HEDS Discussion Paper

No.13/02

How to measure quality of life for cost effectiveness analyses in personality disorders? A systematic review.

Diana Papaioannou[1] BChD, MSc, John Brazier[1] BA, MSc, PhD and Glenys Parry[1] BA, Dip.Clin.Psych., PhD, C.Psychol., FBPsS

[1] School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA University of Sheffield

JPD MS#1541R-07-11 Accepted, FINAL version

How to measure quality of life for cost effectiveness analyses in personality disorders? A systematic review.

Diana Papaioannou[1] BChD, MSc, John Brazier[1] BA, MSc, PhD and Glenys Parry[1] BA, Dip.Clin.Psych., PhD, C.Psychol., FBPsS

[1] School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA University of Sheffield

Corresponding author: Diana Papaioannou, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA. Email: d.papaioannou@sheffield.ac.uk

How to measure quality of life for cost effectiveness analyses in personality disorders? A systematic review.

Abstract

Objective: To assess the construct validity and responsiveness of four generic health related quality of life (HRQL) measures in personality disorders (PDs).

Methods: A systematic review was undertaken. Ten databases were searched and reference lists scrutinised to identify relevant studies. Relevant data were extracted accordingly. A narrative synthesis was performed of the evidence on construct validity including known groups validity (detecting differences in HRQL scores between two different groups), convergent validity (strength of association between generic HRQL and other measures (e.g. symptom) and responsiveness (differences in generic HRQL measure scores in responders/non-responders or correlation with changes in other measures).

Results: Ten studies were identified, with 6 for the EQ-5D, two involving SF-36 and another 2 the SF-12, but none on the SF-6D. Evidence indicated that the EQ-5D, SF-36 and SF-12 were probably valid measures within PDs. Four studies demonstrated that the EQ-5D Index was able to detect changes in patients.

Conclusion: Generic HRQL measures appear appropriate for use in people with PDs in terms of psychometric performance. However, qualitative concerns remain as to whether they fully reflect the impact of the condition.

Introduction

The effectiveness of mental health interventions can not be exclusively measured in terms of their ability to improve clinical outcomes such as reduction in symptom severity. The last decade has seen the increased use of economic evaluation and particularly cost effectiveness analyses to inform resource allocation decisions. This usually takes the form of health technology assessment reports submitted to agencies, for example, the National Institute of Clinical Excellence (NICE) in England, who advise governments on the treatments of choice and make decisions on which treatments to fund at a national level.

The measurement of health-related quality of life (HRQL) fulfils a key role within economic evaluations. Health care policymakers require interventions to be assessed in terms of their cost per Quality Adjusted Life Year (QALY). The QALY is a way of measuring the benefits of health care interventions on a common scale and incorporating improvements in HRQL; thus taking into account improvements from the perspective of the patient. HRQL measures inform the 'quality' aspect of QALYs and the most commonly used method for putting the 'q' into the QALY is generic preference-based HRQL instruments such as the EQ-5D (Brooks, 1996; Rasanen et al., 2006). Generic in this case means relevant to all patient groups, including those suffering from physical as well as those suffering from mental health problems. Preference-based HRQL measures have a multi-dimensional classification to define specific health states to which existing utility weights can be attached (McCrone et al., 2009); thus providing a valuation of each health state. Utilities provide a value that indicates the strength of preference an individual has for a specific health state or outcome (Tolley, 2009). The individual making the valuation can be the patient, clinician or the general population, though in the case of the EQ-5D it is a sample of the general population.

There is strong evidence that interventions, such as psychotherapy, are clinically effective in personality disorders. A review by Bartak, Soeteman, Verheul and Busschbach (2007) found overwhelmingly positive evidence for the clinical effectiveness of psychotherapy for personality disorders, in the form of two meta-analyses (Leichsenring and Leibing, 2003; Perry, Banon and Ianni, 1999), six reviews (Gabbard, 2000; Ogrodniczuk and Piper, 2001; Perry and Bond, 2000;

Sainslow and McGlashan, 1998; Shea 1993; Bateman and Fonagy, 2000) and one Cochrane review (Binke et al., 2006). Treatments for PDs also appear to be cost-effective. Gabbard, Lazar, Hornberger and Spiegel (1997)reviewed the costs of psychotherapeutic treatment for personality disorders, and concluded that psychotherapy would lead to cost savings, attributable to reductions in inpatient treatments and decreases in work impairment.

In general, studies in personality disorders present costs without explicitly relating them to the effects of an intervention (Bartak et al., 2007). Studies typically present the overall cost savings of one treatment in comparison to another (Ranger et al., 2009; Abbass, Sheldon, Gyra and Kalpin, 2008; Stevenson and Meares, 1999) or occasionally qualify the cost of the intervention in terms of an outcome or event avoided (e.g. costs per avoided parasuicide event were presented in six trials included in a health technology assessment on psychological therapies in borderline personality disorder) (Brazier et al., 2006).

More recently, cost-effectiveness analyses have been published that provide policymakers with the preferred outcome of cost per QALY gained. For example, three studies used the EQ-5D to calculate the cost per QALY gained to measure the cost-effectiveness of psychotherapy in cluster B personality disorders (Soeteman et al., 2010), schema- or transference-focused psychotherapy in borderline personality disorder (van Asselt et al., 2008) and cognitive behavioural therapy for borderline personality disorder (Palmer et al., 2006). There has also been one study that used the SF-6D to evaluate the cost effectiveness of psychoanalysis versus psychoanalytic therapy (Berghout, Zevalkink, Hakkaart-van Roijen, 2010). However, the psychometric properties of generic HRQL measures in this patient group are largely unknown, and in general doubts have been raised about the appropriateness of generic measures in mental health (Brazier, 2010).. Whilst, it is claimed that the EQ- 5D and other 'generic' measures such as the SF-6D are applicable to all interventions and patient groups, including many physical conditions where these instruments have managed to pass psychometric tests of reliability and validity (Marra et al., 2005); it is unknown if generic HRQL measures can 'measure what they are intended to measure' (i.e. are valid) and can detect changes in HRQL in patients when a change in severity in personality disorders has occurred (i.e. are responsive).

Systematic reviews of generic HRQL measures have been undertaken to determine what or if any evidence exists for their validity and responsiveness in physical health conditions e.g. vision (Tosh, Brazier, Evans and Longworth, 2012) and mental health conditions e.g. schizophrenia (Papaioannou, Brazier and Parry, 2011). Such reviews have found mixed evidence on the validity and responsiveness of generic HQRL measures in a particular condition. In addition, reviews have also identified a lack of evidence for particular generic HRQL measures e.g.SF-6D.

This is the first attempt to systematically review the literature to identify if evidence exists to demonstrate that generic measures are valid and responsive in measuring HRQL in personality disorders. Four generic HRQL measures were evaluated in this review, two generic health status measures (SF-36 and SF-12) and two preference based measures (EQ-5D and SF-6D).

The measures studied

The SF-36® is a generic health status profile measure consisting of eight dimensions of general health (GH); bodily pain (BP); physical functioning (PF); role-physical (RP), mental health (MH); vitality (V); social functioning (SF) and role-emotional (RE). These eight dimensions also can be used to generate a physical and mental health summary scores. (Ware, Kosinski and Keller, 1994) The SF-12® (Ware, Kosinski and Keller, 1996) is a shortened version of the SF-36, containing 12 SF-36 items, and also produces two weighted summary scores (PCS and MCS).

The EQ-5D valuation questionnaire comprises a five dimensional questionnaire and an EQ-5D Visual Analogue Scale (VAS). Respondents are asked to report their level of problems (no problems, some/moderate problems or severe/extreme problem) on the questionnaire which includes mobility, self-care, usual activities, pain and anxiety/depression, to provide a position on the EQ-5D health state classification. Responses can be converted into one of 243 different health state descriptions (ranging from no problems on any of the dimensions [11111] to severe problems on all five dimensions [33333]) which each have their own preference-based score. Preference-based scores are determined by eliciting preferences i.e. establishing which health states are preferred from a population sample. In order to do so, a method such as time trade off

is used which involves asking participants to consider the relative amounts of time (for example, number of life-years) they would be willing to sacrifice to avoid a certain poorer health state (Tolley, 2009). Utility values from the UK EQ-5D can range from -0.59 to 1, where negative values are felt to be worse than death and a value of 1 indicates perfect health (Dolan, 1997). These health state utility values can be used to calculate quality adjusted life years (QALYs) for cost effectiveness analysis

The EQ-5D VAS reports on the respondents' self-rated valuation of their health stated and thus is based on the preferences of the patients, but is not preference based and not normally used to generate QALYs.

The SF-6D is a preference-based measure of health that can be generated from items of the SF-36 or SF-12 (Brazier, Roberts and Deverill, 2002; Brazier and Roberts, 2004). The SF-6D has a classification that describes health on six multi-level dimensions of physical functioning, role limitations, social functioning, pain, mental health and vitality. There are algorithms for scoring each state based on values obtained from general population surveys using standard gamble (respondents make a series of choices which allow estimation of the strength of preferences regarding a health state). Health state utility values range from 0.29 to 1.0.

Methods

Identification of studies

A literature search was performed to identify relevant research for a wider review including other mental health conditions. The search included searching for 'personality disorders' using database thesaurus and free text terms. Two sets of search terms were combined: terms for each of the four HRQL measures AND terms for the personality disorders. (Search strategies are available from authors). Ten databases were searched for published research: Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, NHS Economics and Evaluations Database, Health Technology Database, Database of Abstracts of Reviews of Effects, MEDLINE, PreMEDLINE, CINAHL, EMBASE, Web of Science. Searches were limited to English Language only but not by any date restriction. All searches were conducted in August 2009, and update searches were conducted in March 2011 for two

conditions (personality disorders and bipolar disorder) to identify new relevant studies. The reference lists of relevant studies were searched for further papers.

Citations identified by the searching process were screening by one reviewer (DP). Studies were eligible for inclusion if they contained HRQL data using one or more of the following instruments: SF-36, SF-12, SF-6D or EQ-5D within the specified population: adults (18 years +) with a personality disorder. HRQL data could be from descriptive systems (i.e. their items and dimensions) or health state utility values generated by the EQ-5D or SF-6D or the EQ-5D VAS. Studies whose primary focus was on individuals with alcohol and/or drug dependency with a comorbid personality disorder were excluded. The outcomes had to include data that allowed measurement of the construct validity (i.e. known groups, convergent) or the responsiveness of the HRQL instrument (s). The full texts of papers were retrieved for any titles or abstracts that appeared to satisfy the inclusion criteria, or for which inclusion or exclusion could not be definitely determined. In addition, 10% of citations were double-screened for the wider systematic (which included other mental health conditions) at abstract-level. Both reviewers were in complete agreement on the decisions made in this sample of citations. Subsequently, the same inclusion and exclusion criteria were used to assess full papers by one reviewer (DP) and where queries over inclusion arose, these were resolved by discussion and consensus with another reviewer (JB).

Data extraction

Data from all included trials were extracted by one reviewer (DP) using a form designed specifically for this review, and piloted on a sample paper. Data extracted included: country of publication, type of disorder, study sample characteristics (numbers, age, gender), other measures used, mean values for HRQL measures (per population and per collection), type and method of validity assessment, type and method of responsiveness assessment and validity and responsiveness data.

Evidence synthesis and meta-analysis

Due to the heterogeneity of study designs, HRQL instruments and methods of determining construct validity and responsiveness, it was not appropriate to perform meta-analysis. Analysis was by narrative synthesis and data were tabulated.

Defining validity and responsiveness

For the purposes of this review, construct validity and responsiveness an HRQL were defined and assessed as follows:

1) Known groups or extreme groups validity was demonstrated when an HRQL measure was able to detect differences in HRQL between two groups who differ in a trait or behavior (Streiner and Norman, 2003). For e.g. HRQL scores in individuals with personality disorders could be hypothesised as being lower than that of healthy individuals.

2) Convergent validity was demonstrated by the relationship an HRQL measure had with other measures of the same construct (Streiner and Norman, 2003). Convergent validity was defined as the correlation between two measures that in theory are associated. The strength of correlation between the two instruments is calculated using statistical tests (for e.g. Pearson's product moment correlation, Spearman's rank correlation).

3) Responsiveness was demonstrated by the extent to which an HRQL measure was able to detect a clinically significant or practically important change over time (Walters, 2009). For example, this might be demonstrated by strong correlation between change scores on a clinical measure and an HRQL measure or by the HRQL being able to detect differences in HRQL between patients defined as responders and non-responders.

The application of these psychometric criteria to preference-based HRQL measures requires some adaptation (Brazier and Deverill, 1999). The purpose of EQ-5D or SF-6D is to identify all differences or changes in health that are important to patients and valued by the general public. An item of the EQ-5D, for example, may fail to pick up small differences in one dimension or miss another health dimension entirely, but if these are not important to patients and not valued by the general population, then it is not a weakness of the instrument. For example, in convergent validity, it would be important to use an instrument hypothesised as being likely to have a strong relationship to preferences to test convergence of a generic preference based HRQL measure

such as the EQ-5D. Similarly, where known groups validity is being tested, the groups in which the HRQL measure must detect a difference in scores, must be hypothesised as being likely to differ in HRQL.

Results

The search for studies for the wider review on a range of mental health conditions retrieved 4115 unique citations (see Figure 1). Of these, 3849 were excluded at the title and abstract stage and 266 were examined in full-text; of which seven articles on personality disorders met the criteria for inclusion in this review. In the update searches, a further 623 unique citations were retrieved, and of these 11 were examined at full-text level. A further three studies on personality disorders were included in this review from the update searches. In total, ten studies including data on personality disorders were included at full-text level.

Study characteristics

Six studies looked at the EQ-5D (Bartak et al., 2010; Bartak et al., 2011a; Bartak et al., 2011b; Soeteman et al., 2005; Soeteman, Verheul and Busschbach, 2008; van Asselt et al., 2009), two at the SF-36 (Narud, Mykletun and Dahl, 2005; Hueston, Mainous and Schilling, 1996) and two at the SF-12 (corresponding to 3 articles) (Jackson and Burgess, 2000; Jackson and Burgess, 2002; Sanderson and Andrews, 2002). No studies were found investigating the validity or responsiveness of the SF-6D in this patient group. Six studies were undertaken in the Netherlands (Bartak et al., 2010; Bartak et al., 2011a; Bartak et al., 2011b; Soeteman et al., 2005; Soeteman, Verheul and Busschbach, 2008; van Asselt et al., 2009). Two studies were undertaken in Australia (Jackson and Burgess, 2002; Sanderson and Andrews, 2002). One study was undertaken in Norway (Narud et al., 2005) and one study in the USA (Hueston et al., 1996). Nine of the ten studies presented data for different personality disorders together (Bartak et al., 2010; Bartak et al., 2011a; Bartak et al., 2011b; Soeteman et al., 2005; Soeteman, Verheul and Busschbach, 2008; Narud et al., 2005; Hueston et al., 1996; Jackson and Burgess, 2002; Sanderson and Andrews, 2002) . One study looked exclusively at individuals with borderline personality disorder (van Asselt et al., 2009).

The number of individuals included within the studies that were diagnosed or screened as having one or more personality disorders ranged from 48 to 1,708. Participants included males and females (proportions can be seen in Table 1). The mean age of participants with personality disorders reported in nine of the ten studies ranged between 29.4 to 45 years. The two reports on the study by Jackson and Burgess did not provide information on age (Jackson and Burgess, 2000; Jackson and Burgess, 2002)

Validity and responsiveness

Two studies investigated the known groups validity of the EQ-5D (Soeteman et al., 2005; Soeteman, Verheul and Busschbach, 200); one study investigated the convergent validity of the EQ-5D (Soeteman et al., 2005) and four studies investigated the responsiveness of the EQ-5D (Bartak et al., 2010; Bartak et al., 2011a; Bartak et al., 2011b; van Asselt et al., 2009). Two studies investigated the known groups validity of the SF-36 (Narud et al., 2005; Hueston et al., 1996) and two studies investigated this property in the SF-12 (Jackson and Burgess, 2002; Sanderson and Andrews, 2002). One study investigated the responsiveness, and convergent validity of the SF-36 (Narud et al., 2005).

<u>EQ-5D</u>

Soeteman et al. (2005) looked at the use of the EQ-5D within a group of individuals described as having 'complex personality problems and personality disorders', although no formal diagnoses are provided. The EQ-5D demonstrated that these individuals had much lower EQ-5D index scores in comparison to a non-clinical population (0.54 vs. 0.85). In addition, EQ-5D scores were moderately correlated with the Global Severity Index scores (0.49). In another study, Soeteman et al. (2008) looked at the use of the EQ-5D within all personality disorders. Having a borderline, narcissistic, obsessive-compulsive, depressive, negativistic or personality disorder mixed had a significant effect on the EQ-5D score. However, it was the number of personality disorders rather than the type of personality disorder that had a large effect on EQ-5D score (p=0.000). When controlling for the number of disorders in the linear regression, only depressive personality disorder maintained a unique statistically significant effect on QoL (p=0.03).

Van Asselt et al. (2009) assessed the responsiveness of the EQ-5D amongst a group of individuals with borderline personality disorder (BPD). Three years after baseline in an RCT where patients had received one of two types of psychotherapy, the EQ-5D-Index and EQ-VAS showed significant moderate correlation (0.487 and 0.404 respectively, $p<0.01$) with the change scores on the Borderline Personality Disorder Severity Index-IV (BPDSI-IV) (a clinical measure of the severity of BPD). The EQ-5D was also able to detect differences in patients who recovered and had not recovered three-years post-baseline according to change scores on the BPDSI-IV ($p=0.000$). Bartak et al. (2010, 2011a, 2011b) conducted three studies in individuals with cluster A (Bartak et al., 2011b), cluster B (Bartak et al., 2011a) and cluster C PDs ((Bartak et al., 2010). However, within each study there were substantial numbers of patients who were diagnosed with PDs from more than one cluster (see table 1 for details) and thus it was unclear if patients were counted twice between the three studies. In the cluster A and B studies, patients were assigned to one of three settings for psychotherapeutic treatment (outpatient, day hospital or inpatient). In the study investigating Cluster C PDs, patients were allocated to one of five different treatment modalities: long-term outpatient treatment, short-term or long-term day hospital treatment or short-term or long-term inpatient treatment.

In each of the three studies, EQ-5D scores were measured pre-and post-treatment per treatment group. Effect sizes were largely moderate to strong when scores pre-and post-treatment were compared, thus the EQ-5D Index appears responsive to change in patients. In addition, the EQ-5D Index was able to detect differences in scores between difference treatment modalities.

SF-36

Narud et al. (2005) found that the SF-36 could distinguish between individuals with personality disorders and age- and gender-adjusted norms ($p<0.001$). However, there were no significant differences in the mean MCS or PCS when patients with one, two or three or more personality disorders were compared. Hueston et al. (1996) found that scores were significantly lower in individuals screened as high risk of a personality disorder compared with individuals at low risk for personality disorder on mental health ($p=0.01$), physical functioning ($p=0.04$), role limitations due to emotional problems ($p=0.03$) and general health ($p=0.05$) dimensions.

Narud et al. (2005)found correlations between the SF-36 and clinical and functioning measures were weak to moderate. Correlation with the Global Assessment of Functioning (GAF) ranged from 0.12 to 0.40, with the Symptom Checklist 90-Revised (SCL-90R) Personality Severity Index (PSI) from 0.12 to 0.38 and the Social Adjustment Scale (SAS) as 0.23 to 0.45. In general, correlations were stronger between the mental health domains and the measures. However, comorbid Axis I disorders explained a significant part of the scores on the MCS (Narud et al., 2005).

Narud et al. (2005)noted significant improvement on clinical and functioning measures, whilst the SF-36 only showed significant improvement on the role limitations due to physical problems and mental health dimensions. Considerable but not significant changes were observed for the bodily pain.

SF-12

As part of the Australian National Survey of Mental Health and Wellbeing undertaken in 1997, SF-12 values were found to be lower in individuals with on or more personality disorders when compared with individuals who had no personality disorders ($p<0.0001$) on both the MCS and PCS (Jackson and Burgess, 2000). These data were re-examined at a later date ((Jackson and Burgess, 2002)and the authors found that an increase in the number of personality disorders reduced the SF-12 MCS (i.e. greater disability), even when controlling for comorbid Axis I disorders.

Sanderson and Andrews (2002) used linear regression to predict the difference in SF-12 scores between individuals with personality disorder diagnoses and individuals with no mental health disorder, whilst controlling for the influence of sociodemographic variables and co-occurring mental and chronic physical conditions. Individuals with personality disorders were 10.6 points lower in SF-12 score. There were no data on the responsiveness of the SF-12 within personality disorders.

Discussion

Six studies contained data on the EQ-5D, two studies involved the SF-36 and two studies contained data on the SF-12. No studies were identified that examined the SF-6D in individuals diagnosed or at high risk of personality disorders. The EQ-5D appears responsive in individuals with personality disorders. Data on other properties such as convergent and known groups validity were very limited. There was also little evidence on the SF-36 or SF-12. Nevertheless, the studies which did exist provided some positive evidence that the measures are valid for use in personality disorders. An exception was Narud et al. (2005)who found that most dimensions on the SF-36 were not able to detect changes in patients in the same way as clinical measures and concluded that this may be because some SF-36 dimensions are not relevant to HRQL so that, even if patients change clinically, this does not translate to a change in health related quality of life.

There is of course a third conclusion to draw from Narud et al. (2005), that the SF-36 dimensions do not capture the aspects of HRQL that are important to patients, and thus can not detect any change in those aspects. This review has been limited to examining quantitative evidence. However, increasingly researchers and regulatory authorities such as the FDA and EMEA (U.S.Department of Health and Human Services Food and Drug Administration CDER & CDRH, 2009; European Medicines Agency, 2005) also require qualitative evidence on the validity of measures in specific patient groups based on interviews with patients. There are studies that have examined the concepts important to quality of life from patients' perspective within mental health conditions in general (Mayers, 2000) and within individual conditions such as bipolar disorder (Michalak, Kolesar and Lam, 2006) and schizophrenia and other psychotic conditions (Cook and Chambers, 2009). These studies have found that individuals with mental health conditions regard concepts such as stigma (Michalak, Kolesar and Lam, 2006; Cook and Chambers, 2009), lack of personal achievement, loneliness and personal safety (Mayers, 2000) as having a big impact on their quality of life. However, such concepts are not incorporated into generic HRQL measures.

Within personality disorders, we know of just one study which investigated the experiences of symptoms, suffering, and life situation in a group of ten individuals with borderline personality disorder (Persieus, Ekdahl, Zsberg and Samuelsson, 2005). Rapid mood swings, putting on a 'mask or normality' when feeling very different on the inside, self-hate, and fear relationships or fear of life/longing for death were all highlighted by study participants as increasing their suffering. Therefore, these are factors, which aren't directly incorporated in generic HRQL measures and are important to consider when measuring HRQL in this patient group. Nonetheless, they may be indirectly reflected in the generic dimensions of these measures (Brazier et al., 2012). Indeed the quantitative evidence reviewed in this paper suggests that these condition specific problems are reflected in the generic dimensions, and that it is reasonable to recommend the EQ-5D in particular for use in this population. However, a concern remains that there is an important qualitative dimension to some of these PD specific problems that are not being measured. It is possible that qualitative evidence may identify key dimensions which could be 'added' onto the EQ-5D to improve its content validity.

Despite the paucity of literature on quality of life within personality disorders (Narud et al. 2005); Narud and Dahl, 2002) and the limited evidence available on the validity and responsiveness of generic HRQL measures, cost-effectiveness analyses using generic preference-based HRQL measures have been undertaken and have demonstrated that interventions and therapies in this patient group can be cost-effective (Soeteman et al., 2010; van Asselt et al., 2008 and Palmer et al., 2006). More cost effectiveness analyses need to be undertaken within this patient group using preference-based measures.

Strengths and limitations

This review is the first to have comprehensively identified studies that report on the construct validity and responsiveness of the EQ-5D, SF-36, SF-12 and SF-6D within personality disorders and to tabulate and give a narrative synthesis of the findings. However, the review has some limitations. Whilst the search for studies was reasonably comprehensive, it was limited to key databases and reference list checking of included studies, and study selection was undertaken by one reviewer. Ideally, further searching could be undertaken in trial registries, conference

proceedings and by citation searching to make the search process more comprehensive. Study quality assessment has not been undertaken and there are issues to consider when interpreting the findings. Four studies had small sample sizes (N<100) (Bartak et al., 2011b; van Asselt et al., 2009; Narud et al., 2005; Hueston et al., 1996) and all but one study (van Asselt et al., 2009) presented overall findings for all personality disorders combined. Given the heterogeneous nature of different types of personality disorders, this may not be appropriate as HRQL can vary according to the type of personality disorder (Cramer, Torgerson and Kringlen, 2006). In addition, the level of co-morbidity with other conditions was reported as high in three studies (Narud et al., 2005; Hueston et al., 1996; Jackson and Burgess, 2002), although two studies did adjust their analyses accordingly. Indeed, Jackson and Burgess (2002) (found that even where co-morbidities existed, personality disorders still had an additional effect on HRQL. Finally, five of the ten studies were cross-sectional in design and evidence could be strengthened by undertaking more longitudinal studies. Nevertheless, this review gives an overall picture of the validity and responsiveness of the HRQL measures within personality disorder.

In conclusion, the limited quantitative evidence available shows that generic HRQL measures, including preference-based measures and in particular the EQ-5D, are responsive within this patient group. However, further research, particularly on validity and on the SF-6D, would strengthen the use of generic HRQL measures within individuals with personality disorders. Incorporating qualitative research and the aspects deemed important by patients is also important to further examine what may be missing. .

Acknowledgements

This study was funded by the Medical Research Council Methodology Board (project number G0801394).

References

Abbass A., Sheldon A., Gyra J., Kalpin A. (2008). Intensive short-term dynamic psychotherapy for DSM-IV personality disorders: a randomized controlled trial. *Journal of Nervous and Mental Disease*.196(3):211-6.

Bartak A., Soeteman D.I., Verheul R., Busschbach J.J.V. (2007). Strengthening the Status of Psychotherapy for Personality Disorders: An Integrated Perspective on Effects and Costs. *Canadian Journal Of Psychiatry.* 52(12):803-10.

Bartak A., Spreeuwenberg M.D., Andrea H., Holleman L., Rijnierse P., Rossum B.V., et al. (2010). Effectiveness of different modalities of psychotherapeutic treatment for patients with cluster C personality disorders: results of a large prospective multicentre study. *Psychotherapy and Psychosomatics* 79(1):20-30.

Bartak A., Andrea H., Spreeuwenberg M.D., Ziegler U.M., Dekker J., Rossum B.V., et al. (2011a). Effectiveness of outpatient, day hospital, and inpatient psychotherapeutic treatment for patients with cluster B personality disorders. *Psychotherapy and Psychosomatics.* 80(1):28-38.

Bartak A, Andrea H, Spreeuwenberg MD, Thunnissen M, Ziegler UM, Dekker J, et al. (2011b). Patients with cluster A personality disorders in psychotherapy: an effectiveness study. *Psychotherapy and Psychosomatics* .80(2):88-99.

Bateman A.W., Fonagy P. (2000) Effectiveness of psychotherapeutic treatment of personality disorder. *British Journal of Psychiatry*.177:138-43.

Berghout C.C., Zevalkink J., Hakkaart-van R.L. (2010). A cost-utility analysis of psychoanalysis versus psychoanalytic psychotherapy. *International Journal of Technology Assessment in Health Care.*26(1):3-10.

Binks C.A., Fenton M., McCarthy L., Lee T., Adams C.E., Duggan C.(2006). Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systetmatic Reviews*.(1):CD005652.

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

Brazier J., Roberts J., Deverill M. (2002).The estimation of a preference based measure of health from the SF-36. *Journal of Health Economics*.21:271-92.

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

Brazier J., Tumur I., Holmes M., Ferriter M., Parry G., Dent-Brown K, et al. (2006). Psychological therapies including dialectical behaviour therapy for borderline personality disorder: a systematic review and preliminary economic evaluation. *Health Technology Assessment.*10(35):iii, ix-iii,117.

Brazier J. (2010) Is the EQ-5D fit for purpose in mental health? *British Journal of Psychiatry.*197, 348–349.

Brazier J., Rowen D., Mavranezouli I., Tsuchiya A., Young T., Yang Y. (2012) Developing and testing methods for deriving preference-based measures of health from condition specific measures (and other patient based measures of outcome). *Health Technology Assessment*. Forthcoming

Brooks R (1996). EuroQol: the current state of play. *Health Policy* ;37(1):53-72.

Cook S., Chambers E. (2009). What helps and hinders people with psychotic conditions doing what they want in their daily lives. *British Journal of Occupational Therapy*.72(6):238-48.

Cramer V., Torgerson S., Kringlen E. (2006). Personality disorders and quality of life. A population study. *Comprehensive Psychiatry.*47(178):184.

Dolan P.*(1997).* Modelling valuation for EuroQol health states*. Medical Care.* 35*:* 351–363*.*

European Medicines Agency Pre-authorisation Evaluation of Medicines for Human Use. (2005). Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products.

Gabbard G.O., Lazar S.G., Hornberger J., Spiegel D. (1997) The economic impact of psychotherapy: a review. *American Journal of Psychiatry.*154(2):147-55.
Gabbard G.O.(2000). Psychotherapy of personality disorders. Journal of Psychotherapy Practice and Research .9(1):1-6.
Hueston W.J., Mainous A.G., , Schilling ,(1996). Patients with personality disorders: functional status, health care utilization, and satisfaction with care. *Journal of Family Practice.*42(1):54-60.
Jackson H.J, Burgess P.M. (2000) Personality disorders in the community: a report from the Australian National Survey of Mental Health adn Wellbeing. *Social Psychiatry & Psychiatric Epidemiology.*35(12):531-8.
Jackson HJ, Burgess PM. (2002). Personality disorders in the community: results from the Australian National Survey of Mental Health and Wellbeing Part II. Relationships between personality disorder, Axis I mental disorders and physical conditions with disability and health consultations. *Social Psychiatry & Psychiatric Epidemiology.*37(6):251-60.
Leichsenring F., and Leibing E. (2003). The effectiveness of psychodynamic therapy and cognitive behavior therapy in the treatment of personality disorders: a meta-analysis. *American Journal of Psychiatry*.160(7):1223-32.
Marra C.A., Rashidi A.A., Guh D., Kopec J.A., Abrahamowicz M., Esdaile J.M., et al. (2005). Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? *Quality of Life Research*.14(5):1333-44.
Mayers C.(2000). Quality of Life: Priorities for People with Enduring Mental Health Problems. *British Journal of Occupational Therapy*.63(12):591-7.
McCrone P., Patel A., Knapp M., Schene A., Koeter M., Amaddeo F., et al. (2009). A comparison of SF-6D and EQ-5D utility scores in a study of patients with schizophrenia. *Journal of Mental Health Policy and Economics.*12(1):27-31.
Michalak E.E., Kolesar S., Lam R.W. (2006). Bipolar disorder and quality of life: a patient-centered perspective. *Quality of Life Research.*15(25):37.
Narud K., Dahl A.A. (2002). Quality of life in personality and personality disorders. *Current Opinion in Psychiatry.*15(2):131-3.
Narud K., Mykletun A., Dahl A.A.,(2005). Quality of life in patients with personality disorders seen at an ordinary psychiatric outpatient clinic. *BMC Psychiatry*.5:10.
Ogrodniczuk J.S, Piper W.E. (2001). Day treatment for personality disorders: a review of research findings. *Harvard Review of Psychiatry.*9(3):105-17.
Palmer S., Davidson K., Tyrer P., Gumley A., Tata P., Norrie J., et al. (2006).The cost-effectieveness of cognitive behavior therapy for borderline personality disorder: Results from the BOSCOT trial. .20(5):466-81.
Papaioannou, D., Brazier J., Parry, G,(2011). How valid and responsive are generic health status measures, such as EQ-5D and SF-36, in Schizophrenia? (A systematic review) . *Value Health* . 2011;14:907–920
Perry J.C., Banon E., Ianni F. (1999). Effectiveness of psychotherapy for personality disorders. *American Journal of Psychiatry.* 156(9):1312-21.
Perry J.C, Bond M. (2000). Empirical studies of psychotherapy for personality disorder. In: Gunderson JG, Gabbard GO, editors. *Psychotherapy for personality disorders. Review of psychiatry.*Washington (DC): American Psychiatric Publishing.

Persieus K.I., Ekdahl S., Zsberg M., Samuelsson M. (2005). To Tame a Volcano: Patients with Borderline Personality Disorder and Their Perceptions of Suffering. *Archives of Psychiatric Nursing.*19(4):160-8.
Ranger M., Tyrer P., Miloseska K., Fourie H., Khaleel I., North B., et al. (2009). Cost-effectiveness of nidotherapy for comorbid personality disorder and severe mental illness: randomized controlled trial. *Epidemiologia e psichiatria sociale.*18(2):128-36.
Räsänen P., Roine E., Sintonen H., Semberg-Konttinen V., Ryynänen O., Roine R. (2006). Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. *International Journal of Technology Assessment in Health Care.* 22:235-41.
Sainslow C.A., McGlashan T.H. (1998) Treatment outcome of personality disorders. *Canadian Journal Of Psychiatry.*;43(3):250.
Sanderson K, Andrews G.(2002). . Prevalence and severity of mental health-related disability and relationship to diagnosis. Psychiatric Services.53(1):80-6.

Shea M.T. (1993). Psychosocial treatment of personality-disorders. J*ournal of Personality Disorders*.7:167-80.
Soeteman D.I., Verheul R., Delimon J., Meerman A.M., van den Eijnden E., Rossum B.V., et al. (2010). Cost-effectiveness of psychotherapy for cluster B personality disorders. *British Journal of Psychiatry.* 196(5):396-403.
Soeteman D.I., Timman R., Trijsburg R.W., Verheul R., Busschbach J.J.,, et al. (2005). Assessment of the burden of disease among inpatients in specialized units that provide psychotherapy. *Psychiatric Services* .56(9):1153-5.
Soeteman D.I., Verheul R., Busschbach J.J.V. (2008). The burden of disease in personality disorders: diagnosis-specific quality of life. *Journal of Personality Disorders* .22(3):259-68.
Stevenson J., Meares R. (1999). Psychotherapy with borderline patients: II. A preliminary cost benefit study. *Australian and New Zealand Jouranl of Psychiatry*.33(4):473-7.
Streiner D.L., Norman G.R. (2003). Health Measurement Scales: A practical guide to their development and use. USA: Oxford University Press
Tolley K .(2009). What are health utilities? . Retrieved from: http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/Health-util.pdf [Accessed July 13, 2012]
Tosh J., Brazier J. , Evans, P.and Longworth, L.(2012). A Review of Generic Preference-Based Measures of Health-Related Quality of Life in Visual Disorders. *Value in health* 15 (1): 118-127
U.S.Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER) Center for Devices and Radiological Health (CDRH). (2009). Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.
van Asselt A.D., Dirksen C.D., Arntz A., Giesen-Bloo J.H., van D.R., Spinhoven P., et al. (2008). Out-patient psychotherapy for borderline personality disorder: cost-effectiveness of schema-focused therapy v. transference-focused psychotherapy. *British Journal of Psychiatry.* 192(6):450-7.
van Asselt A.D., Dirksen C.D., Arntz A., Giesen-Bloo J.H., Severens J.L., , et al. (2009).The EQ-5D: A useful quality of life measure in borderline personality disorder? *European Psychiatry: the Journal of the Association of European Psychiatrists* .24(2):79-85.
Walters S.J. (2009). Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation: A Practical Guide to analysis and interpretation. WileyBlackwell

Ware J.E., Kosinski M., Keller S.D.,(1994) SF-36® Physical and mental health summary scales: a user's manual. Boston: MA: The Health Institue.
Ware J.E., Kosinski M., Keller S.D.,(1994) SF-36® Physical and mental health summary scales: a user's manual. Boston: MA: The Health Institue
Ware J.E., Kosinski M., Keller S.D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care.*34(3):220-33.

.

Figure 1: Flow diagram of study identification- see separate file (JPEG)

Table 1: Validity and responsiveness of the SF-36, SF-12 and EQ-5D in personality disorders.

Study details	Population (N) and characteristics	Measurement and Practical Properties	
		Properties measured	**Details of validity or responsiveness**
SF-36			
Narud, K (2005)(42) Norway	Personality disorders N=91 (48 females and 43 males) (Mean age= 36.6 (10.5) Age range 19 to 74 years	Known groups validity convergent validity Responsiveness	The SF-36 could distinguish between individuals with PDs and age- and gender-adjusted norms ($p<0.001$). There were no significant differences in the mean MCS or PCS when patients with one, two or three or more PDs were compared. Weak to moderate correlation with clinical (SCL-90R) and functioning measures (GAF, SAS) on the MCS mostly, correlations ranged from 0.12 to 0.45. Only the role limitations due to physical problems and mental health dimensions showed significant changes post-treatment.
Hueston, W J (1996)(43) USA	Patients 'at risk' of PDs: determined by completion of the SCID for DSM-III-R for personality disorders N=93 N=65 patients at high risk of one or more PDs; Mean age= 44.7 (15.3); 17 males and 48 females N=28 patients at low risk of PDs; Mean age= 39.7 (15.1); 6 males and 22 females	Known groups validity	Scores were significantly lower individuals at high risk of personality disorder compared with individuals at low risk for personality disorder on the following SF-36 dimensions: Mental health ($p=0.01$); Physical functioning ($p=0.04$); Role limitations due to emotional problems ($p=0.03$) ; General health ($p=0.05$)
SF-12			
Jackson, H I (2000)(44)/ Jackson, H I (2002)(45) Australia	N=10, 641 4705 males and 5936 females Age data not reported. N=704 (319 males and 385 females) had at least one PD	Known groups validity	The SF-12 could distinguish between individuals with or without personality disorders. However, further analyses indicated this effect may be due to comorbid Axis I conditions.
Sanderson, K (2002)(46) Australia	Sample taken from Australian National Mental Health and Well-being survey (N=10,641)- a nationally representative household survey of mental disorders in	Known groups validity	Linear regression was used to predict the difference in SF-12 scores between individuals with personality disorder diagnoses and

Study details	Population (N) and characteristics	Measurement and Practical Properties	
		Properties measured	**Details of validity or responsiveness**
	adults. 5214 males and 5427 females Participants in survey mean age= 45 years No mental disorder: n=9, 902 Any personality disorder: n=564		individuals with no mental health disorder, whilst controlling for the influence of sociodemographic variables and co-occurring mental and chronic physical conditions. Individuals with personality disorders were 10.56 points lower in score.
EQ-5D			
Bartak, A (2011)(38) Netherlands	DSM-IV-TR axis II cluster A personality disorders N=57 17 males and 40 females Mean age= 29.4 (+/-8.2) **PD diagnosis** 'Pure' cluster A, n=9; Cluster A and B n= 7; Cluster A and C n= 18 Cluster A, B and C n=23 Paranoid PD n= 49; Schizoid PD n=5; Schizotypal PD n=4 All received psychotherapeutic treatment in one fo three settings: Outpatient n=20 Day hospital n=19 Inpatient n=18	Responsiveness	Effect sizes for the EQ-5D Index scores pre- and post- treatment were moderate for the outpatient and inpatient groups (0.47 and 0.59 respectively) and strong for the day hospital group (0.85). For the EQ-5D VAS, the effect sizes were strong for the day hospital group and inpatient group (1.03 and 0.73 respectively) whilst weak for the outpatient group (0.04) When the EQ-5D Index and VAS change scores were compared across the three treatment groups, the Index showed moderate to strong effect sizes (outpatient vs. day hospital 0.83, outpatient vs. inpatient 0.68 and day hospital vs. inpatient 0.83). For the EQ-5D VAS, the effect sizes were weak for two comparisons (outpatient vs. day hospital 0.02 and day hospital vs. inpatient 0.21). The effect size was strong when outpatient vs. inpatient VAS scores were compared (0.99).
Bartak A (2011)(37) Netherlands	DSM-IV-TR axis II cluster B personality disorders N=207 70 males and 147 females Mean age= 31.3 (SD 8.5)	Responsiveness	Effect sizes were weak to strong when EQ-5D Index scores were compared pre- and post-treatment (outpatient 0.37, day hospital 0.72 and inpatient 0.80)

Study details	Population (N) and characteristics	Measurement and Practical Properties	
		Properties measured	**Details of validity or responsiveness**
	PD diagnosis Pure cluster B n=84; Cluster B and C n=93; Cluster B and A n=7; Cluster A, B and C n=23 Borderline PD n=160; Narcissistic PD n=47; Histrionic PD n=26; Antisocial PD n=18 All received psychotherapeutic intervention in one of three settings: Outpatient n=46; Day hospital n=81; Inpatient n=80		When the change scores of the EQ-5D Index were compared across the three treatment groups effect sizes were weak for the comparisons of outpatient vs. inpatient (0.16) and day hospital vs. inpatient (0.18), whilst strong for outpatient vs. day hospital (0.71).
Bartak, A (2010)(36) Netherlands	DSM-IV-TR axis II cluster C personality disorders N=371; 110 males and 261 females; Mean age= 33.5 (SD 9.5) Pure cluster C n=247; Cluster C and B n=88; Cluster C and A n=15 Cluster, C, B, and A n= 21 Avoidant PD n=235; Obsessive-compulsive PD n= 183; Dependent PD= 84 All received one of five modalities of psychotherapeutic interventions Long outpatient n=68; Short day hospital n=77;Long day hospital n=74;Short inpatient n=59; Long inpatient n=93	Responsiveness	Effect sizes for the pre- and post-EQ-5D Index scores were strong for the Long day hospital and Short inpatient treatment groups (0.90 and 1.21 respectively) moderate for the Long inpatient and long outpatient treatment groups (0.67and 0.74 respectively) and weak for the short day hospital treatment group (0.37). EQ-5D scores also improved significantly more in the short-term inpatient group than in 2 other groups: the short-term day hospital group (β= 0.15, p=0.0009, 95% CI 0.06-0.23) and the long-term inpatient-group (β=0.11, p=0.0113, 95% CI 0.03-0.19)
Soeteman, D I (2005)(39) Netherlands	Complex personality problems or personality disorders N=1, 651 (541 male and 1110 female) Mean age= 31.9 years (SD=9.2) Age range 18 to 61 years	Known groups validity Convergent validity	Substantial differences in EQ-5D scores between a nonclinical population and the study population. Moderate correlation with a functioning measure (GSI).
Soeteman, D I (2008)(40) Netherlands	N=1, 708 (605 males and 1103 females). Mean age= 33.7 years (SD=9.9) Age range= 18 to 67 years	- Known groups validity	The greater the number of personality disorders, the worse the EQ-5D score (p=0.000)

Study details	Population (N) and characteristics	Measurement and Practical Properties	
		Properties measured	**Details of validity or responsiveness**
Van Asselt, A D (2009)(41) Netherlands	Borderline personality disorder N= 48 ; Mean age of completers was 31 years (SD=8.55); 43 females and 5 males.	Responsiveness	Able to detect difference between patients who had recovered and not recovered three years post baseline. Also showed moderate correlation with change scores on a severity index (BPDSI-IV)

BPDSI-IV = Borderline Personality Disorder Severity Index-IV; GAF= Global Assessment of functioning ; GSI=Global Severity Index; SAS= Social adjustment scale; SCID= Structured Clinical Interview; SCL-90R= Symptom checklist 90-Revised