PREFERENCE-BASED ASSESSMENT

How Valid and Responsive Are Generic Health Status Measures, such as EQ-5D and SF-36, in Schizophrenia? A Systematic Review

Diana Papaioannou, BChD, MSc1,*, John Brazier, BA, MSc, PhD2, Glenys Parry, BA, DipClinPsych, PhD, CPsychol, FBPsS3

1School of Health and Related Research, University of Sheffield, Sheffield, UK

ABSTRACT

Objectives: Generic health status measures such as the short form health survey (SF-36) and EuroQol-5D (EQ-5D) are increasingly being used to inform health policy. They are claimed to be applicable across disease areas and have started to be used within mental health research. This review aims to assess the construct validity and responsiveness of four generic health status measures in schizophrenia, including the preference-based SF-6D and EQ-5D. Method: A systematic review of the literature was undertaken. Ten databases were searched from inception to August 2009 and reference lists scrutinized to identify relevant studies. Studies were appraised and data extracted. A narrative synthesis was performed of the evidence on construct validity including known groups validity (detecting a difference in health-related quality of life (HRQL) scores between two different groups such as samples from the general population and people with schizophrenia), convergent validity (strength of association between generic HRQL and other measures (e.g., symptom or functional), and responsiveness. Responsiveness was considered by: 1) differences in generic HRQL measure scores in responders/non-responders and 2) correlation between changes on generic HRQL measures and changes in specific measures obtained from patients and clinicians. Results: Thirty-three studies were identified that provided data on the validity and/or responsiveness of the instruments. Most of the evidence concerns the SF-36 and EQ-5D, and for these instruments there was evidence for known group validity. The evidence for convergent validity and responsiveness was mixed, with studies presenting contradictory results. Conclusion: Although the evidence base is limited in a number of important respects, including problems with the measures used to develop constructs in the validation studies, it is sufficient to raise doubts about the use of generic measures of health like the EQ-5D and SF-36 in patients with schizophrenia.

Keywords: EQ-5D, generic health status measures, health-related quality of life, preference-based measures, quality of life, schizophrenia, SF-36, SF-12, SF-6D.

Copyright © 2011, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc. Open access under - CC BY license.

Introduction

Generic health status measures such as short form health survey (SF-36) and EuroQol-5D (EQ-5D) are increasingly being used to inform health policy. The last decade has seen the increased use of economic evaluation, particularly the use of cost-effectiveness analyses by agencies such as National Institute for Health and Clinical Excellence (NICE) to inform resource allocation decisions [1], where interventions are assessed in terms of their cost per quality adjusted life year (QALY). The QALY provides a way of measuring the benefits of health care interventions, including improvements in health-related quality of life (HRQL) usually measured using a generic measure like EQ-5D. There has been, however, only a limited use of generic measures of health in mental health [2].

It is claimed that the EQ-5D and other generic preference-based measures such as the SF-6D [3] are applicable to all interventions and patient groups. This claim has support in many physical conditions where these instruments have managed to pass psychometric tests of reliability and validity [4]. For other conditions the claim has not been substantiated, such as in relation to visual impairment in macular degeneration [5] and hearing loss [6]. Doubts have also been raised about the appropriateness of generic measures in mental health [7]. One solution would be to use disease-specific HRQL measures, for example there have been attempts to derive preference-based measures from the positive and negative syndrome scale (PANSS) and clinical outcomes in routine evaluation – outcome measure (CORE-OM) [8,9] in mental health. There are concerns, however, about the comparability of such disease-specific scales and in the United Kingdom, health technology assessment submissions to NICE are expected to follow the details outlined in the reference case analysis described by the NICE methods guide. This clearly stipulates that wherever possible and appropriate, the EQ-5D is the favored measure for generating utility values [1], thus allowing a common metric to assess health care interventions. Alternative measures may be used where the EQ-5D has been empirically demonstrated to be inappropriate in terms of their validity and responsiveness to change and several studies have been undertaken providing such evidence.

In order to provide a reasoned assessment of the appropriateness of generic HRQL measures in patients with schizophrenia, we have under-
Methods

Measures being evaluated

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 [10]. The SF-12 [11] is a shortened version of the SF-36, containing 12 of the 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 provide a position on the EQ-5D health state classification and 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. 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]) and each one has its own preference-based score. Preference-based scores are determined by eliciting preferences: establishing which health states are preferred from a population sample. In order to do so, a method such as trade off is used and 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 [12]. 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. The EQ-5D VAS reports on the respondent’s self-rated valuation of his or her health status; thus, it is based on the preferences of the patient, 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 [3,13]. The SF-6D has a classification that describes health on six multilevel 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. These health state utility values can be used to calculate QALYs for cost-effectiveness analysis.

Inclusion and exclusion criteria

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 old) with schizophrenia or schizophrenia-related disorders (e.g., schizophreniform disorder or schizoaffective disorder). HRQL data could be from descriptive systems (i.e., their items and dimensions), health state utility values generated by the EQ-5D or SF-6D, or the EQ-5D VAS. Studies with the primary focus on individuals with alcohol and/or drug dependency with comorbid schizophrenia or schizophrenia-related disorder were excluded. The outcomes had to include data that allowed measurement of the construct validity (i.e., known groups or convergent) or the responsiveness of the HRQL instrument(s). Responsiveness data had to be in the form of effect sizes, standardized response means (SRMs), or correlation with change scores on symptom measures. Studies that only provided data on other psychometric properties such as reliability, face validity, and content validity were not included.

Identification of studies

As part of a wider review of HRQL measures in mental health funded by the Medical Research Council (MRC), this review focused on the construct validity and responsiveness of the four generic HRQL measures within schizophrenia. Other reviews were carried out, each focusing on one mental health condition, as part of the wider review. A literature search was performed to identify relevant research for all mental health conditions being investigated within the wider review 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 each mental health condition (search strategies are available from authors). Ten databases were searched for published research from inception: 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, and Web of Science. Searches were limited to English language only, but not by date restriction. All searches were conducted in August 2009. The reference lists of relevant studies were searched for further articles.

Citations identified by the searching process were screened by one reviewer (DP) using the inclusion criteria. The full texts of articles 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. The same inclusion and exclusion criteria were used to assess full articles and any queries over inclusion were resolved by discussion and consensus between two reviewers (DP/JB).

Data extraction

Data from all included trials were extracted using a form designed specifically for this review, and piloted on one paper [6]. Data extracted included: country of publication, type of disorder, study sample characteristics (numbers, age, gender), other measures used, mean scores on HRQL measures, type and method of validity assessment, type and method of responsiveness assessment, and validity and responsiveness data. Extractions were performed by one reviewer (DP). Where duplicate publications reported on similar data, the most complete and recent data were extracted.

Quality assessment

There is no formal method for assessing the quality of these studies (i.e., there are no quality assessment checklists). The methods described by Fitzsimmons et al. [14] were used to evaluate HRQL data in their systematic review on the use and validation of HRQL instruments within older cancer patients. This included whether tests of statistical significance were applied, differences between treatment groups were reported (where applicable, e.g., in known groups validity), clinical significance discussed, and missing data were documented. We also report on response and completion rates where these are provided.

Evidence synthesis and meta-analysis

Due to the large degree of heterogeneity between studies (including types of study designs, HRQL instruments, population characteristics 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. All analyses were performed based on the HRQL instrument, with data analysis grouped by type of validity (convergent/discriminant or known groups) or responsiveness measured.

Defining validity and responsiveness

Validity

Construct validity is defined as the extent to which an instrument measures the construct it is designed to measure and in the settings
it is designed for [15,16]. Construct validity can be measured by known or extreme groups where in theory, in two groups who differ in a trait or behavior, one group is expected to score significantly higher or lower compared with the other group [16]. Care must be taken to ensure that the groups are hypothesized to have different scores and for preference-based measures care must be taken to ensure that patients and the general public would have clear preferences for one over the other [17]. Convergent validity assesses the relationship of the instrument of interest to other measures of the same construct to which it should be related [16]. Convergent validity is the correlation between two measures that in theory are associated. Again, the instrument being used to test convergence of EQ-5D and SF-6D must be a good indicator of the trait or behavior, such as another preference-based measure may be hypothesized to be likely to have a strong relationship to preferences. The strength of correlation between the two instruments was calculated using statistical tests (Pearson’s product moment correlation or Spearman’s rank correlation). We have used the following categories for evidence of correlation: >0.6, very strong; ≥0.5 to <0.6, strong; <0.5 to ≥0.3, moderate; and <0.3, weak. Statistical significance was also attached to correlations (P < 0.05).

Responsiveness
Walters [15] defined responsiveness as the extent to which an instrument can detect a clinically significant or practically important change over time. Any change must be perceptible and important to patients, and something that would be valued by the general public. Responsiveness can be measured in a number of ways by effect size statistics [15] standardized in different ways, such as dividing through by the SD at baseline or SD of the change in scores over time (i.e., standardized response means). Within this review, Cohen’s [18] categories for magnitude of effect size were used: ≥0.80, large; <0.80 and ≥0.50, moderate; and 0.30 to <0.50, small.

The application of these psychometric criteria to preference-based measures requires some adaptation [17]. 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 condition-specific 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. Equally, the EQ-5D may fail to reflect clinical differences, but these may not be important to patients. Thus, the tests of construct and convergent validity and responsiveness need to be applied with care.

Results

Study characteristics
The search for studies for the wider review retrieved 4115 unique citations (Fig. 1). Of these, 3849 were excluded at the title and abstract stage and 266 full articles were examined. Another 12 studies were identified through reference list checking. Thirty-three studies were identified that provided data on the validity and/or responsiveness of the EQ-5D, SF-36, SF-12, or SF-6D (Tables 1–3 and Appendix found at doi:10.1016/j.jval.2011.04.006) within individuals diagnosed with schizophrenia, schizophreniform disorder, or schizoaffective disorder. Six studies were undertaken...
<table>
<thead>
<tr>
<th>Study</th>
<th>Population characteristics</th>
<th>Properties measured</th>
<th>Source and types of measures used to test convergent validity and/or responsiveness</th>
<th>Details of validity or responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auquier (2003) [35]</td>
<td>France DSM-IV schizophrenia Inpatients and outpatients (numbers not reported). N=207 (141 males and 66 females). Mean age, 37.5 (SD, 10.9) (range 18–70 years).</td>
<td>Convergent validity</td>
<td>Patient-completed</td>
<td>Correlations with EQ-5D descriptive system health states and SQoL dimensions ranged from 0.06 (SQoL family relationships) to 0.56 (SQoL self-esteem). Generally moderate correlations, overall correlation with S-QOL index was moderate and significant: 0.48, P &lt; 0.05</td>
</tr>
<tr>
<td>Badia (1999) [49]</td>
<td>Country not reported DSM-IV schizophrenia (classification not reported). N=approx 2949 (n=2128 olanzapine; n=821 risperidone or haloperidol; small numbers on other antipsychotics). No age, gender or inpatient/outpatient status reported.</td>
<td>Responsiveness</td>
<td>No measures reported</td>
<td>EQ-VAS and EQ-5D index recorded large effect sizes (0.98 and 1.13, respectively) for olanzapine-treated patients pre- and post-treatment and moderate to large effect sizes for other antipsychotics (0.58 to 0.75 for VAS and 0.78 to 0.96 for index).</td>
</tr>
<tr>
<td>Barton (2009) [41]</td>
<td>UK Non-affective psychosis diagnosis (criteria not specified). Includes: schizophrenia, schizoaffective disorder, bipolar disorder, and psychotic depression. Participants had to screen positive for psychotic symptoms and in relative remission (≤4 on PANSS). N=77 (55 males, 22 females). Mean age, 28.9 years; range 18–52.50/77 had a diagnosis of non-affective psychosis. Inpatient/outpatient status not reported.</td>
<td>Known groups validity. Convergent validity. Responsiveness</td>
<td>Clinician-completed</td>
<td>EQ-5D Index scores showed at least a minimally important clinical difference (MID) (defined as &gt;0.03) between those with milder and more severe scores on symptom and functioning measures. Correlations between the EQ-5D Index and three symptom measures (BAI, BDI, BHS) were moderate to very strong (0.360–0.656). A significant but weak correlation was found with a measure the GAF(0.263). Non-significant and weak correlations were seen with the PANSS, QLS and SOFAS. Mean EQ-5D scores were higher for those who improved than those who did improve on 6 of 7 symptom or functioning measures. The difference in means between improvers and non-improvers was equal to or greater than the MID (0.03).</td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Study</th>
<th>Population characteristics</th>
<th>Properties measured</th>
<th>Source and types of measures used to test convergent validity and/or responsiveness</th>
<th>Details of validity or responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konig (2007) [30] Germany</td>
<td>ICD-10: Schizophrenia, schizotypal or delusional disorders. 49.4% outpatient; 41.6% inpatient; 9.0% day clinic. N=166 (97 males, 69 females). Mean age, 40.5 (SD, 11.1); range 21–80 years.</td>
<td>Convergent validity</td>
<td>Clinician-completed i) Symptoms PANSS, SCL-90R &amp; CGI-S ii) Functional GAF, GARF, SOFAS &amp; HoNOS Patient-completed i) Quality of life-generic TTO direct utility &amp; WHOQOL-BREF</td>
<td>Effect sizes (calculated using the mean values of symptom and functioning measures between individuals who answered “yes” or “no” for each EQ-SD dimensions) were mostly moderate to large for symptom measures (0.37–1.29) and functioning measures (0.24–1.4). Effect sizes for the for the pain/discomfort dimension were smaller. Moderate correlations recorded between EQ-SD VAS and index and symptom measures (0.34–0.73), functioning measures (0.20–0.65), and generic quality of life measures (0.47–0.57).</td>
</tr>
<tr>
<td>Konig (2009) [31] Germany</td>
<td>ICD-10: Schizophrenia, schizotypal or delusional disorders. 51.7% outpatient; 38.5% inpatient, and 9.8 day clinic. N=143 (83 males and 60 females). Mean age, 40.4 (SD, 11.6).</td>
<td>Convergent validity</td>
<td>Clinician-completed i) Symptoms PANSS, SCL-90R, CGI-S, and BRAMES ii) Functional GAF, GARF, SOFAS, and HoNOS Patient-completed i) Quality of life-generic TTO direct utility &amp; WHO-QOL-BREF</td>
<td>Correlation with the TTO direct elicitation of utility values and the EQ-SD VAS and EQ-SD index (UK and German) were weak in correlation (0.25). However, the TTO method did not correlate well with a number of theoretically related measures. Moderate correlation (0.343) with EQ-SD index and a symptom measure (BPRS) at baseline. Weak correlation (0.29) with changes in symptom measure after treatment. Where improvement on BPRS was at least 25%, EQ-SD SRM was small in size (0.39). Where deterioration on BPRS was at least 25% or improvement on BPRS &lt;25%, EQ-5DSRMs were very small (0.17 and 0.05 respectively). EQ-SD index and EQ-SD VAS both demonstrated moderate to strong association with one symptom (CGI-S) and one functional measure (GAF), range 0.34–0.54, P &lt; 0.001.</td>
</tr>
<tr>
<td>McCrone (2009) [20] The Netherlands, Germany, UK, and Italy</td>
<td>SCAN interview diagnosed schizophrenia (classification scheme not specified). “Chronic high disability sample” based on number of years on medication, number of psychiatric inpatient days last year, and GAF score. N=409 (245 males and 164 females). Mean age, 41.5 (SD, 11.5); no range reported.</td>
<td>Convergent validity. Responsiveness</td>
<td>Clinician-completed i) Symptoms CGI-S ii) Functional GAF</td>
<td>Correlation with the TTO direct elicitation of utility values and the EQ-SD VAS and EQ-SD index (UK and German) were weak in correlation (0.25). However, the TTO method did not correlate well with a number of theoretically related measures. Moderate correlation (0.343) with EQ-SD index and a symptom measure (BPRS) at baseline. Weak correlation (0.29) with changes in symptom measure after treatment. Where improvement on BPRS was at least 25%, EQ-SD SRM was small in size (0.39). Where deterioration on BPRS was at least 25% or improvement on BPRS &lt;25%, EQ-5DSRMs were very small (0.17 and 0.05 respectively). EQ-SD index and EQ-SD VAS both demonstrated moderate to strong association with one symptom (CGI-S) and one functional measure (GAF), range 0.34–0.54, P &lt; 0.001.</td>
</tr>
<tr>
<td>Prieto (2004) [38] Spain</td>
<td>ICD-10 Schizophrenia. N=2657 (1691 males and 966 females). Not stated if inpatient or outpatient N=2128 on olanzapine; n=417 on risperidone; n=112 on haloperidol. Mean age, 35.32 (SD, 11.57); range not reported.</td>
<td>Convergent validity</td>
<td>Clinician-completed i) Symptoms CGI-S ii) Functional GAF</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Population characteristics</td>
<td>Properties measured</td>
<td>Source and types of measures used to test convergent validity and/or responsiveness</td>
<td>Details of validity or responsiveness</td>
</tr>
<tr>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>Scalone (2008) [47] Italy</td>
<td>N= 637 (n=551 with schizophrenia n=86 with schizophreniform disorder ). 414 males and 223 females; 18–40 years old (no mean age reported). Inpatient/outpatient status not reported.</td>
<td>Convergent validity</td>
<td>Clinician-completed i) Symptom PANSS, CGI-S ii) Functional GAF</td>
<td>Weak to moderate correlations between QOL scores (EQ-5D and SF-36) and symptom measures (PANSS and CGI-S) ranging from 0.189–0.393.</td>
</tr>
<tr>
<td>van de Willige (2005) [40] The Netherlands</td>
<td>DSM-IV schizophrenia (described as chronic sample). Auditory hallucinations for &gt; 2 years after adequate treatment. Use of at least 2 antipsychotic drugs. Inpatients and outpatients-numbers not reported. N=76 (42 males and 34 females). Mean age, 36 years (SD, 11.2).</td>
<td>Responsiveness</td>
<td>Clinician-completed i) Symptom PANSS, AHRS ii) Functional GSDS iii) Quality of life-generic WHOQOL-BREF</td>
<td>Differences in EQ-5D descriptive system scores between baseline and follow-up were statistically significant for the daily functioning domain (Z=1.79, P &gt; 0.05 &lt; 0.10) and anxiety/depression domain (Z=3.53, P &lt; 0.001). Moderate correlations between changes on EQ-5D VAS and changes in PANSS (total and subscales) (0.34–0.47, P &lt; 0.01 and P &lt; 0.0005). Correlations between changes on EQ-5D index and changes in PANSS existed only on PANSS positive symptoms subscale (0.53, P &lt; 0.001). Moderate to strong correlations with 3 of 4 AHRS subscales and the EQ-5D VA (0.46–0.50, P &lt; 0.001). The EQ-5D index was only correlated with one AHRS subscale and this was weak (distress, 0.25, P &lt; 0.01). Moderate correlations with social function (GSDS) on both the EQ-5D VAS (0.27–0.46, P ranges &lt; 0.01 and &lt; 0.001) and EQ-5D index (0.29–0.39, P ranges &lt; 0.05 and &lt; 0.005). WHOQoL-Bref dimensions correlated for the most part moderately to strongly with the EQ-5D VAS (0.27–0.60) and EQ-5D index (0.25–0.58).</td>
</tr>
</tbody>
</table>

*Note: other measures used in the study, but not used to test convergent validity or responsiveness, are not listed.

AHRs, auditory hallucinations rating scale; BAI, Beck anxiety inventory; BDI, Beck depression inventory; BHS, Beck hopelessness scale; BPRS, brief psychiatry rating scale; BRAMES, Bech–Rafaelsen melancholia scale; CDSS= Calgary depression scale for schizophrenia; CGI-S, clinical global impression-severity; EQ-5D, EuroQol-5D; ESRS, extrapyramidal symptom rating scale; GAF, global assessment of functioning; GARF, global assessment of relational functioning scale; GSDS= Groningen social disabilities schedule; HoNOS, health of the nation outcome scales; PANSS, positive and negative syndrome scale; QLS = quality of life scale; QoLI, quality of life inventory; SCL-90R, symptom checklist-90-R; SF-36, short form health survey; SOFAS, social and occupational functioning assessment scale; S-QOL, schizophrenia quality of life questionnaire; TTO, time trade off; VAS, visual analogue scale; WHO-QOL-BREF= WHO quality of life-BREF.
Table 2 – Summary of evidence for SF-36 by property
(more detailed evidence is presented in the Appendix

<table>
<thead>
<tr>
<th>Property</th>
<th>Number of studies</th>
<th>%</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known groups validity</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Convergent validity</td>
<td>7</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

**KEY**

✓ Evidence suggests property exists (e.g., statistically significant difference in scores for known groups validity or moderate to strong correlations for convergent validity).

? Mixed evidence for property.

X Evidence suggests property does not exist (e.g., weak correlations for convergent validity).

Known groups validity

Barton et al. [41] demonstrated known groups validity for the EQ-5D index whose scores differed according to the severity of disease. Clinically significant differences in EQ-5D index scores (defined as >0.03) were found between individuals defined as “severe” or “less severe” on seven symptoms or functioning measures, which included the PANSS, Hamilton depression rating scale, and global assessment of functioning (GAF).

Convergent validity

**Symptom measures.** Correlation with the EQ-5D and measures of symptoms or symptom severity such as the PANSS, symptom checklist-90-revised (SCL-90R), clinical global impression severity of illness scale (CGI-S), and brief psychiatry rating scale (BPRS) were modest or occasionally strong in three studies [20,30,38]. Two studies, however, found associations with the PANSS measures as nonexistent or mostly weak [41,47]. Moderate to strong associations between EQ-5D index scores and depression or anxiety symptom measures were recorded in one study [41].

Functioning and other quality of life measures. Association with the functioning measure, GAF, was mixed – it was non-existent in one study [41] and moderate or strong in two studies [30,38]. Similar association between the EQ-5D index and the social and occupational functioning assessment scale (SOFAS) was non-existent in one study [41], and it was moderate with UK and German versions of the EQ-5D index (0.44 and 0.42, respectively, P < 0.001) [30]. The EQ-5D index was moderately to strongly associated with the health of the nation outcome scales (HoNOS) and weakly to moderately correlated with the global assessment of relational functioning (GARF); whereas EQ-5D health state scores were mostly moderately to strongly associated with these measures [30].

Most moderate and significant correlations were found between the EQ-5D descriptive system health states score and the schizophrenia quality of life questionnaire (S-QoL) [35]. Barton et al. [41] found no association between the EQ-5D index and another schizophrenia-specific HRQL measure, the quality-of-life scale (QLS) [41]. Konig et al. [30] found no association between the EQ-5D index and direct utilities elicited by the time trade off method.

**Responsiveness**

Responsiveness data were also mixed. Two studies demonstrated that the EQ-5D VAS and EQ-5D index were responsive to change in patients [41,49]. There was weak evidence to suggest that EQ-5D index scores were associated with changes >25% on the BPRS and there was little or no association found when changes on the BPRS were <25% [20]. Furthermore, van de Willige et al. [40] found that EQ-5D index scores did not respond to changes in most symptom or functioning measures, only showing significant (but not always moderate or strong) correlations with the PANSS positive subscale, auditory hallucinations rating scale (AHRS), and Groningen social disabilities schedule (GSDS) (–0.39, P < 0.005). The EQ-5D dimension and VAS scores appeared to perform better than the EQ-5D index in the same study [40].

**SF-36**

Fourteen studies examined the construct validity of the SF-36 using convergent validity [8,21,23,28,32,35–37,39,44,46–48] and 12 studies examined the construct validity of the SF-36 using known groups validity [19,22–25,28,29,32–34,45,46]. Nine studies investigated the responsiveness of the SF-36 [21,26,27,32,34,36,44,48,50]. (See Table 2 for concise version of SF-36 validity and responsiveness evidence and Appendix found at doi:10.1016/j.jval.2011.04.006 for further details on the evidence for the validity and responsiveness of the SF-36.).

---

**EQ-5D**

Seven studies examined the convergent validity of the EQ-5D [20,30,31,35,38,41,47] and one study examined the construct validity of the EQ-5D by the known groups method [41] (Table 1). Four studies investigated the responsiveness of the EQ-5D [20,40,41,49]. Seven studies investigating the EQ-5D used population preferences to generate an index value [20,30,31,38,40,41,49].
<table>
<thead>
<tr>
<th>Study</th>
<th>Patient characteristics</th>
<th>Properties measured</th>
<th>Source and types of measures used to test convergent validity and/or responsiveness/</th>
<th>Details of validity or responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-12 validity and responsiveness</td>
<td>DSM-IV psychosis (not defined). 50 participants (male/female not provided). No mean age or range reported. Inpatient/outpatient status not reported.</td>
<td>Known groups validity</td>
<td>Not applicable</td>
<td>Linear regression demonstrated that individuals with psychosis were significantly ( P &lt; 0.001 ) more likely to report disability on the SF-12 than individuals with no mental health disorder. SF-12 scores were around 12 points lower in individuals with psychosis.</td>
</tr>
<tr>
<td>SF-6D validity and responsiveness</td>
<td>SCAN interview diagnosed schizophrenia (classification scheme not specified). “Chronic high disability sample” based on number of years on medication, number of psychiatric inpatient days last year, and GAF score. N=409 (245 males and 164 females). Mean age, 41.5 (SD, 11.5); no range reported.</td>
<td>Convergent validity. Responsiveness</td>
<td>Clinician-completed ii) Symptoms BPRS</td>
<td>Moderate correlation (0.314) with a symptom measure (BPRS) at baseline. Weak correlation (0.22) with changes in symptom measure after treatment. Where improvement on BPRS was at least 25%, SRM was moderate in size (0.39). Where deterioration on BPRS was at least 25% or improvement on BPRS &lt;25%, SRM was very small (0.27 and 0.02, respectively).</td>
</tr>
</tbody>
</table>

BPRS, brief psychiatry rating scale; GAF, global assessment of functioning; SF-6D, short form 6D (preference-based) generated from items of the SF-36 or SF-12; SF-12, short form 12 (shortened SF-36); SRM, standardized response mean.
Table 4 – Quality assessment of included studies.

<table>
<thead>
<tr>
<th>Study details</th>
<th>Properties measured</th>
<th>Statistical significance tested for properties measured</th>
<th>Difference between groups</th>
<th>Clinical significance addressed or discussed</th>
<th>Missing HRQL data documented *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auquier (2003) [35]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported (for SF-36)</td>
</tr>
<tr>
<td>Badia (1999) [49]</td>
<td>Responsiveness</td>
<td>Not reported</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Barton, G. (2009) [41]</td>
<td>Known groups and convergent validity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Partly – numbers presented for each analysis which demonstrate some non-completion, but no detail on EQ-5D completion.</td>
</tr>
<tr>
<td>Bebbington (2009) [19]</td>
<td>Known groups validity</td>
<td>Yes</td>
<td>Not reported but demographics adjusted for in analysis</td>
<td>Not reported</td>
<td>Partly – SF-36 domains were scored if participants completed 50% of a domain. Numbers varied between dimensions. However, we are not told how complete each dimension is.</td>
</tr>
<tr>
<td>Bobes (1997) [37]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Dunayevich (2007) [50]</td>
<td>Responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Folsom (2009) [24]</td>
<td>Known groups validity</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Jarema (2001) [48]</td>
<td>Convergent validity Responsiveness</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Kebede (2004) [33]</td>
<td>Known groups validity</td>
<td>Not reported</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Kebede (2005) [34]</td>
<td>Known groups and convergent validity.</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Konig (2007) [30]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Partly – states some missing values for some variables and such patients are excluded. Does not state what EQ-5D values are missing.</td>
</tr>
<tr>
<td>Konig (2009) [31]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Partly – states some missing values for some variables and such patients are excluded. Does not state what EQ-5D values are missing.</td>
</tr>
<tr>
<td>Law (2005) [46]</td>
<td>Known groups and convergent validity</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Lenert (2005) [8]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>McCrone (2009) [20]</td>
<td>Convergent validity and responsiveness</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Meijer (2002) [39]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Yes</td>
</tr>
<tr>
<td>Milliken (2007) [44]</td>
<td>Convergent validity and responsiveness</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Nasrallah (2004) [25]</td>
<td>Known groups validity and responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Norholm (2007) [45]</td>
<td>Known groups validity</td>
<td>Yes</td>
<td>Not reported but age-matched sample used to compare scores</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Study details</th>
<th>Properties measured</th>
<th>Statistical significance tested for properties measured</th>
<th>Difference between groups</th>
<th>Clinical significance addressed or discussed</th>
<th>Missing HRQL data documented *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phillips (2006) [26]</td>
<td>Convergent validity and responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Prieto (2004) [38]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported but ceiling effects discussed</td>
<td>Not reported</td>
</tr>
<tr>
<td>Pukrop (2003) [32]</td>
<td>Known groups and convergent validity and responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Pyne (2003) [27]</td>
<td>Responsiveness</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Reine (2005) [36]</td>
<td>Convergent validity and responsiveness</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Revicki (1999) [21]</td>
<td>Convergent validity and responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Russo (1998) [28]</td>
<td>Known groups and convergent validity</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Sanderson (2002) [43]</td>
<td>Known groups validity</td>
<td>Not reported</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Scalone (2008) [47]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Sciolla (2003) [29]</td>
<td>Known groups validity</td>
<td>Yes</td>
<td>Yes</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Strakowski (2005) [22]</td>
<td>Known groups validity</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
<tr>
<td>Tunis (1999) [23]</td>
<td>Known groups and convergent validity and responsiveness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Partly – missing SF-36 values were mentioned by authors but actual percentages were not reported.</td>
</tr>
<tr>
<td>van de Willige (2005) [40]</td>
<td>Responsiveness</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Yes</td>
<td>Not reported</td>
</tr>
<tr>
<td>Wilkinson (2000) [42]</td>
<td>Convergent validity</td>
<td>Yes</td>
<td>Not applicable</td>
<td>Not reported</td>
<td>Not reported</td>
</tr>
</tbody>
</table>

* Actual missing values from instrument NOT lost to follow-up.
Known groups validity
Eleven studies compared SF-36 scores with normative values. Normative values were taken mostly from general population samples and published figures, although some studies recruited a sample of “normal participants” to compare SF-36 scores [24,29,32]. Almost all studies found statistically significant differences in SF-36 summary (PCS and MCS) and dimension scores between individuals with schizophrenia and normative values; this could be up to 80 points in difference on the MCS and its dimensions and up to 50 points in difference on the PCS and its dimensions. Two exceptions were Sciolia et al. [29] and Norholm et al. [45], where statistically significant differences were noted for all dimensions except bodily pain.

One study investigated the effect of the presence of side effects on SF-36 scores. Scores were between two and five points lower on the PCS and MCS for individuals with some side effects (e.g., subjective rigidity or anticholinergic effect) when compared with those who did not have those side effects; these differences were statistically significant [19]. This was not the case, however, for all side effects; for example the MCS and PCS did not differ between participants presenting with subjective akathisia and weight gain (among others).

Convergent validity
Symptom measures. Five studies found mostly weak or non-existent correlations with symptom measures such as the PANSS, Scale for the Assessment of Negative Symptoms (SANS), Extrapyramidal Symptom Rating Scale (ESRS), BPRS, and CGI-S [23,32,36,46,47]. There was some evidence of stronger association with the PANSS in two studies [44,48] and the BPRS in another study [28]. Correlations with measures of depression such as the Montgomery-Åsberg depression rating Scale (MADRS) and Calgary depression scale for schizophrenia (CDSS) were weak in two studies [28,46] and moderate to strong in another study [23].

Functioning and other quality-of-life measures. Correlation with the GAF was recorded as weak to moderate in one study [39] and very strong in another study [44]. The SOFAS was correlated very strongly with the SF-36 [44]. Strong and statistically significant (P < 0.001) correlations were reported with two schizophrenia-specific HRQL measures [35,42]. Revicki et al. [21] described very weak correlations between the SF-36 and schizophrenia-specific QLS total score. Correlations with generic HRQL measures like the World Health Organization quality of life instruments (WHOQoL-BREF, WHOQoL-100, and WHOQoL-26) were mostly moderate to very strong (i.e., >0.3) [37,46].

Responsiveness
Little evidence existed to demonstrate that when changes were recorded on the PANSS, this correlated with changes on the SF-36, with the association being mostly weak and nonsignificant in four studies [26,27,48,50]. Pyne et al. [27] also found weak correlations with changes on the CDSS (–0.27, P < 0.01) and the extrapyramidal symptoms rating scale (ESRS) (–0.22, P < 0.05).

Responsiveness was also measured with other measures or by methods other than calculating correlation between change scores, but similarly this evidence was weak. Effect sizes calculated for patients judged to have improved or not improved according to CGI-S scores were all nonsignificant apart from for social functioning, which was small in size [36]. Milliken et al. [44] found higher MCS scores in remitted versus nonremitted participants, but this was only a trend and not statistically significant (P = 0.063). Revicki et al. [21] reported that the total MCS indicated statistically significant contributions for changes in the PANSS positive scale and the MADRS. Although Pukrop et al. [32] found that improvement in negative symptoms significantly impacted the role physical and role emotional dimensions (and also remained significant when controlling for improvement in negative symptoms), no such interactions remained significant for any dimensions when controlling for improvement in positive symptoms. However, Kebede et al. [34] found that SANS and Scale for the Assessment of Positive Symptoms (SAPS) scores were inversely related with improvements in physical and social functioning domains and role limitations due to emotional problems.

SF-12
Data were limited to one study containing known groups validity evidence, and revealed that individuals with psychosis were significantly (P < 0.001) more likely to report disability on the SF-12 than individuals with no mental health disorder [43] (Table 3).

SF-6D
Data were limited to one study that demonstrated moderate correlation between the SF-6D index and the symptom measure BPRS (–0.344, no P value) [20] (Table 3). When changes occurred on the BPRS, however, changes in the SF-6D were correlated only weakly (–0.22, no P value) and appeared only able to respond to changes on the BPRS greater than 25%. Data for the SF-6D scores were normally distributed, thus there was no evidence for floor or ceiling effects.

Distributional properties of the measures
Only five studies reported distributional properties of the measures: three for the EQ-5D [20,30,38]; and one study each for the SF-36 [23] and SF-6D [20]. Scores were found to be normally distributed for the SF-36 [23] and SF-6D [20]; thus, there was no evidence of floor or ceiling effects. The three studies which report on the distributional properties of the EQ-5D [20,30,38], however, found that the EQ-5D index showed a moderate ceiling effect (for example, Konig et al. reported 21% of respondents achieved the maximum score) [30]. This ceiling effect could potentially limit the responsiveness of the measure. In contrast, two of the three studies found that the EQ-5D VAS was normally distributed [30,38].

Discussion
Thirty-three studies were identified that examined the validity and/or responsiveness of four generic HRQL measures, although very limited data were found for the generic health status measure SF-12 and the preference-based SF-6D. The studies were undertaken in a variety of countries, mostly in Europe and North America, illustrating the wide use of such measures internationally.

The majority of the evidence (25 studies) examined the validity and responsiveness of the SF-36. Although there appears to be strong evidence that the SF-36 is able to distinguish between general population norms and scores of people with schizophrenia (known groups validity), the evidence for convergent validity and responsiveness is less certain. Similar findings existed for the EQ-5D, with mixed evidence for the properties of convergent validity and responsiveness. Indeed, when strong associations were found between individual EQ-5D health state dimensions (e.g., anxiety/depression or self-care) and symptom or functioning measures, this did not necessarily translate into comparable changes in overall EQ-5D index scores such as utility values [30,40]. For psychiatric research, it may be that the physical health domains are overly stressed and with less emphasis on mental health, the total EQ-5D index scores may not be accurately represented [40].

There was some evidence that associations with measures of depression were comparatively stronger than those with symptom measures of schizophrenia (e.g., PANSS) [23,30,36,41].
may indicate that the generic HRQL measures were only able to
detect the component of HRQL or that depression is the only com-
ponent of HRQL within schizophrenia that is important within the
context of HRQL measurement. The issue is whether schizophrenia
has quality-of-life implications not adequately described by the
five dimensions of the EQ-5D. This is an important issue that
needs to be explored further using a range of research methods,
including qualitative interviews with patients.

Types of measures
When testing association between measures for convergent valid-
ity (or change scores in responsiveness), there are good reasons to
predict that stronger and more consistent correlations might exist
between generic HRQL measures and functioning (e.g., GAF, SO-
FAS) or mental health/schizophrenia-specific HRQL (e.g., QLS)
measures than purely symptom-based measures such as the
PANSS. These types of measures are more likely to measure simi-
lar concepts to that of generic HRQL measures and due to this
degree of overlap, we could reasonably assume that these mea-
sures would correlate well with generic HRQL measures. By their
very nature, symptom measures are measuring different concepts
to HRQL measures, so it might be reasonable to predict that it is
less likely that a strong correlation might exist. Similarly, one
might expect a greater degree of association between subjective
measures (completed by patients) and generic HRQL measures
than with objective symptom measures (typically completed by
clinicians).

Re-examining the evidence accounting for the type of measure
used to assess convergent validity (symptom vs. functioning or
HRQL measures; subjective vs. objective measures), for whichever
type of measure the evidence for convergent validity remains
uncertain in this population. Ten studies suggested no or uncer-
tain evidence for a correlation between symptom measures and
generic HRQL measures [8,20,23,32,36,41,46–48] whereas four re-
vealed moderate to strong correlations [28,30,41,44]. Functioning
and schizophrenia HRQL measures did not fare much better, with
four studies indicating strong evidence for convergent validity
[35,38,42,44] and four describing uncertain or no evidence of such
a relationship [21,36,39,41]. Of the seven studies that used objec-
tive measures to test an association, four reported a strong evi-
dence for convergent validity [35,37,42,46] and three found no
such evidence [30,31,41].

Thus, it seems there is a wider issue regarding what types of
measures might reasonably be expected to correlate strongly with
generic HRQL measures. It is difficult to determine how strongly
correlated in theory generic HRQL measures should be with symp-
tom and/or other measures and there is little guidance on what
constitutes reasonable correlation. Indeed, Walters [15] noted that
some would say it is impossible to prove validity of HRQL instru-
ments because no “gold standard” exists. Although a number of
different concepts or constructs will be the same or similar be-
tween HRQL and other measures, there will of course be some
areas where there is no overlap. Also, as discussed previously,
where health dimensions and changes appear to have been
missed by preference-based HRQL measures, these may not actu-
ally be important to patients or valued by the general population;
thus it cannot be determined as a weakness of the measure. This
needs to be explored in further research.

Strengths and limitations
This review comprehensively identified studies that reported on
the construct validity and responsiveness of four generic HRQL
measures (SF-6D, SF-12, SF-36, and EQ-5D), and then tabulated and
provided a narrative synthesis of the findings. The review has
some limitations, mainly due to compromising on some elements
of the review process due to the large scope of the project. The
search for studies was reasonably comprehensive, but it was lim-
ited to key databases and reference list checking of included stud-
ies, and study selection was undertaken by one reviewer. Ideally,

Further research
There is very limited evidence of validity or responsiveness for the
SF-12 and SF-6D and, though they are derivatives of the SF-36, they
have a limited item coverage (12 and 11, respectively) and may not
perform as well. Therefore, further research needs to be directed
toward demonstrating these properties for these instruments.

In conclusion, the evidence found in this review on the validity and
responsiveness of these four measures in this population and pro-
vides a starting point for future more focused reviews and future
primary research.

Conclusion
In conclusion, the evidence found in this review on the validity and
responsiveness for a number of widely used generic measures in
patients with schizophrenia has been mixed. Although the evidence
base is limited in a number of important respects (including prob-
lems with the measures used to develop constructs in the validation
studies), it is sufficient to raise doubts about the use of generic mea-

VALUE IN HEALTH 14 (2011) 907–920
sures of health like the EQ-5D and SF-36 in patients with schizophrenia. This suggests that agencies, such as NICE, which advise on reimbursement of health costs, should be willing to consider evidence on health state utility values based on other methods.

Source of financial support: This study was funded by the Medical Research Council Methodology Board (project number G0801394).

**Supplemental Materials**

Supplemental material accompanying this article can be found in the online version as a hyperlink at doi:10.1016/j.jval.2011.04.006, or by hard copy of article, at www.valueinhjournal.com/issues (select volume, issue, and article).

**References**


