIDP-IV in the personality disorders sample from the SCEPTRE study.

Brief Psychiatric Rating Scale Expanded (BPRS-E). The BPRS[33] was developed to assess symptom change in psychiatric in-patients and is one of the most widely used measures of psychotic and affective symptoms. The expanded version, BPRS-E, has 24 items developed for use in patients with schizophrenia and was used in the current study. The BPRS-E is administered using semi-structured interviews, with items scored from 1 (not present) to 7 (extremely severe). EQ-5D and SF-6D were assessed alongside the BPRS-E in the schizophrenia sample from the QUATRO study.

Positive and Negative Syndrome Scale (PANSS). The interviewer-administered PANSS[34] was developed to evaluate positive, negative and other symptom dimensions in schizophrenia by combining the 18 items of the BPRS with the 12 items of the Psychopathology Rating Schedule. The 30 items are scored from 1 (absent) to 7 (extreme) and result in 3 subscales: positive, negative, and general psychopathology. EQ-5D was assessed alongside PANSS in the schizophrenia sample from the MATISSE study.

Analysis

Construct validity

Construct validity assesses the extent to which a measure reflects differences in HRQoL hypothesised to exist in a population and is important in relation to preference-based PROMs as generic utility values used in economic evaluation should reflect HRQoL factors linked to the condition or treatment being evaluated.

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

Construct validity is assessed in light of the fact that there is no gold standard for the measurement of HRQoL in mental health. This is linked to the heterogeneity of mental health conditions, and the difficulty in generating an indicator that assesses the full impact of the condition on people's lives. Therefore, we can assess a range of indicators of construct validity but cannot fully prove the validity of an instrument.

To assess construct validity we examined the two related empirical tests of convergent validity and known group differences.

Convergent validity

The convergence between the generic preference-based PROMs and the condition-specific instruments was tested using Pearson's correlation coefficients and locally weighted scatterplot smoothing (LOWESS)[35] techniques. Strong correlations indicate that the preference-based PROMs can measure mental health-related factors that are also assessed by the validated condition-specific instruments. Correlations are considered weak if scores are <0.3, moderate if scores are $\geqslant 0.3$ and <0.7, and strong if scores are $\geqslant 0.7$.

LOWESS is a form of non-parametric regression that attempts to capture general patterns in the relationship between two measures without making assumptions about the actual relationship between the variables. LOWESS plots a line on a scatterplot on the central tendency between the two variables, thereby visualising the relationship between these variables across the full scoring range.

In the common mental health problems and mixed diagnosis groups, the convergent validity of the EQ-5D was assessed in comparison to the HADS-T, HADS-A and HADS-D. The SF-6D was assessed in comparison to the CORE-OM clinical and dimension scores.

For personality disorders, tests of convergence between the EQ-5D and SIDP-IV were not carried out, as the SIDP-IV assesses 14 personality disorders individually on a 4-point scale, and therefore correlating each disorder indicator with the EQ-5D index score was inappropriate. For schizophrenia, EQ-5D was assessed in comparison to the PANSS and the BPRS-E, and the SF-6D was assessed in comparison to the BPRS-E.

Known group validity

Known group validity was assessed by testing whether the generic preference-based PROMs discriminated between condition-specific severity groups. For the common mental health problems and mixed diagnosis samples, the known group validity of the EQ-5D was assessed using HADS-A and HADS-D cut-off points indicating probable anxiety or depression (scores $\geqslant 11$). For the SF-6D, known group validity was assessed using CORE-OM clinical cut-off points (where a score >10 indicates clinical concerns).

For the personality disorders sample, EQ-5D known group validity was tested using diagnosis categories. These were defined as those with and without a personality disorder diagnosis, and also the number of personality disorders diagnosed.

For schizophrenia, the validity of the EQ-5D and SF-6D in the QUATRO sample used BPRS-E cut-offs (31 for 'mildly ill', 41 for 'moderately ill', 53 for 'markedly ill' and 70 for 'extremely ill').[36] For the MATISSE sample, PANSS cut-offs (58 for 'mildly ill', 75 for 'moderately ill', 95 for 'markedly ill' and 116 for 'severely ill')[37] were used.

One-way ANOVA was used to assess the magnitude of differences in the preference-based PROM scores across the severity groups. Standardised effect sizes across severity subgroups were assessed (calculated as the difference in mean scores between two adjacent severity subgroups divided by the standard deviation of scores for the milder of the two subgroups). Effect sizes of <0.2 are considered small, 0.5 moderate, and 0.8 large.[38]

Responsiveness

To test responsiveness we assessed the sensitivity of the EQ-5D and SF-6D to change in mental health in comparison with the condition-specific PROMs. Responsiveness is important in economic evaluation as any change in health must be reflected by change in utility (or preferences), and subsequent change in QALYs. For example, if HRQoL changes following an intervention, but the generic measure does not pick up this change, then this will not be reflected in QALYs gained despite improvements in HRQoL. This could wrongly influence funding decisions.

To measure responsiveness we examined floor and ceiling effects. Floor (lowest possible score) and ceiling (highest possible score) effects affect the ability of the measure to detect deterioration or improvements in health respectively.

We also examined the magnitude of change in scores before and after an intervention. We accept that this is a crude indicator of change. However, for each study, we assessed whether evidence of health change between baseline and follow-up would be expected based on the intervention and the published results. Evidence of a change in mental health should be reflected by change in the preference-based PROM score. The magnitude of change was assessed using the standardised response mean (SRM) statistic (calculated by dividing the mean change on the measure by the standard deviation of the change). Standardised response means of <0.2 are considered small, 0.5 moderate, and 0.8 large.[38] Responsiveness analysis was not carried out for the mixed diagnosis (POPMACT) sample as only baseline data were available.

Results

Sample characteristics

Demographic characteristics available for each data-set are displayed in online Table DS1. The POPMACT sample has significantly lower EQ-5D and HADS scores than the AHEAD sample, indicating higher levels of quality of life impairment and anxiety and depression (online Table DS2). SF-6D and CORE-OM scores indicate that the SDO-IAPT sample displays lower levels of quality of life and functioning than the PoNDER sample. For the schizophrenia samples, baseline EQ-5D scores indicate that both the MATISSE and QUATRO samples have similar quality of life levels. Those in the personality disorder (SCEPTRE) sample display lower quality of life than the schizophrenia sample. Across the samples, completion rates are high (above 95%).

Convergent validity

Common mental health problems

The correlations between the EQ-5D and HADS indicate a moderate level of convergence (Table 1). The SF-6D was correlated with the CORE-OM clinical score and functioning, well-being, and symptoms domain scores in the moderate to strong range across both the SDO-IAPT and PoNDER samples. The correlation with the risk domain score was moderate for the SDO-IAPT sample and low for the PoNDER sample. All correlations were significant ($P<0.01$), and were negative as a high score on the generic preference-based PROM and a low score on the condition-specific measure indicate better health status. These results support hypothesis 1.

Online Fig. DS1 displays scatterplots of the relationship between the generic and condition-specific measures and the

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

Table 1 Convergent validity of the EQ-5D and SF-6D

	EQ-5D				SF-6D		
	AHEAD	POPMACT	QUATRO	MATISSE	SDO-IAPT	PoNDER	QUATRO
Common mental disorders							
HADS							
Total	−0.36*	−0.49*			–	–	
Anxiety	−0.35*	−0.39*			–	–	
Depression	−0.22*	−0.46*			–	–	
CORE-OM score							
Clinical	–	–			−0.61*	−0.51*	
Functioning	–	–			−0.51*	−0.46*	
Symptoms	–	–			−0.64*	−0.53*	
Well-being	–	–			−0.51*	−0.45*	
Risk	–	–			−0.37*	−0.16	
Schizophrenia							
BPRS-E							
Total			−0.42*	–			−0.29*
Disorganisation			−0.22*	–			−0.13*
Depression			−0.43*	–			−0.34*
Negative symptoms			−0.21*	–			−0.12*
Positive symptoms			−0.31*	–			−0.20*
PANSS							
Total			–	−0.16*			–
Positive symptoms			–	−0.12			–
Negative symptoms			–	−0.05			–
General symptoms			–	−0.21*			–

*Significant at 0.01.
EQ-5D, EuroQoL-5D; AHEAD, Assessing Health Economics of Antidepressants; POPMACT, Prevention of Parasuicide with Manual Assisted Cognitive behaviour Therapy; QUATRO, Quality of Life following Adherence Therapy for People Disabled by Schizophrenia and their Carers; MATISSE, Multicenter study of Art Therapy in Schizophrenia: Systematic Evaluation; SDO-IAPT, Improving Access to Psychological Therapies cohort study; PoNDER, psychological interventions for postnatal depression; HADS, Hospital Anxiety and Depression Scale; CORE-OM, Clinical Outcomes in Routine Evaluation – Outcome Measure; BPRS-E, Brief Psychiatric Rating Scale Expanded; PANSS, Positive and Negative Syndrome Scale.

LOWESS fit lines. The lines demonstrate that the relationship between the EQ-5D and HADS differed across the severity scale, where the concordance between the measures is better at the less severe end of the scale. The relationship between the SF-6D and CORE-OM was more consistent across the severity scale, and was similar for both the SDO-IAPT and PoNDER samples.

Common mental health problems and personality disorders

The correlations between the EQ-5D and HADS indicate a moderate level of convergence ($P<0.01$; Table 1) supporting hypothesis 1. Again, the LOWESS fit line for the POPMACT data indicates that the relationship between the EQ-5D and HADS differed across the severity scale, where the concordance between the measures was higher at the less severe end of the scale.

Schizophrenia

The correlations between EQ-5D and condition-specific measures varied across the two schizophrenia samples. Correlations with the BPRS-E in the QUATRO sample were moderate for the total score and the depression and positive symptom dimensions. However, they were weak for the other dimensions (Table 1). Correlations with the PANSS in the MATISSE sample were weak, indicating little convergence. The correlations between SF-6D and BPRS-E follow a similar pattern to those of the EQ-5D, although the correlations were smaller in magnitude, with weak correlations across most of the dimensions apart from depression (Table 1). This indicates little convergence and supports hypothesis 3.

The LOWESS lines for the QUATRO sample (those who completed both EQ-5D and SF-6D) demonstrate a tendency for the generic preference-based PROM scores to increase as scores on the BPRS-E decrease (equivalent to less severe problems on both measures, see online Fig. DS2). However, a score of 1 on EQ-5D was associated with a wide range of BPRS-E scores. There was a trend towards a linear relationship between the EQ-5D and PANSS.

Known group validity

Common mental health problems

EQ-5D index scores were significantly higher in the ‘no case’ group (a score of 0–10) than the ‘probable case’ group (a score of 11+) as measured by both the HADS-A and HADS-D ($P=0.002$). In both the SDO-IAPT and PoNDER samples, the SF-6D index score was significantly higher in the non-clinical population compared with the clinical group as measured by CORE-OM (both $P<0.001$; online Table DS3). These significant findings support hypothesis 1.

Common mental health problems and personality disorders

For the POPMACT sample, the EQ-5D index scores were significantly higher in the ‘no case’ group than the ‘probable case’ group for both the HADS-A ($P<0.001$) and HADS-D ($P<0.001$), supporting hypothesis 1.

Personality disorders

For the SCEPTRE data, EQ-5D scores varied according to the number of diagnoses, with lower scores for those with one or more personality disorders (Table DS3). However, these differences were not statistically significant ($P=0.202$). There was a significant difference in EQ-5D scores between samples with different types of personality disorder ($P=0.042$). There is limited support for hypothesis 2.

Schizophrenia

EQ-5D scores were significantly higher for those with a lower level of severity measured by both the BPRS-E ($P<0.001$) and the

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

PANSS ($P = 0.003$) in both schizophrenia samples (Table DS3). Effect sizes across the severity subgroups were moderate in size for the BPRS-E and small for the PANSS. This indicates that to some extent the EQ-5D can identify known severity groups.

SF-6D scores significantly discriminated between BPRS-E severity groups. Effect sizes indicate that the difference between the mild and moderate severity groups was small. These findings indicate that there is a level of construct validity for the generic preference-based PROMs, but this varies across samples. Hypothesis 3 is not confirmed.

Responsiveness

Common mental health problems

For the AHEAD sample at baseline, EQ-5D and HADS displayed no evidence of floor or ceiling effects. However, at follow-up there was evidence of a large ceiling effect for EQ-5D and a moderate ceiling effect for HADS-D (Table 2). The SRM for EQ-5D was moderate and for the HADS was large. This demonstrates that the HADS was more responsive in the AHEAD sample, where significant change for both measures based on the results of the study would be expected.[15]

The SF-6D displayed a small ceiling effect for the PoNDER data. The SRM was in the large range, in contrast to the CORE-OM domains, which were in the small range. Change based on the psychotherapy interventions tested may be expected, and therefore provides evidence for SF-6D responsiveness.[16]

The SRM statistics for the SF-6D and CORE-OM in the SDO-IAPT sample were in the moderate range (where change based on the psychological therapies delivered as part of IAPT may be expected).[17] Therefore, there was evidence that the responsiveness of SF-6D was in the same range as the CORE-OM for depression, and may be more responsive in the PoNDER postnatal depression sample. Overall, the results reported in this section provide some support for hypothesis 1.

Personality disorders

In the SCEPTRE sample, EQ-5D displays minimal floor and ceiling effects. Responsiveness is also good, which reflects the

Table 2 Responsiveness of generic and condition-specific measures[a]

	% at floor		% at ceiling				
Measure	T_0	T_1	T_0	T_1	Mean change (s.d.)	SRM[a]	*t*-test
Common mental disorders							
EQ-5D							
AHEAD ($n = 164$)							
EQ-5D	0	0	2.19	34.15	0.17 (0.38)	0.45	
HADS total	0	0	0	0	−10.74 (8.83)	−1.22	
HADS anxiety	0	0	0	0	−4.81 (4.98)	−0.97	
HADS depression	0	0	0.62	14.79	−5.93 (5.67)	−1.05	
SF-6D							
SDO-IAPT ($n = 390$)							
SF-6D	0	0	0	1.54	−0.06 (0.12)	0.50	
CORE-OM clinical score	0	0	0	0	−4.71 (6.71)	−0.70	
Functioning score	0.41	0	0.82	1.50	−0.37 (0.75)	−0.49	
Symptoms score	1.22	1.24	0.20	0.50	−0.58 (0.84)	−0.70	
Well-being score	7.46	2.72	0.81	3.95	−0.57 (0.97)	−0.59	
Risk score	0.2	0	39.27	54.48	−0.18 (0.55)	−0.32	
PoNDER ($n = 1697$)							
SF-6D	0	0	0	18.33	0.17 (0.13)	1.31	
CORE-OM clinical score	0	0	3.48	7.82	−0.58 (4.69)	−0.12	
Functioning score	0	0.06	12.35	17.24	−0.04 (0.57)	−0.07	
Symptoms score	0	0	8.60	16.13	−0.10 (0.57)	−0.18	
Well-being score	0	0.06	20.14	29.77	−0.10 (0.76)	−0.13	
Risk score	0.04	0	90.23	89.55	−0.01 (0.20)	−0.05	
Personality disorders							
SCEPTRE ($n = 679$), EQ-5D	0	0	4.0	21.6	0.170 (0.29)	0.58	0.000
Schizophrenia							
QUATRO ($n = 328$)							
EQ-5D	0	0	16.8	20.7	0.035 (0.29)	0.12	0.026
SF-6D	0	0.3	0.6	0.9	0.014 (0.12)	0.12	0.027
BPRS-E	1.2	4.3	0	0	−7.60 (13.06)	−0.58	0.000
BPRS-E positive	17.1	26.8	0	0	−3.04 (5.70)	−0.53	0.000
BPRS-E negative	21.3	35.1	0	0	−1.37 (4.06)	−0.34	0.000
BPRS-E disorganisation	20.1	36.9	0	0	−1.62 (4.22)	−0.38	0.000
BPRS-E depression	0	15.9	0	0	−1.90 (5.41)	−0.35	0.000
MATISSE ($n = 321$)							
EQ-5D	0	0	16.8	20.2	−0.005 (0.29)	−0.02	0.767
PANSS	0	0	0	0	−3.41 (20.85)	−0.14	0.004
PANSS positive	2.5	3.4	0	0	−0.93 (6.17)	−0.15	0.007
PANSS negative	2.2	4.0	0	0.3	−0.78 (6.48)	−0.11	0.031
PANSS general symptoms	0.3	0	0	0	−1.21 (10.65)	−0.10	0.042

T_0, baseline; T_1, follow-up; EQ-5D, EuroQoL-5D; AHEAD, Assessing Health Economics of Antidepressants; HADS, Hospital Anxiety and Depression Scale; SDO-IAPT, Improving Access to Psychological Therapies cohort study; CORE-OM, Clinical Outcomes in Routine Evaluation – Outcome Measure; PoNDER, psychological interventions for postnatal depression; SCEPTRE, Study on Cost-Effectiveness of Personality Disorder Treatment; QUATRO, Quality of Life following Adherence Therapy for People Disabled by Schizophrenia and their Carers; BPRS-E, Brief Psychiatric Rating Scale Expanded; MATISSE, Multicenter study of Art Therapy in Schizophrenia: Systematic Evaluation; PANSS, Positive and Negative Syndrome Scale.
n = those who completed both measures at all time points.
a. Standardised response mean (SRM) size: small, >0.2 ⩽0.5; moderate, >0.5 <0.8; large, ⩾0.8.

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

significant change expected following the psychotherapy administered,[20] with moderate SRMs at 12 months. This finding suggests that EQ-5D can respond to change over time, providing support for hypothesis 2.

Schizophrenia

For the QUATRO sample, EQ-5D and SF-6D display no evidence of a floor effect, but EQ-5D has a large ceiling effect at both time points (Table 2). Although adherence therapy was not found to improve quality of life relative to health education,[21] mean change for EQ-5D and SF-6D is statistically significant. However, the SRMs are <0.2 (below the clinically significant range), and the BPRS-E has larger SRMs, indicating that the preference-based PROMs were less responsive. This provides some support for hypothesis 3.

In the MATISSE sample, the EQ-5D has no floor effect but a large ceiling effect. Mean change for EQ-5D is not statistically significant and has a small SRM. The PANSS demonstrates statistically significant mean change, but the SRMs are still in the low range. This indicates that neither the EQ-5D nor PANSS are responsive in the MATISSE schizophrenia sample. The results of the trial (which found that the intervention (art therapy) did not improve outcomes in comparison to the control group),[22] suggest that neither the PANSS nor the EQ-5D would be expected to demonstrate responsiveness, and therefore this result should be interpreted with caution.

Discussion

Seven data-sets were used to examine the psychometric validity of the EQ-5D and SF-6D across a range of mental health conditions. The results suggest that the generic preference-based PROMs are valid for use in common mental health problems, and there is some evidence of responsiveness to change over time. Our hypothesis that the measures will display construct validity and responsiveness in common mental health problems and mixed diagnosis samples (hypothesis 1) was supported. For personality disorders, the results were also positive, as EQ-5D was shown to discriminate between severity groups, and respond to change over time, supporting hypothesis 2. In comparison, the evidence in schizophrenia was less clear. There was some support for construct validity across related domains and some evidence of discriminative properties. However, responsiveness was low. Our hypotheses that the generic measures would not display a high level of validity or responsiveness in schizophrenia (hypothesis 3) was supported to some extent.

Evidence for the psychometric validity of the preference-based PROMs in common mental health problem patient samples is consistent with previous empirical work in mild depression and anxiety samples.[6,8,9] Both descriptive systems include questions that are relevant to depression and anxiety, and have a level of sensitivity to the condition. There were some differences between the performance of EQ-5D and SF-6D, but direct comparisons were difficult because the analysis of each measure was carried out using different samples. The growing evidence base regarding the validity of the instruments indicates that EQ-5D and SF-6D can be considered valuable for use in the economic evaluation of interventions for a range of common mental health problems.

The positive results found for the personality disorders sample are in line with past work which found that the EQ-5D correlates with condition-specific indicators, [12] and is responsive.[13] This indicates that the EQ-5D has a level of validity for use in the assessment of interventions for personality disorders. We also found that EQ-5D differs between different types of personality disorders, but this is difficult to interpret without further information about the characteristics of the conditions and the sample. We compared EQ-5D with a diagnosis instrument completed by clinicians, and it would be informative to use a self- or interviewer-administered condition-specific PROM as a comparator (in line with the other analysis carried out).

Past work has found mixed evidence for the performance of generic preference-based PROMs in schizophrenia.[10] We have established evidence for and against validity, with mixed evidence regarding the ability of the measures to reflect schizophrenia-specific symptoms. EQ-5D may be related to some condition-specific domains (e.g. depression) but not others (e.g. positive symptoms). This is linked to the classification system that may not be sensitive to schizophrenia-specific dimensions. Direct comparisons between the EQ-5D and SF-6D were possible for the QUATRO study, which found that neither instrument converges with the condition-specific measure. The intervention (adherence therapy) was also not shown to be better than health education, but the condition-specific measure was more responsive. In the MATISSE study, change may not be expected based on the intervention (art therapy), and neither the generic nor condition-specific indicators displayed responsiveness. The low level of responsiveness found for EQ-5D may be due to the large ceiling effect, which impairs its ability to detect change over time and reflects the lack of overlap between the descriptive system and schizophrenia-specific symptoms. However, SF-6D does not display the same ceiling effect characteristics. The mixed evidence regarding the schizophrenia sample means that the EQ-5D and SF-6D should be used with caution in this condition

Implications

The results of this study highlight a range of issues for clinicians and decision makers, who are involved in the use and interpretation of generic preference-based PROMs across mental health conditions. There are also issues raised for people with mental health problems who complete the measures as part of their ongoing care.

First, clinicians using PROMS should also be aware of the issues surrounding the sensitivity of both generic and condition-specific instruments. This is important if clinical decisions, the assessment of health change over time, and the assessment of performance are being linked to an individual's self-reported health status as measured by instruments where the level of validity is unclear.

Second, those interpreting the results who are responsible for the assessment and commissioning of treatments and interventions – and also researchers and guideline developers – should be aware of the limitations of the measures in certain conditions, and consider this in the decision-making process. Generic preference-based PROMs are used to assess effectiveness in economic evaluation, and so understanding their appropriateness in these conditions is important. The validity issues raised here are important for the comparability of interventions and the subsequent allocation of resources, and there is no consensus on the most valid outcome measure to use. The possible limitations of these generic health measures raises the question of whether those designing studies and subsequently assessing the cost-effectiveness of interventions should consider using measures developed for mental health populations. Generic measures such as the EQ-5D allow for comparability of interventions across conditions and so are useful in decision-making. However, they may favour interventions for physical conditions where the measures are found to be valid. Therefore, there would seem to

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

be a case for developing a preference-based measure for use in mental health populations that better reflects their concerns.[14,39]

A range of issues for patients completing the measures are also highlighted. The extent to which these instruments reflect the reality of a patient's condition (in terms of the dimensions included and the associated severity levels) needs to be assessed. If they do not, then patients may be reluctant to complete measures that lack face validity and/or the information gained from such measures may not accurately reflect their experience. Furthermore, clinicians and patients may differ in what they think the key areas of health to assess are, and therefore the tools used may not provide the most holistic assessment possible.

Psychometric analysis of preference-based PROMs is one method of assessing validity, and should be considered alongside other types of evidence to establish a more detailed picture. For example, this work should be considered alongside systematic reviews, which allow evidence of validity and responsiveness across a range of studies to be synthesised.[9,10] Qualitative work assessing the content validity and acceptability of the instruments from the patient perspective can be used to highlight domains that are missing from the preference-based PROMs.[14] This allows for insight into the performance of the instruments and will inform future work to increase the sensitivity and validity of measurement across a range of mental health conditions.

There are a number of ways in which the validity and sensitivity of preference-based PROMs for use in mental health could be improved. A five-level version of the EQ-5D (EQ-5D-5L) has also been developed,[40] and it is possible that this version may be more sensitive to different severity levels, and therefore change across time. Further research should assess the validity of EQ-5D-5L in patients with mental health conditions. Mental health-specific preference-based PROMs could also be developed either using standard instrument development procedures or by adapting an existing condition-specific instrument. This has been done for general mental health conditions using the CORE-OM.[41] Alternatively, 'bolt on' dimensions for the generic preference-based PROMs can be developed to directly assess particular conditions.

Limitations

This study has a number of limitations. First, as in much psychometric validation, there is no 'gold standard' measure of HRQoL against which to compare the generic preference-based PROMs. The lack of a gold standard is linked to both the heterogeneity of conditions and the multiplicity of perspectives on the impact of mental health conditions. For example, in relation to schizophrenia, although a clinician might put greater emphasis on positive and negative symptoms of psychosis, the person with the illness or their family members might be more concerned with social functioning or employment. The lack of a gold standard means that the comparisons between the measures are limited to the level of validity and relevance of the comparison indicator. This means that this study provides a guide to the performance of the measure, but can only be assessed in light of the overlap between the measures, which may be restricted due to the limited focus of the generic measures on mental health. Therefore, the results are open to interpretation and opinion. In this study it can be argued that the generic preference-based PROMs are compared against indicators that have some level of validity in the populations tested,[24,30,31] and this allows for inferences to be drawn. However, the different scope of the condition-specific and generic measures used here suggests that some level of divergence is to be expected. The same concerns apply when testing responsiveness and it is important to consider whether the level of change reported by the instrument is meaningful.

The inferences that can be drawn from the results are also limited to the mental health conditions and the samples included in the seven data-sets, and need to be interpreted with caution. The differing levels of performance reflect systematic variance attributable to the different types of data, patient populations, and study designs. For responsiveness analysis, it is important to note that significant change can only be inferred based on the intervention tested in the trial, which may not be found to significantly improve outcomes. Generalisability to other mental health samples with similar diagnoses is therefore unclear, and comparisons between the generic preference-based PROMs are difficult. Further work into the performance of the EQ-5D and SF-6D in mental health conditions, including direct comparisons, should focus on replicating the current analysis on different mental health conditions using a range of condition-specific indicators. This analysis could further inform decisions about which measure should be recommended for use in different conditions. It is also possible that the preference-based PROMs are picking up comorbidities but this was difficult to test in the data available as indicators of other conditions (including physical conditions) were not available. The impact of comorbidities on utility scores in mental health populations should be assessed in future work.

In summary, we have reported the first work to test the psychometric performance of two widely used, generic preference-based measures of HRQoL across a range of mental health problems using data from a variety of sources. The study adds to the evidence base about the mental health conditions where the measures can be used in the economic evaluation of new and emerging interventions. It also highlights possible areas where new preference-based measures, or additions to existing measures, would improve the measurement of HRQoL in mental health.

Acknowledgements

The authors would like to thank Professor Jan van Busschbach, Professor Simon Dixon, Professor Simon Gilbody, Derek King, Dr Louise Longworth, and Associate Professor Jane Morrell for their support in accessing the data-sets used in this study.

Brendan Mulhern, MRes, **Clara Mukuria**, PhD, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, UK; **Michael Barkham**, PhD, Centre for Psychological Services Research, Department of Psychology, University of Sheffield, UK; **Martin Knapp**, PhD, Centre for the Economics of Mental and Physical Health, King's College London, and Personal Social Services Research Unit, London School of Economics and Political Science, UK; **Sarah Byford**, PhD, Centre for the Economics of Mental and Physical Health, King's College London, UK; **Djøra Soeteman**, PhD, Center for Health Decision Science, Harvard School of Public Health, Boston, Massachusetts, USA; **John Brazier**, PhD, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, UK

Correspondence: Brendan Mulhern, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, Sheffield S1 4DA, UK. Email: b.mulhern@sheffield.ac.uk

First received 10 Oct 2012, final revision 6 Mar 2014, accepted 14 Mar 2014

References

1 National Institute for Health and Clinical Excellence. *Guide to the Methods of Technology Appraisal*. NICE, 2008.

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

3 Dolan P. Modeling valuations for EuroQol health states. *Med Care* 1997; **35**: 1095–108.

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

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press

5 Brazier JE, Roberts J. Estimating a preference-based index from the SF-12. *Med Care* 2004; **42**: 851–9.

6 Brazier J. Measuring and valuing mental health for economic evaluation. *J Health Serv Res Policy* 2008; **13** (suppl 3): 70–5.

7 Brazier J. Is the EQ-5D fit for purpose in mental health? *Br J Psychiatry* 2010; **197**: 348–9.

8 Lamers LM, Bouwmans CA, van Straten A, Donker MC, Hakkaart L. Comparison of EQ-5D and SF-6D utilities in mental health. *Health Econ* 2006; **15**: 1229–36.

9 Peasgood T, Brazier J, Papaioannou D. A systematic review of the validity and responsiveness of EQ-5D and SF-6D for depression and anxiety. Health Economics and Decision Science (HEDS) paper 12/15 (available at http://eprints.whiterose.ac.uk/id/eprint/74659).

10 Papaioannou D, Brazier J, Parry G. How valid and responsive are generic health status measures, such as EQ-5D and SF-6D, in schizophrenia? A systematic review. *Value Health* 2011; **14**: 907–20.

11 Barton GR, Hodgekins J, Mugford M, Jones PB, Croudace T, Fowler D. Measuring the benefits of treatment for psychosis: validity and responsiveness of the EQ-5D. *Br J Psychiatry* 2009; **195**: 170–7.

12 Soeteman DI, Timman R, Trijsburg RW, Verheul R, Busschbach JVV. Assessment of the burden of disease among inpatients in specialized units that provide psychotherapy. *Psychiatr Serv* 2005; **56**: 1153–5.

13 Soeteman DI, Verheul R, Delimon J, Meerman AMMA, van den Eijnden E, Rossum BV, et al. Cost-effectiveness of psychotherapy for cluster B personality disorders. *Br J Psychiatry* 2010; **196**: 396–403.

14 Brazier JE, Connell J, Papaioannou D, Mukuria C, Mulhern B, O'Cathain A, et al. Validating generic preference-based measures of health in mental health populations and estimating mapping functions for widely used specific measures. *Health Technol Assess* 2014; in press.

15 Kendrick T, Peveler R, Longworth L, Baldwin D, Moore M, Chatwin J, et al. Cost-effectiveness and cost-utility of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine: randomised controlled trial. *Br J Psychiatry* 2006; **188**: 337–45.

16 Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, et al. Clinical effectiveness of health visitor training in psychologically informed approaches for depression in postnatal women: pragmatic cluster randomised trial in primary care. *BMJ* 2009; **338**: a3045.

17 Mukuria C, Brazier J, Barkham M, Connell J, Hardy G, Hutten R, et al. Cost-effectiveness of an Improving Access to Psychological Therapies service. *Br J Psychiatry* 2013; **202**: 220–7.

18 Byford S, Knapp M, Greenshields J, Ukoumunne OC, Jones V, Thompson S, et al. Cost-effectiveness of brief cognitive behaviour therapy versus treatment as usual in recurrent deliberate self-harm: a rational decision making approach. *Psychol Med* 2003; **33**: 977–86.

19 Tyrer P, Tom B, Byford S, Schmidt U, Jones V, Davidson K, et al. Differential effects of manual assisted cognitive behavior therapy in the treatment of recurrent deliberate self-harm and personality disturbance: the POPMACT study. *J Pers Disord* 2004; **18**: 102–16.

20 Gray R, Leese M, Bindman J, Becker T, Burti L, David A, et al. Adherence therapy for people with schizophrenia: European multicentre randomised controlled trial. *Br J Psychiatry* 2006; **189**: 508–14.

21 Crawford MJ, Kilaspy H, Barnes TRE, Barrett B, Byford S, Clayton K. Group art therapy as an adjunctive treatment for people with schizophrenia: multicentre pragmatic randomised trial. *BMJ* 2012: **344**: e846.

22 Soeteman D, Verheul R, Busschbach J. The burden of disease in personality disorders: diagnosis-specific quality of life. *J Pers Disord* 2008; **22**: 259–68.

23 Canadian Agency for Drugs and Technologies in Health. *Guidelines for the Economic Evaluation of Health Technologies: Canada* (3rd edn). Canadian Agency for Drugs and Technologies in Health, 2006.

24 Pharmaceutical Benefits Advisory Committee. *Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee (PBAC)*. Commonwealth of Australia, 2008.

25 Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand* 1993; **67**: 361–70.

26 Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale: an updated literature review. *J Psychosom Res* 2002; **52**: 69–77.

27 Barkham M, Evans C, Margison F, McGrath G, Mellor-Clark J, Connell J, et al. The rationale for developing and implementing core batteries for routine use in service settings and psychotherapy outcome research. *J Ment Health* 1998; **7**: 35–47.

28 Barkham M, Gilbert N, Connell J, Marshall C, Twigg E. Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *Br J Psychiatry* 2005; **186**: 239–46.

29 Evans C, Mellor-Clark J, Margison F, Barkham M, Audin K, Connell J, et al. CORE: Clinical Outcomes in Routine Evaluation. *J Ment Health* 2000; **9**: 247–55.

30 Evans C, Connell J, Barkham M, Margison F, McGrath G, Mellor-Clark J, et al. Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM. *Br J Psychiatry* 2002; **180**: 51–60.

31 Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validations of PHQ-9 and CORE-OM. *Br J Gen Pract* 2007; **57**: 650–2.

32 Pfohl B, Blum N, Zimmerman M. *Structured Interview for DSM-IV Personality: SIDP-IV*. American Psychiatric Press, 1995.

33 Ventura J, Lukoff D, Nuechterlein K, Liberman R, Green M, Shaner A. Brief Psychiatric Rating Scale (BPRS) Expanded Version (4.0) scales, anchor points and administration manual. *Int J Methods Psychiatr Res* 1993; **3**: 227–43.

34 Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophr Bull* 1987; **13**: 261–76.

35 Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 1979; **74**: 829–36.

36 Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel R. Clinical implications of Brief Psychiatric Rating Scale scores. *Br J Psychiatry* 2005; **187**: 366–71.

37 Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel RR. What does the PANSS mean? *Schizophr Res* 2005; **79**: 231–8.

38 Cohen J. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Erlbaum, 1988.

39 Connell J, Brazier J, O'Cathain A, Lloyd-Jones M, Paisley S. Quality of life of people with mental health problems: a synthesis of qualitative research. *Health Qual Life Outcomes* 2012; **10**: 138.

40 Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011; **20**: 1727–36.

41 Mavranezouli I, Brazier J, Young T, Barkham M. Using Rasch analysis to form plausible health states amenable to valuation: the development of CORE-6D from a measure of common mental health problems. *Qual Life Res* 2011; **20**: 321–33.

EXTRA CONTENT ONLINE

https://doi.org/10.1192/bjp.bp.112.122283 Published online by Cambridge University Press