The University Of Sheffield.

School Of Health And Related Research.

Health Economics and Decision Science (HEDS)

Discussion Paper

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

Brendan Mulhern (MRes),1 Clara Mukuria (PhD),1 Prof. Michael Barkham (PhD),2 Prof. Martin Knapp (PhD),3, 4 Sarah Byford (PhD),3 Djøra Soeteman (PhD), 5 Prof. John Brazier (PhD)1

DP 13/04

HEDS Discussion Paper

No. 13.04

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

Brendan Mulhern (MRes),1 Clara Mukuria (PhD),1 Prof. Michael Barkham (PhD),2 Prof. Martin Knapp (PhD),3, 4 Sarah Byford (PhD),3 Djøra Soeteman (PhD), 5 Prof. John Brazier (PhD)1

1 Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, Sheffield, S1 4DA.
2 Centre for Psychological Services Research, Department of Psychology, University of Sheffield, S10 2TP.
3 Centre for the Economics of Mental and Physical Health, King's College London, Box P024, De Crespigny Park, London, SE5 8AF.
4 Personal Social Services Research Unit, London School of Economics and Political Science, Houghton Street, London WC2A 2AE. 5 Center for Health Decision Science, Harvard School of Public Health, Boston, Massachusetts 02115.

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

Brendan Mulhern (MRes),[1] Clara Mukuria (PhD),[1] Prof. Michael Barkham (PhD),[2] Prof. Martin Knapp (PhD),[3, 4] Sarah Byford (PhD),[3] Djøra Soeteman (PhD),[5] Prof. John Brazier (PhD)[1]

1 Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, Sheffield, S1 4DA.
2 Centre for Psychological Services Research, Department of Psychology, University of Sheffield, S10 2TP.
3 Centre for the Economics of Mental and Physical Health, King's College London, Box P024, De Crespigny Park, London, SE5 8AF.
4 Personal Social Services Research Unit, London School of Economics and Political Science, Houghton Street, London WC2A 2AE.
5 Center for Health Decision Science, Harvard School of Public Health, Boston, Massachusetts 02115.

Corresponding author:
Brendan Mulhern, Health Economics and Decision Science, School of Health and Related Research, University of Sheffield, Regent Court, Sheffield, S1 4DA.
Tel no: 0114 222 0794
e-mail: b.mulhern@sheffield.ac.uk

Full author details are reported at the end of the manuscript

Abstract

Background: Generic preference based measures (EQ-5D and SF-6D) can be used in the economic evaluation of mental health interventions. However there are inconsistent findings regarding the psychometric properties of the instruments.

Aims: To investigate the psychometric performance of the measures across a range of mental health conditions using seven existing datasets.

Methods: The feasibility, construct validity and responsiveness of the EQ-5D and SF-6D were assessed in comparison to condition specific indicators.

Results: Strong evidence for validity and responsiveness in common mental health and personality disorder samples was found. The psychometric performance in schizophrenia was more inconsistent.

Conclusions: EQ-5D and SF-6D can be used in the economic evaluation of interventions for common mental health problems and personality disorders with some confidence. In schizophrenia, the measurement of quality of life may be improved by developing a condition-specific preference based measure.

Declaration of interests: John Brazier developed SF-6D. Michael Barkham developed the CORE-OM.

Introduction

Cost utility analysis (CUA) can be used to assess the cost effectiveness of interventions across mental health conditions, and is employed by agencies such as the National Institute for Health and Clinical Excellence (NICE) to inform the allocation of scarce resources.[1] CUA uses the common metric of Quality Adjusted Life Years (QALYs) as the effect outcome measure - which combines values for the quantity and quality of life into a single score and allows comparisons across treatments for different conditions. To derive a value for health related quality of life (HRQL), or utility, generic preference-based measures of health (PBM) such as the EuroQol EQ-5D (EQ-5D)[2,3] or Short Form-6D (SF-6D)[4,5] can be used. The utility score is derived from the preferences of the general population and is anchored on the 0-1 dead-full health scale (where a score below zero is equivalent to a state worse than dead). Generic PBMs can be used in clinical trials alongside condition specific patient-reported outcome measures (PROMs) in order to assess both the comparative and cost effectiveness of interventions.

With the significant increase in the use of economic appraisal for funding and reimbursement decision making, there has been interest in establishing the psychometric validity of generic PBMs for use in mental health disorders. It has been found that both 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 populations[10] and individuals with psychosis[11] found mixed evidence on the validity of generic PBMs. For personality disorders research indicates that the EQ-5D may be related to condition specific indicators and be sensitive to changes in HRQL.[12]

The inconsistent findings regarding the psychometric properties of the EQ-5D and SF-6D in mental health disorders suggests that further work is needed to establish the validity of the measures. The aim of this study is to investigate the psychometric performance of both the EQ-5D and SF-6D across a range of mental health conditions including common mental health problems, schizophrenia, and personality disorders. Seven large datasets were used to assess the feasibility, validity, and responsiveness of the instruments to change over time in comparison to widely used and validated condition-specific PROMs. The current study complements prior work by pooling data from multiple sources and combining the evidence in an overview of the psychometric strengths and weaknesses of the EQ-5D and SF-6D for mental health disorders.

Methods

Identification of datasets

A literature search was conducted to identify studies that have used the EQ-5D and/or the SF-6D in measuring treatment efficacy in anxiety, depression, schizophrenia and personality disorders[9,10] as well as have included a condition specific measure. In total 69 authors of relevant studies were contacted with the request to use their datasets in the analysis. Of those, 12 datasets were received (17% of those requested), and these were reviewed for acceptable condition-specific comparison measures or clinical indicators. Datasets from seven studies were selected for use in these analyses including (1) assessing health economics of antidepressants (AHEAD), (2) psychological interventions for postnatal depression (PONDeR), (3) improving access to psychological therapies cohort study (IAPT), (4) a trial of cognitive behaviour therapy versus treatment as usual for recurrent self harm (POPMACT), (5) quality of life following adherence therapy (QUATRO), (6) multi-centre study of art therapy in schizophrenia – systematic evaluation (MATISSE), and (7) the study on the cost-effectiveness of personality disorder treatment (SCEPTRE). The first three studies included samples with common mental health problems (n=3,512), the fourth study included mixed common mental health and personality disorder diagnoses leading to self-harm (n=480), the next two studies (fifth and sixth) included schizophrenia (n=826) and the seventh study personality disorders (n=932)). Of the five datasets excluded, three were excluded as they focused on general population samples, and two were excluded as they did not include a comparison measure of interest. The seven datasets are described in **Table 1**.

Measures

The generic PBMs were compared to a condition specific measure in each dataset. The measure pairs used in the analyses are detailed in **Table 1**.

Generic preference based measures

EQ-5D

The EQ-5D[2,3] is a widely used generic PBM and 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). A selection of the 243 possible health states was valued by the general population using Time Trade Off (TTO) to produce a utility score for each health state (range -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 PBM that generates 18,000 health states across six dimensions (physical functioning, role limitations, social functioning, pain, mental health, and vitality), with between four and six response options. The questionnaire is a short version of the SF-36/SF-12. The utility scale for the SF-6D was derived by valuing 249 states using Standard Gamble (SG) 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,[21] and the Australian Pharmaceutical Benefits Advisory Committee.[22]

Condition specific measures

Hospital Anxiety and Depression Scale (HADS)

The HADS[23] is a 14 item self-report measure that contains two seven 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+ indicates 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.[24] In this study, HADS was used to assess the performance of the EQ-5D in two samples of people with mild and moderate anxiety and depression from the AHEAD and POPMACT trials.

Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM)

The CORE-OM[25-29] is a self-report measure developed in the UK for routine use in psychological services. 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). The psychometric validity of the CORE-OM has been demonstrated.[30,31] In this study the CORE-OM was used to assess the psychometric performance of the SF-6D in two samples of people with mild and moderate anxiety and depression from the PONDeR and IAPT trials.

Brief Psychiatric Rating Scale Expanded (BPRS-E)

The BPRS[32] was developed to assess symptom change in psychiatric inpatients and is one of the most widely used measures of psychotic and affective symptoms. The expanded version, BPRS–E which has 24 items developed for use in schizophrenia patients, was used in the current study. The BPRS-E is administered using semi-structured interviews and includes 24 items scored from 1 (not present) to 7 (extremely severe). In this study the

BPRS-E was used to assess the psychometric performance of the EQ-5D and SF-6D in a sample of patients with schizophrenia from the QUATRO trial.

Positive and Negative Syndrome Scale (PANSS)

The PANSS[33,34] 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 with detailed instructions on completion by interview. The 30 items are scored from 1 (absent) to 7 (extreme) and result in 3 subscales: positive, negative, and general psychopathology. The PANSS was used to assess the performance of the EQ-5D in a sample of patients with schizophrenia from the MATISSE trial.

Structured Interview for DSM-IV Personality (SIDP-IV)

Personality disorder diagnoses were assessed using the SIDP-IV.[35] 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 the 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. The SIDP-IV was used to assess the performance of the EQ-5D in a sample of patients with personality disorders from the SCEPTRE trial.

Analysis

Feasibility

The feasibility of administering the measures to respondents was assessed in terms of the level of completion of each measure at baseline. The baseline assessment was used since this was available for both the generic and the condition specific measures. Completion is a simplistic measure of feasibility from the patient's perspective, but provides an indication of the acceptability of the instruments using the level of missing data as a proxy. Completion rates of 95% or more were considered high.[36]

The analysis was carried out for both the common mental health condition group- where the measures were completed using self-report- and also for the schizophrenia and personality disorders samples- where the measures were interviewer-administered, but may still lead to missing data. It is important to note that although the level of completion may act as a proxy for feasibility, the results need to be interpreted with caution as it is not always clear how many questionnaires respondents have completed before those assessed in this study, the impact of fatigue on missing data, and other pressures placed on respondents to complete the measures.

Validity

Validity assesses how well an instrument measures what it was intended to measure, and is assessed in comparison to other instruments and clinical indicators that have been validated for use in the field. The validity of an instrument is assessed in light of the fact that there is no gold standard for the measurement of HRQL in mental health. This means that we can assess a range of indicators of validity, but cannot fully prove the validity of an instrument. We assessed validity by carrying out tests of discriminant or known group validity, and convergent validity.

Convergent validity

The convergence between the generic PBMs and the condition specific instruments was tested using Pearson's correlation coefficients and locally weighted scatterplot smoothing (LOWESS)[37] techniques. Good correlations indicate that the PBMs can measure mental health-related factors that are assessed by the validated condition-specific instruments. Correlations are considered weak if scores are <0.3, moderate if scores are ≥0.3 and <0.7, and strong if scores are ≥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, and demonstrates the relationship between the measures across the scoring range. LOWESS plots a line on a scatterplot on the central tendency between the two variables thereby visualising the relationship between these variables.

In the common mental health condition and mixed diagnosis groups, the convergent validity of EQ-5D was assessed in comparison to the HADS-T, HADS-A and HADS-D, and the SF-6D was assessed in comparison to the CORE-OM clinical and dimension scores. In the schizophrenia analysis, the 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. For personality disorders, tests of convergence between the EQ-5D and SIDP-IV were not carried out, as the SIDP-IV assesses fourteen personality disorders individually on a four point scale, and we did not believe that correlating each disorder indicator with the EQ-5D index score was appropriate.

Discriminant validity

The discriminant or known group validity analysis assesses the ability of the generic PBMs to discriminate between condition-specific severity groups. For the common mental health and mixed diagnosis samples, the discriminant validity of EQ-5D was assessed using HADS-A and HADS-D cut-off points indicating probable anxiety or depression (a score of ≥ 11). For SF-6D, discriminant validity was assessed using CORE-OM clinical cut off points (where a score of > 10 indicates clinical concerns)

One-way ANOVA was used to assess the magnitude of differences in the PBM scores across the severity groups. Standardised effect sizes across severity sub-groups were assessed (calculated as the difference in mean scores between two adjacent severity sub-groups divided by the standard deviation of scores for the milder of the two sub-groups). Effect sizes of less than 0.2 are small, 0.5 moderate, and 0.8 large.[38] However, care must be taken when comparing these between preference-based measures, since more is not necessarily better in terms of effect sizes (that simply indicate whether the generic PBM reflects what appears to be an important difference).[39]

The discriminant validity of the EQ-5D and SF-6D in the QUATRO schizophrenia sample used BPRS-E cut-offs (31 for 'mildly ill' , 41 for 'moderately ill', 53 for 'markedly ill' and 70 for 'extremely ill').[40] For the MATISSE sample, PANSS cut-offs (58 for "mildly ill", 75 for "moderately ill", 95 for "markedly ill" and 116 for "severely ill")[41] were used. For the SCEPTRE analysis, the discriminant validity of the EQ-5D was tested using diagnosis categories (defined as those with and without a personality disorder diagnosis, and also the number of personality disorders diagnosed).

Responsiveness

The responsiveness analysis assessed the sensitivity of EQ-5D and SF-6D to change in health status in comparison to the condition-specific PROMS. This included assessing floor and ceiling effects and the magnitude of the change in scores between two study time points. Floor (lowest possible score) and ceiling (highest possible score) effects impact the ability of the measure to detect deterioration or improvements in health respectively. The magnitude of change in scores is assessed before and after an intervention. We accept that this is a crude indicator of change. However for each study there was evidence of change between baseline and follow up. Where there has been an overall change then this should be reflected in by a significant change in the generic PBM score.

The magnitude of change reflected in the measures between the time points 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),[42] and the effect size. Again, effect sizes of less than 0.2 are small, 0.5 moderate, and 0.8 large.[38] Responsiveness analysis was not carried out for the mixed common mental health problem and personality disorder sample as only baseline data was available for the POPMACT study.

Results

Sample characteristics

Demographic characteristics available for each dataset are displayed in Table 1. The POPMACT sample has lower EQ-5D and HADS scores than the AHEAD sample, indicating higher levels of quality of life impairment and anxiety and depression (**Table 2**). SF-6D and CORE-OM scores indicate that the IAPT sample displays lowers levels of quality of life and functioning than the PONDeR sample. For the schizophrenia sample, baseline EQ-5D scores indicate that those in the MATISSE and QUATRO samples have similar quality of life levels. Those in the personality disorder (SCEPTRE) sample display lower quality of life as measured by the EQ-5D than the schizophrenia sample

Feasibility

Completion rates for assessment of feasibility are reported in Table 2.

Common mental health conditions

The AHEAD dataset demonstrates that both the EQ-5D and HADS had completion rates in the high range (97.86% to 99.08%) at baseline, for those participants who returned a questionnaire. Across the IAPT and PONDeR datasets, the completion rates for SF-6D and CORE-OM for those who returned a questionnaire at baseline were between 93.74% and 97.86%.

Common mental health and personality disorders

The EQ-5D and HADS were fully completed by more than 99% of the POPMACT sample who were interviewed

Schizophrenia and personality disorders

The EQ-5D completion rates (for those taking part at baseline where the measures were collected via interview) ranged between 96.33% and 98.56%. This was slightly higher than the SF-6D which had a completion rate of 93.64%. The condition specific PANSS and BPRS-E also had high completion rates (98.56% and 99.27%, respectively).

Personality disorders

The completion rate of the interviewer completed EQ-5D was 99.34%.

Convergent validity

Common mental health conditions

The correlation between the EQ-5D and HADS-T, HADS-A and HADS-D indicate a moderate level of convergence (**Table 3**). Negative correlations were produced, as a high score on the generic PBM and a low score on the condition specific measure indicates better health status. The SF-6D is correlated with the CORE-OM clinical score and functioning, wellbeing and symptoms domain scores in the moderate to strong range across both samples. The correlation with the risk domain score was moderate for the IAPT sample and low for the PONDeR sample. All correlations were significant ($p<0.01$).

Figure 1 displays scatterplots of the relationship between the generic and condition specific measures and the LOWESS fit lines. The lines demonstrate that the relationship between the EQ-5D and HADS differed across the severity scale (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 IAPT and PONDeR samples.

Common mental health and personality disorders

The correlation between the EQ-5D and HADS-T, HADS-A and HADS-D indicates a moderate level of convergence ($p<0.01$; **Table 3**). 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; while they were weak for the other dimensions (**Table 4**). Correlations with the PANSS and 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 4**). There was therefore poor evidence of convergence for SF-6D.

The LOWESS lines for the QUATRO sample (that completed both EQ-5D and SF-6D) demonstrate a tendency for the generic PBM scores to increase as scores on the BPRS-E decrease (equivalent to less severe problems on both measures). However, the EQ-5D displayed a large ceiling effect, meaning that a score of 1 on EQ-5D was associated with a wide range of BPRS scores. There was a trend towards a linear relationship between the EQ-5D and PANSS, and a large EQ-5D ceiling effect.

Personality disorders

Tests of convergence between the EQ-5D and SIDP-IV were not carried out.

Discriminant validity

Common mental health conditions

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 for the AHEAD sample (p=0.002). In both the IAPT and PONDeR samples, the SF-6D index score was significantly higher in the non clinical population in comparison to the clinical group as measured by CORE-OM (both P<0.001; **Table 5**).

Common mental health 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).

Schizophrenia

EQ-5D scores were significantly higher for those with a lower level of severity as measured by both the BPRS-E (p<0.001) and the PANSS (p=0.003) in the two schizophrenia samples (**Table 6**). Effect sizes across the severity sub-groups were moderate in size for the BPRS-E and small for the PANSS indicating that the EQ-5D can discriminate between severity groups to some extent.

The SF-6D scores significantly discriminated between BPRS-E severity groups, with scores in the most severe group higher than those for the EQ-5D. Effect sizes indicate that the difference between the mild and moderate severity groups was small.

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 6**). However, these

differences were not statistically significant (P=0.202). There is a significant difference in EQ-5D scores between samples with different types of personality disorder (*p*=0.042), but this is difficult to interpret.

Responsiveness

Common mental health conditions

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 7**). The SRM for EQ-5D was in the moderate range and for the HADS was large. This demonstrates that the HADS was more responsive in the AHEAD sample.

The SF-6D displayed a small ceiling effect for the PONDeR data. The SRM statistics for the SF-6D and CORE-OM in the IAPT validation sample were in the moderate range. For the PONDeR sample, the SF-6D SRM was in the large range, in contrast to the CORE-OM dimensions which were in the small range. Therefore, there was evidence that the responsiveness of SF-6D was in the same range as the CORE-OM for depression, and may even be more responsive in post-natal depression.

Common mental health and personality disorders

Responsiveness analysis was not carried out, as only baseline data was available.

Schizophrenia

For the QUATRO and MATISSE samples, the EQ-5D displays no evidence of floor effects at baseline, but there is evidence of a large ceiling effect at both time points (**Table 7**). Mean change in the QUATRO sample is statistically significant, but the effect sizes and SRM statistics are less than 0.2 (below the clinically significant range). The BPRS-E has a larger effect size and SRM statistic which indicates that EQ-5D was less responsive in this particular sample. The SF-6D displays no evidence of floor or ceiling effects in the QUATRO sample. Mean change on the SF-6D is smaller than the EQ-5D but the effect size and SRM statistics were consistently below 0.2.

In the MATISSE sample, mean change for EQ-5D is not statistically significant, leading to a small effect size and SRM statistics. The PANSS demonstrates statistically significant mean change, however, the effect sizes are in the low range. The small change demonstrated indicates that neither the EQ-5D nor PANSS are responsive in the MATISSE schizophrenia sample.

Personality disorders

In the SCEPTRE sample, EQ-5D displays minimal floor and ceiling effects and shows good responsiveness with moderate effect sizes and SRMs at 12 months.

Discussion

Seven datasets were used to examine the psychometric validity of the EQ-5D and SF-6D across a range of mental health conditions in comparison to widely used condition specific measures. The results suggest that the generic PBMs are valid for use in common mental health conditions and mixed diagnoses groups in comparison to existing measures of mental health, and there is some evidence of responsiveness to change in health status over time. For personality disorders, the results were also positive, as EQ-5D was shown to discriminate between severity groups, and respond to change over time. 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 to change was low.

Evidence for the psychometric validity of the EQ-5D and SF-6D in common mental health patient samples is consistent with previous empirical work in mild depression and anxiety samples.[6,8,9] A probable explanation for these positive findings is that both descriptive systems include mental health specific questions that are relevant to depression and anxiety. Therefore the measures may have a level of sensitivity to the conditions and some level of association with the widely used comparison measures. We have also established some evidence that the EQ-5D is valid in a sample with common mental health and personality disorder diagnoses leading to self harm. This group could be seen as a moderately severe sample (which is supported by the higher HADS scores in comparison to the AHEAD sample). There were some differences between the performance of the EQ-5D and SF-6D in common mental health samples, but direct comparisons were difficult because the analysis of each measure was carried out using different samples with different characteristics. The growing evidence base regarding the validity of the instruments indicates that EQ-5D and SF-6D can be considered valuable for the use in the economic evaluation of interventions for common mental health disorders.

The positive results found for the personality disorders sample is in line with past work in the area[12] which found that the EQ-5D correlates with condition specific indicators, and responds to change over time. This indicates that EQ-5D has some level of validity for use in the assessment of interventions for personality disorders. We compared EQ-5D to a

diagnosis instrument completed by clinicians, and it would be useful to use a self or interviewer administered PROM as a comparator.

Past work has found mixed evidence for the performance of generic PBMs in schizophrenia.[10] In this study we have established evidence for and against the validity of the generic PBMs in schizophrenia, and there was mixed evidence regarding the ability of the measures to reflect schizophrenia-specific symptoms. The EQ-5D may be related to some condition specific domains (for example depression) but not others (such as positive symptoms), and again this may be linked to the classification system which directly assesses anxiety and depression, but may not be sensitive to other schizophrenia-specific domains. Direct comparisons between the EQ-5D and SF-6D were only possible for the QUATRO study, which found that neither instrument converges with the condition specific measure (but this may not be expected), and neither instrument responds to change over time at the same level that is reflected in the condition-specific indicators. The low level of responsiveness for EQ-5D may be due to the large ceiling effect at baseline which may impair its ability to detect change over time. The mixed evidence regarding the schizophrenia sample means that the EQ-5D and SF-6D should be used with caution in these groups, and further research in other samples to investigate psychometric performance in more detail is warranted.

Psychometric analysis of the PBMs is one method of assessing validity, and should be considered alongside other types of evidence to establish a detailed picture of the performance of these measures. For example this work should be considered alongside systematic reviews[9,10] and qualitative work assessing the content validity and acceptability of the instruments from the patient perspective. This allows for detailed 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 sensitivity of the instruments could be improved. This includes the development of 'bolt on' dimensions for the generic PBMs to directly assess particular conditions. Alternatively, condition specific PBMs could be developed either using standard instrument development procedures or by adapting an existing condition specific instrument, examples of which are available for other neurological and mental health conditions.[43-46] A five level version of the EQ-5D has been developed,[47] and it is possible that this version may be more sensitive to different severity levels and change across time. Further research could assess the validity of the five-level version in patients with mental health conditions.

This study has a number of limitations. Firstly, as in much psychometric validation, there is no 'gold standard' measure of HRQL against which to compare the generic PBMs. This means that the analysis provides a guide to the performance of the measure, but is limited by the validity of the comparison indicator and the constructs tested by this instrument. This means that the results are open to interpretation and opinion. In this study it can be argued that the generic PBMs are compared against indicators that have some level of validity in the populations tested,[24,30,31] and this allows some inferences to be drawn. However, the different scope of the condition specific and generic measures and the different development procedures used, suggests that some level of divergence is to be expected. Furthermore, the condition specific measures used here assess specific symptoms, in comparison to the generic PBMs which include dimensions of health-related quality of life. Therefore, we may not expect especially close concordance between them. The same concerns apply when testing responsiveness, and it is important to consider if the measures of health change are themselves valid.

Secondly, we used the level of missing data as a form of proxy for the feasibility of the measure in mental health populations. This approach can be criticised, as the external pressures and expectations felt by respondents in trials to complete the measures is unclear. There are a range of reasons why measures are incomplete that do not specifically relate to the measure, including severity of the mental health condition, fatigue, lack of motivation, or the position of the questionnaires in the study.

The inferences that can be drawn from the results are also limited to the mental health conditions included in the seven datasets, and the generalisability to other populations should be investigated. Moreover, the differing levels of performance in terms of construct validity, convergent validity, and responsiveness also reflect the systematic variance attributable to different types of data used in this study, the different patient populations, and different study designs. Furthermore, only one dataset included both EQ-5D and SF-6D. This means that the level of transferability to other mental health samples with similar diagnoses but different characteristics is unclear, and full comparisons between the generic PBMs are not possible. Further work into the performance of the EQ-5D and SF-6D in mental health conditions should focus on replicating the current analysis on different mental health conditions using different condition-specific measures and indicators, and directly comparing the generic PBMs. It is also possible that due to their generic nature, the PBMs are picking up co morbidities, however this is difficult to test in the data available, as

indicators of other conditions (including physical conditions) were not available. It would be useful to attempt to assess the impact of comorbidities on utility scores in mental health populations.

In summary, we have reported the first work to test the psychometric performance of two widely used generic preference based measures of health related quality of life across a range of populations with mental health disorders 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 HRQL in mental health.

References

1 National Institute of Health and Clinical Excellence. *Guide to the Methods of Technology Appraisal*. London: 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.

5 Brazier JE, Roberts J. Estimating a preference-based index from the SF-12. *Med Care* 2004; **42(9)**: 851-59.

6 Brazier, J. Measuring and valuing mental health for economic evaluation. *Health Services Research and Policy* 2008; **13(S3)**: 70-75.

7 Brazier J. Is the EQ-5D fit for purpose in mental health? *Br J Psychiatry* 2010; **197**: 348-349

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**: 1229-236.

9 Peasgood T, Brazier JE. The psychometric validity of EQ-5D and SF-6D in anxiety and depression: A systematic review. *HEDS Discussion Paper*, **12**/**15**.

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(6)**: 907-20.

11 Barton, G.R., Jones, P.B., Croudace, T., Fowler, D. Measuring the benefits of treatment for psychosis: validity and responsiveness of the EQ-5D. *Br J Psychiatry* 2009; **195**: 170-177.

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. *Psychiatric Services* 2005; **56(9)**.

13 Kendrick T, Peveler R, Longworth L, Baldwin D, Moore M, Chatwin J, Thornett A, Goddard J, Campbell M, Smith H, Buxton M, Thompson C. Cost-effectiveness and cost-utility of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine. *Br J Psychiatry* 2006; **188**: 337-45.

14 Morrell CJ, Slade P, Warner R, Paley G, Dixon S, Walters SJ, Brugha T, Barkham M, Parry G, Nicholl, J. 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.

15 Parry G, Barkham M, Brazier J, Dent-Brown K, Hardy G, Kendrick T, Rick J, Chambers E, Chan T, Connell J, Hutten R, de Lusignan S, Mukuria C, Saxon D, Bower P, Lovell K. *An*

evaluation of a new service model: Improving Access to Psychological Therapies demonstration sites 2006-2009. Final report. NIHR Service Delivery and Organisation programme 2011.
16 Byford S, Knapp M, Greenshields J, Ukoumunne OC, Jones V, Thompson S, Tyrer P, Schmidt U, Davidson K, Catalan J. 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.
17 Tyrer P, Tom B, Byford S, Schmidt U, Jones V, Davidson K, Knapp M, MacLeod A, Catalan J; POPMACT Group. Differential effects of manual assisted cognitive behavior therapy in the treatment of recurrent deliberate self-harm and personality disturbance: the POPMACT study. *Journal of Personality Disorders* 2004; **18(1)**: 102-16.
18 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.
19 Crawford M, Killaspy H, Kalaitzaki E, Barrett B, Byford S, Patterson S. et al. (2010). The MATISSE study: a randomised trial of group art therapy for people with schizophrenia. *BMC Psychiatry* 2010; **10**: 65.
20 Soeteman D, Verheul R, Busschbach J. (2008). The burden of disease in personality disorders: diagnosis-specific quality of life. *J Personal Disord* 2008; **22**: 259-68.
21 Canadian Agency for Drugs and Technologies in Health. *Guidelines for the Economic Evaluation of Health Technologies: Canada (3rd Edn)*. Ottawa: Canadian Agency for Drugs and Technologies in Health 2006.
22 Pharmaceutical Benefits Advisory Committee. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (PBAC)*. Commonwealth of Australia 2008.
23 Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand* 1993; **67**: 361-70.
24 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.
25 Barkham M, Evans C, Margison F, McGrath G, Mellor-Clark J, Connell J, Milne D. The rationale for developing and implementing core batteries for routine use in service settings and psychotherapy outcome research. *Journal of Mental Health* 1998; **7**: 35-47.
26 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.
27 Barkham M, Margison F, Leach C, Lucock M, Mellor-Clark J, Evans C, Benson L, Connell J, Audin K, McGrath G. Service profiling and outcome benchmarking using the CORE-OM:

Towards practice based evidence in the psychological therapies. *J Consult Clin Psychol* 2001; **69**: 184-96.
28 Evans C, Connell J, Barkham M, Margison F, McGrath G, Mellor-Clark J, Audin K. Towards a standardised brief outcome measure: psychometric properties and utility of the CORE-OM. *Br J Psychiatry* 2002; **180(1)**: 51-60.
29 Evans C, Mellor –Clark J, Margison K, Barkham M, Audin K, Connell J, McGrath G. CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health* 2000; 9(3): 247-55.
30 Barkham, M., Mullin, T., Leach, C., Stiles, W.B., & Lucock, M. Stability of the CORE-OM and the BDI-I prior to therapy: Evidence from routine practice. *Psychology and psychotherapy: Theory, Research and Practice* 2007; 80: 269-78.
31 Gilbody S, Richards D, Barkham M. Diagnosing depression in primary care using self-completed instruments: UK validations of PHQ-9 and CORE-OM. *B J Gen Pract* 2007; **57(541)**: 650-52.
32 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. *International Journal of Methods in Psychiatric Research* 1993; **3**: 227-43.
33 Kay SR, Fiszbein A, Opler LA. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr Bull* 1987; 13: 261-76.
34 Kay SR, Opler LA, Lindenmayer JP. Reliability and validity of the positive and negative syndrome scale for schizophrenics. *Psychiatry Res* 1988; 23: 99-110.
35. Pfohl B, Blum N, Zimmerman M. *Structured Interview for DSM-IV Personality: SIDP-IV*. Washington, DC: American Psychiatric Press; 1995.
36 Nunnally, JC, Bernstein, IH. *Psychometric theory, 3rd ed*. New York: McGraw Hill; 1994)
37 Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* 1979; **74(368)**: 829–36.
38 Cohen J. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum 1988.
39 Brazier JE, Deverill M. A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Econ* 1999; **8(1)**: 41-52.
40 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.
41 Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel RR. What does the PANSS mean? *Schizophr Res* 2005; **79**: 231-38.
42 Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM: On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res* 2003; **12**: 349-62.

43 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(3)**: 321-33.
44 Mulhern B, Rowen D, Jacoby A, Marson T, Snape D, Hughes D, Latimer N, Baker G, Brazier J. (2012). The development of a QALY measure for epilepsy: NEWQOL-6D. *Epilepsy and Behavior*; **24(1)**: 36-43.
45 Mulhern B, Smith SC, Rowen D, Brazier JE, Knapp M, Lamping DL, Loftus V, Young TA, Howard RJ, Banerjee S. Improving the measurement of QALYs in dementia: developing patient- and carer-reported health state classification systems using Rasch analysis. *Value Health* 2012; **15(2)**: 323-33.
46 Rowen D, Mulhern B, Banerjee S, van Hout B, Young T, Knapp M, Smith S, Lamping D, Brazier J. Estimating preference based single index measures for dementia using DEMQOL and DEMQOL-Proxy. *Value Health* 2012; **15(2)**: 346-56.
47 Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G, Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res* 2011; **20(10)**: 1727-36.

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 datasets used in this study.

Table 1: Summary of datasets and measures used for analysis

Dataset	Description	Time points used	PBM	Condition specific	Characteristics		
					N (baseline)	Age (m,sd)	Female (%)
Common mental health							
Assessing Health Economics of Antidepressants (AHEAD)[13]	RCT of the cost effectiveness of antidepressants (three time points). Sample includes common mental health concerns - depression, mixed anxiety and depressive disorders (MADD), and phobias	Baseline 6 months 12 months	EQ-5D	HADS	327	43.1 (15.4)	67.0
Psychological interventions for postnatal depression (PONDeR)[14]	Study of two psychologically informed interventions for women with post-natal depression	Baseline 6 months 12 months	SF-6D	CORE-OM	2,640	31.5 (5.1)	100
Improving Access to Psychological Therapies cohort study (IAPT)[15]	Evaluation of the outcomes from two IAPT demonstration sites. Sample reports common mental health condition.	Baseline 4 months 8 months	SF-6D	CORE-OM	527	40.9 (14.2)	72.3
Common mental health and personality disorders							
POPMACT[16,17]	RCT of manual-assisted cognitive therapy (MACT) versus treatment as usual (TAU) in recurrent self-harm. Sample includes patient with diagnoses of common mental health conditions and personality disorders leading to self harm	Baseline	EQ-5D	HADS	480	32.0 (11.2)	
Schizophrenia/personality disorders							
Quality of Life following Adherence Therapy (QUATRO)[18]	Multicounty RCT of adherence therapy in patients with a clinical diagnosis of schizophrenia for needed continuing antipsychotic medication for at least a year from assessment, and had evidence of clinical instability in the year before assessment. Previous work comparing EQ-5D and SF-6D utilities using this data (McCrone et al., 2009).		EQ-5D SF-6D	BPRS-E	409	41.5 (11.5)	59.9
Multi-centre study of Art Therapy In Schizophrenia - Systematic Evaluation (MATISSE)[19]	RCT of group art therapy for people with a clinical diagnosis of schizophrenia.	Baseline 12 months	EQ-5D	PANSS	417	41.0 (11.5)	33.3
Study of cost effectiveness of personality disorder treatment (SCEPTRE)[20]	Dutch study of adult patients with personality disorders (UK EQ-5D tariff used for comparability).	Baseline 12 months	EQ-5D	DSM-IV personality disorder category	932	35.1 (9.8)	68.1

Table 2: Descriptive statistics and completion rates

		N (completing measure)	Mean	SD	Completion (%)
Common mental health – EQ-5D					
AHEAD					
EQ-5D	Baseline	320	0.604	0.264	97.86
	6m	174	0.752	0.257	
	12m	164	0.777	0.249	
HADS-A	Baseline	324	13.11	3.48	99.08
	6m	202	8.78	3.54	
	12m	169	8.30	3.59	
HADS-D	Baseline	324	10.50	3.87	99.08
	6m	202	4.94	3.92	
	12m	169	4.34	3.99	98.83
Common mental health – SF-6D					
IAPT					
SF-6D	Baseline	504	0.613	0.13	95.6
	4 months	425	0.645	0.14	
	8 months	390	0.668	0.15	
CORE-OM clinical score	Baseline	494	20.06	7.81	93.74
	4 months	409	16.58	8.53	
	8 months	403	15.09	8.83	
PONDeR					
SF-6D	Baseline	2600	0.669	0.09	97.82
	6 months	2614	0.822	0.14	
	12 months	1697	0.839	0.13	
CORE-OM clinical	Baseline	2640	5.23	4.97	99.32
	6 months	2641	4.73	4.91	
	12 months	1713	4.52	4.85	
Common mental health and personality disorders					
POPMACT					
EQ-5D	Baseline	476	0.503	0.320	99.20
HADS-A	Baseline	479	14.13	3.94	99.80
HADS-D	Baseline	478	11.22	4.58	99.58
Schizophrenia					
MATISSE					
EQ-5D	Baseline	409	0.676	0.271	98.08
	12 months	357	0.678	0.297	
PANSS Total score	Baseline	411	79.45	24.19	98.56
	12 months	334	76.15	27.11	
QUATRO					
EQ-5D	Baseline	394	0.679	0.291	96.33
	12 months	367	0.710	0.286	
SF-6D	Baseline	383	0.668	0.125	93.64
	12 months	367	0.682	0.134	
BPRS-E Total	Baseline	406	45.17	13.02	99.27
	12 months	371	37.71	10.54	
Personality disorders					
SCEPTRE					
EQ-5D	Baseline	899	0.566	0.284	99.34
	12 months	693	0.741	0.249	

Table 3: Convergent validity of EQ-5D and SF-6D in common mental health, and joint diagnosis samples

	Common mental health			Joint diagnosis
	AHEAD	**IAPT**	**PONDeR**	**POPMACT**
	EQ-5D index	SF-6D index	SF-6D index	EQ-5D index
HADS-T	-0.36**	-	-	-0.49**
HADS-A	-0.35**	-	-	-0.39**
HADS-D	-0.22**	-	-	-0.46**
CORE-OM				
Clinical score	-	-0.61**	-0.51**	-
Functioning score	-	-0.51**	-0.46**	-
Symptoms score	-	-0.64**	-0.53**	-
Wellbeing score	-	-0.51**	-0.45**	-
Risk score	-	-0.37**	-0.16	-

**= significant at 0.01

Table 4: Convergent validity of EQ-5D and SF-6D in schizophrenia

	QUATRO		**MATISSE**
	EQ-5D index	SF-6D index	EQ-5D index
BPRS-E total	-0.42**	-0.29**	-
BPRS-E disorganisation	-0.22**	-0.13**	-
BPRS-E depression	-0.43**	-0.34**	-
BPRS-E negative symptoms	-0.21**	-0.12**	-
BPRS-E positive symptoms	-0.31**	-0.20**	-
PANSS total	-	-	-0.16**
PANSS positive	-	-	-0.12
PANSS negative	-	-	-0.05
PANSS general symptoms	-	-	-0.21**

**= significant at 0.01

Table 5: Discriminant validity of EQ-5D and SF-6D in common mental health and joint diagnosis samples

Data and indicator	Groups		n	
Common mental health – EQ-5D (AHEAD)				
HADS-A caseness	No case	Mean (sd)	98	0.671 (0.25)
	Probable case	Mean (sd)	219	0.573 (0.27)
		P value		0.002
		ES		0.37
HADS-D caseness	No case	Mean (sd)	163	0.677 (0.24)
	Probable case	Mean (sd)	154	0.525 (0.27)
		P value		0.000
		ES		0.60
Common mental health - SF-6D (IAPT)				
CORE-OM clinical	Non clinical	Mean (sd)	53	0.740 (0.11)
	Clinical	Mean (sd)	422	0.597 (0.12)
		P value		0.000
		ES		1.24
SF-6D sample (PONDeR)				
CORE-OM clinical	Non clinical	Mean (sd)	2241	0.683 (0.08)
	Clinical	Mean (sd)	399	0.595 (0.07)
		P value		0.000
		ES		1.16
Common mental health and personality disorders - EQ-5D (POPMACT)				
HADS-A caseness	No case	Mean (sd)	84	0.718 (0.28)
	Probable case	Mean (sd)	392	0.457 (0.31)
		P value		0.000
		ES		0.88
HADS-D caseness	No case	Mean (sd)	210	0.622 (0.30)
	Probable case	Mean (sd)	265	0.410 (0.30)
		P value		0.000
		ES		0.71

Table 6: Discriminant validity of EQ-5D and SF-6D in schizophrenia and personality disorders

			EQ-5D			SF-6D	
Variable	Groups	N	Mean (sd)	ES	N	Mean (sd)	ES
Schizophrenia							
QUATRO							
BPRS-E	None/mild (24-31)	56	0.831 (0.16)		56	0.727 (0.09)	
	Moderate (32 – 41)	112	0.744 (0.23)	0.54*	112	0.682 (0.12)	0.47
	Marked (42 – 53)	119	0.652 (0.28)	0.40*	119	0.664 (0.12)	0.16
	Severe (>53)	88	0.543 (0.36)	0.39	88	0.625 (0.13)	0.31
				p=0.000			*p=0.000*
MATISSE							
PANSS	Normal/mild (30 -58)	86	0.747 (0.24)				
	Moderate (59 – 75)	114	0.693 (0.23)	0.23			
	Marked (76 – 95)	112	0.660 (0.27)	0.14			
	Severe (>95)	92	0.606 (0.33)	0.20			
				p=0.003			
Personality disorders							
Number of diagnosis	No Personality Disorder	84	0.648 (0.24)				
	1 Personality Disorder	248	0.606(0.27)	0.18			
	2 Personality Disorder	95	0.549 (0.30)	0.21			
	3 Personality Disorder	42	0.493 (0.29)	0.18			
	≥4 Personality Disorders	41	0.416 (0.27)	0.27			
				p=0.202			
Diagnosis [a]	None	113	0.657 (0.23)				
	Borderline	41	0.581 (0.29)	0.32[b]			
	Avoidant	69	0.638 (0.25)	0.08 [b]			
	Obsessive-compulsive	55	0.578 (0.27)	0.34 [b]			
	Depressive	60	0.525 (0.28)	0.56 [b]			
	Not otherwise specified	142	0.616 (0.29)	0.17 [b]			
				p=0.042			

* *p<0.05* in test of difference between adjacent severity groups
a – Sample of those with a single diagnosis
b – compared to the group with no personality disorder

Table 7: Responsiveness of generic and condition specific measures

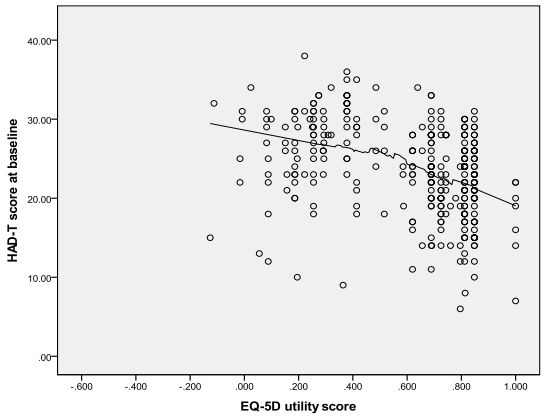
Measure	% at floor		% at ceiling		Mean change	ES	SRM	T-test
	T0	T1	T0	T1				
Common mental health								
EQ-5D								
AHEAD (n=164)								
EQ-5D	0	0	2.19	34.15	0.17 (0.38)	0.64	0.45	
HADS-T	0	0	0	0	-10.74 (8.83)	-1.85	-1.22	
HADS-A	0	0	0	0	-4.81 (4.98)	-0.70	-0.97	
HADS-D	0	0	0.62	14.79	-5.93 (5.67)	-0.68	-1.05	
SF-6D								
IAPT (n=390)								
SF-6D	0	0	0	1.54	-0.06 (0.12)	0.46	0.50	
CORE-OM clinical score	0	0	0	0	-4.71 (6.71)	0.60	-0.70	
Functioning score	0.41	0	0.82	1.50	-0.37 (0.75)	0.44	-0.49	
Symptoms score	1.22	1.24	0.20	0.50	-0.58 (0.84)	0.62	-0.70	
Wellbeing score	7.46	2.72	0.81	3.95	-0.57 (0.97)	0.63	-0.59	
Risk score	0.2	0	39.27	54.48	-0.18 (0.55)	0.22	-0.32	
PONDeR (n=1,697)								
SF-6D	0	0	0	18.33	0.17 (0.13)	1.89	1.31	
CORE-OM clinical score	0	0	3.48	7.82	-0.58 (4.69)	-1.16	-0.12	
Functioning score	0	0.06	12.35	17.24	-0.04 (0.57)	-0.06	-0.07	
Symptoms score	0	0	8.60	16.13	-0.10 (0.57)	-0.17	-0.18	
Wellbeing score	0	0.06	20.14	29.77	-0.10 (0.76)	-0.13	-0.13	
Risk score	0.04	0	90.23	89.55	-0.01 (0.20)	-0.07	-0.05	
Schizophrenia								
QUATRO (n=328)								
EQ-5D	0	0	16.8	20.7	0.035 (0.29)	0.12	0.12	0.026
SF-6D	0	0.3	0.6	0.9	0.014 (0.12)	0.12	0.12	0.027
BPRS-E	1.2	4.3	0	0	-7.60 (13.06)	-0.58	-0.58	0.000
BPRS-E positive	17.1	26.8	0	0	-3.04 (5.70)	-0.52	-0.53	0.000
BPRS-E negative	21.3	35.1	0	0	-1.37 (4.06)	-0.33	-0.34	0.000
BPRS-E disorganisation	20.1	36.9	0	0	-1.62 (4.22)	-0.42	-0.38	0.000
BPRS-E depression	0	15.9	0	0	-1.90 (5.41)	-0.34	-0.35	0.000
MATISSE (n=321)								
EQ-5D	0	0	16.8	20.2	-0.005 (0.29)	-0.02	-0.02	0.767
PANSS	0	0	0	0	-3.41 (20.85)	-0.16	-0.14	0.004
PANSS positive	2.5	3.4	0	0	-0.93 (6.17)	-0.15	-0.15	0.007
PANSS negative	2.2	4.0	0	0.3	-0.78 (6.48)	-0.12	-0.11	0.031
PANSS general symptoms	0.3	0	0	0	-1.21 (10.65)	-0.11	-0.10	0.042
Personality disorders								
SCEPTRE (n =679)								
EQ-5D	0	0	4.0	21.6	0.170 (0.29)	0.61	0.58	0.000

N: Those who completed both measures at all time points

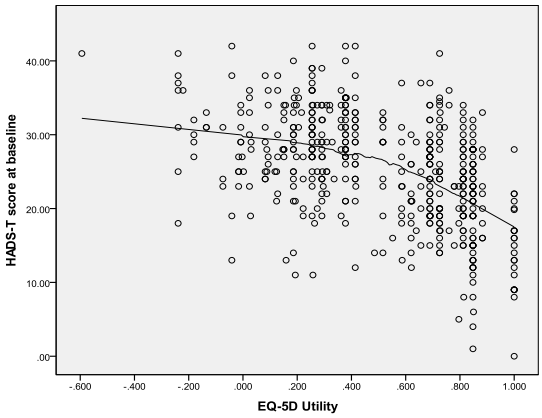
ES/SRM size – small: >0.2 ≤0.5, moderate: >0.5 <0.8, large: ≥0.8; T0: Baseline, T1: Follow up

Figure 1: Scatterplots/LOWESS lines for the common mental health and mixed diagnosis conditions

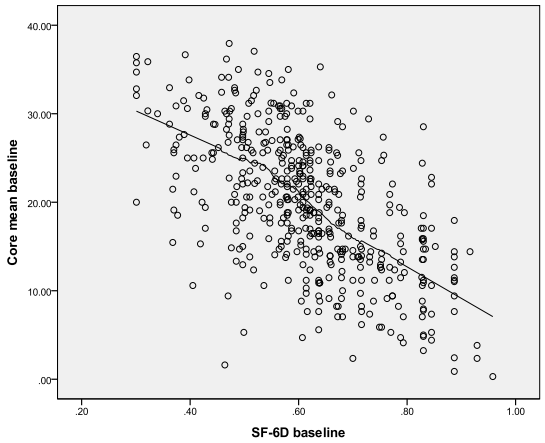
AHEAD



POPMACT



IAPT



PONDeR

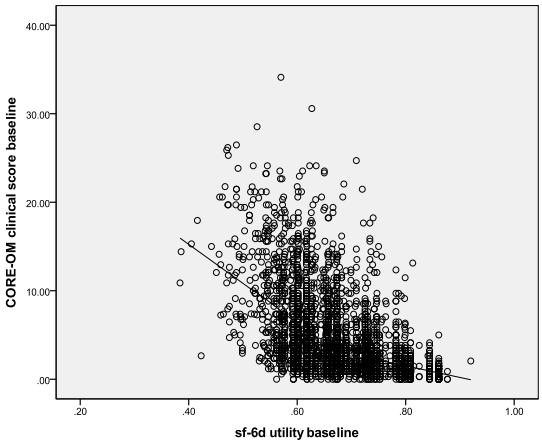
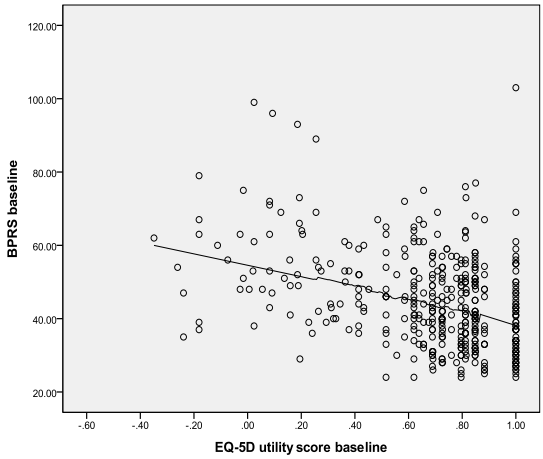
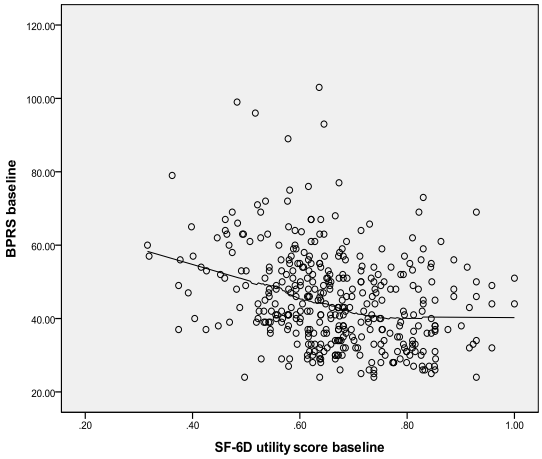


Figure 2: Scatterplots/LOWESS lines for the schizophrenia and personality disorder samples

QUATRO (EQ-5D)



QUATRO (SF-6D)



MATISSE