Patient-reported Health Instruments Group

A STRUCTURED REVIEW OF PATIENT-REPORTED MEASURES IN RELATION TO SELECTED CHRONIC CONDITIONS, PERCEPTIONS OF QUALITY OF CARE AND CARER IMPACT

Report to the Department of Health
November 2006

National Centre for Health Outcomes Development

health outcome indicators

A STRUCTURED REVIEW OF PATIENT-REPORTED MEASURES IN RELATION TO SELECTED CHRONIC CONDITIONS, PERCEPTIONS OF QUALITY OF CARE AND CARER IMPACT

Ray Fitzpatrick
Ann Bowling
Elizabeth Gibbons
Kirstie Haywood
Crispin Jenkinson
Anne Mackintosh
Michele Peters

Patient-reported Health Instruments Group
National Centre for Health Outcomes Development (Oxford site)
Unit of Health-Care Epidemiology
Department of Public Health
University of Oxford

November 2006

Copies of this report can be obtained from:

Elizabeth Gibbons
Research Officer
Patient-reported Health Instruments Group
National Centre for Health Outcomes Development (Oxford site

Unit of Health-Care Epidemiology

Department of Public Health

University of Oxford
Old Road
Headington
Oxford OX3 7LF

tel: +44 (0)1865 289402
e-mail: elizabeth.gibbons@dphpc.ox.ac.uk

Alternatively, it can be downloaded free of charge from the PHIG website:

http://phi.uhce.ox.ac.uk/

CONTENTS

EXECUTIVE SUMMARY

Aims of the report

The aims of the set of reviews included in this report are to identify self-reported health instruments (both generic and disease-specific) for use in asthma, Chronic Obstructive Pulmonary Disease (COPD), diabetes, epilepsy, heart failure and stroke; to assess evidence relating to the development and evaluation of these instruments and make recommendations wherever possible about the most appropriate instruments for use in the NHS in relation to groups rather than individuals (for example, for the purposes of audit, quality assurance, evaluation and research). An addition aim is to carry out two reviews with accompanying recommendations in relation to instruments (i) to assess patients' perceptions of quality of care and (ii) carer impact, both reviews focusing exclusively on long-term conditions.

Introduction (Chapter 1)

The context for the reviews is briefly explained. A range of policy initiatives have resulted in patients and the public being more central to the ways in which services are developed and delivered. One important means of increasing patient and public involvement is through patient-reported health instruments. The proliferation of such instruments requires that complex considerations are involved in the choice of instrument. In particular the evidence for measurement properties and feasibility of use of instruments need to be considered. A set of criteria for assessing evidence regarding instruments is described. It should be pointed out that no advice exists in the literature as to how weigh conflicting and contrasting evidence for different properties of instruments, so some considerable judgment is required in overall assessment of the evidence to determine overall comparative performance of instruments. 'Appropriateness' is one of the key criteria. This criterion has to rely on users' judgements of the degree of fit of the content of an instrument to a specific intended application; something that, a priori, cannot be determined solely by reviewing evidence of formal measurement properties. Users have to judge the degree of fit of the content of an instrument to any specific given application and context. The evidence of this review is intended to support and complement such judgements of appropriateness.

Methods (Chapter 2)

A search strategy was designed to retrieve references relating to the eight reviews included in this report. Where appropriate the strategy was based on searches of a bibliography of over 12,000 records relating to published instrument evaluations developed by the National Centre for Health Outcomes Development (NCHOD) at the University of Oxford and publicly available as the Patient-reported Health Instruments (PHI) website (http://phi.uhce.ox.ac.uk/). Additional searching was carried out by hand searching of relevant journals and searches from reference lists of included articles. Included articles were abstracted and assessed according to a standard protocol. A total of 398 studies were included in the review.

Generic instruments (Chapter 3)

To avoid duplication of material across chapters, chapter 3 provides a description of the generic instruments considered in any of the reviews.

Asthma (Chapter 4)

Fifty articles were found that provided useful information on measurement properties for the generic and asthma-specific instruments included in the review. Five generic instruments were identified which were evaluated with patients with asthma: SF-36, SF-12, EuroQol -EQ-5D, Sickness Impact Profile and the Health Utilities Index. Nine asthma-specific instruments were included in the review: The Juniper collection of Asthma Quality of Life Questionnaires: AQLQ, MiniAQLQ, AQLQ(S), Acute AQLQ, ACQ, and the ACD; The Marks Asthma Quality of Life Questionnaire (MAQLQ), Living With Asthma Questionnaire and the St. George's Respiratory Questionnaire.

Recommendations

The SF-36 is recommended as a generic instrument for the broad evaluation of health-related quality of life for people with asthma.

Among asthma-specific instruments, particularly the AQLQ Juniper collection and the MAQLQ are recommended, with different versions of the AQLQ instruments selected for particular purposes.

Chronic obstructive pulmonary disease (Chapter 5)

Forty-six articles provided useful evidence of measurement properties of the instruments included in the review for people with COPD. Seven generic instruments were identified that had been evaluated with people with COPD: SF-36; SF-20; SF-12; EuroQol -EQ-5D; Sickness Impact Profile; Dartmouth COOP and Nottingham Health Profile. Five COPD patient-reported health instruments were included in the review: Breathing Problems Questionnaire, Chronic Respiratory Questionnaire, Functional Performance Inventory, Seattle Obstructive Lung Disease Questionnaire and the St. George's Respiratory Questionnaire.

Recommendations

The SF-36 is recommended as a generic instrument for the broad evaluation of health-related quality of life for people with COPD.

Among COPD-specific instruments, particularly the CRQ and SGRQ are recommended.

Diabetes (Chapter 6)

Ninety-one articles provided useful evidence of the measurement properties of the instruments included in the review. Six generic instruments were identified which were evaluated with patients with diabetes: SF-36; SF-12; Sickness Impact Profile;

Health Utilities Index; Quality of Well-Being Scale and the EuroQol -EQ-5D. Six diabetes-specific instruments were assessed: Appraisal of Diabetes Scale/ADS; Audit of Diabetes-Dependent Quality of Life/ADDQoL; Diabetes 39/D-39; Diabetes Health Profile/DHP; Diabetes Quality of Life Measure/DQOL and the Diabetes Quality of Life Clinical Trial Questionnaire/DQLCTQ.

Recommendations

Of generic instruments, the SF-36 is recommended.

There is insufficient positive evidence strongly to single out any particular disease-specific instrument in diabetes. Of the large number of such instruments, ADQOL, DHP and DQOL may warrant more attention to establish the case for a disease-specific instrument.

Epilepsy (Chapter 7)

Seventy-one articles provided useful evidence of measurement properties of the instruments included in the review. Seven generic instruments were identified which were have been assessed for use with epilepsy: SF-36; SF-12; EuroQol -EQ-5D; Health Utilities Index; Q Twist; Nottingham Health Profile and the Sickness Impact Profile. Eight epilepsy-specific instruments were identified which were evaluated with patients with epilepsy: Epilepsy Surgery Inventory-55; Katz Adjustment Scale; Liverpool Quality of Life (LQOL) Battery and Seizure Severity Scale; Quality of Life in Epilepsy-89 (QOLIE-89); Quality of Life in Epilepsy-31; Quality of Life in Epilepsy-10; Side Effects and Life Satisfaction (SEALS) Inventory; and the Washington Psychosocial Seizure Inventory.

Recommendations

The SF-36 is recommended as a generic instrument for use with patients with epilepsy.

Of epilepsy-specific instruments, the ESI-55 and the QOLIE-89 are recommended although the ESI-55 needs testing to be used outside of the specific surgical context.

Heart failure (Chapter 8)

Eighty-nine articles provided evidence of measurement properties of the instruments included in the review. Four generic instruments were identified which were evaluated with patients with Heart Failure: SF-36; SF-12; Sickness Impact Profile; EuroQol -EQ-5D. Four heart failure-specific instruments were identified which were evaluated with patients with various cardiovascular conditions resulting in heart failure: Chronic Heart Failure Questionnaire; Kansas City Cardiomyopathy Questionnaire; MacNew (ex QLMI – Quality of Life after Myocardial Infarction Questionnaire); and the Minnesota Living with Heart Failure Questionnaire.

Recommendations

This review supports the use of the SF-36 and the SF-12 as generic instruments.

The MLHFQ is recommended as a heart-failure specific instrument.
However, some possible remaining problems may exist with wording and content validity due to the narrow focus of the instrument.

Stroke (Chapter 9)

A total of 48 articles provided useful evidence regarding the instruments included in the review. Six generic instruments were identified in the review which had been evaluated with people who have experienced a stroke: SF-36, SF-12, SF-6D, EuroQol- EQ-5D, Health Utilities Index, and the Nottingham Health Profile. Seven stroke-specific instrument were identified which had been evaluated with patients with stroke: Stroke Impact Scale (SIS); Stroke Specific Quality of Life Scale (SS-QOL); Subjective Index of Physical and social Outcomes (SIPSO); Barthel Index; Frenchay Activities Index; Nottingham Extended ADL scale; and the London Handicap Scale.

Recommendations

Overall, the SF-36 is recommended. There is evidence to support the EuroQol EQ-5D as a brief, reasonably well acceptable measure of general health in stroke, although the evidence is more limited.

At the present stage of development no single multi-dimensional health instrument has sufficient information available to justify recommendation. Both the SIS and the SIPSO seem highly promising but further evidence is required for both measures. It seems that, at least for the time being, interview and self completion versions of the Barthel Index, Frenchay Activities Index and Nottingham Extended ADL Scale would appear the most appropriate condition-specific instruments

Carer impact (Chapter 10)

A total of seventy-five articles were included which reported instruments to assess experiences of those who care for individuals with long-term conditions.

Six generic health instruments were included which had been evaluated with carers: SF-36, SF-12, GHQ, Health Utilities Index Mark 2 (HUI2), Reintegration to Normal Living Index and the Ferrans and Power Quality of Life Index. Such measures provide indirect evidence of the experiences of carers by assessing broad aspects of health that may be related to caring.

Seven general carer instruments (providing direct evidence through their focus on questions about the caring experience) have evidence of measurement properties from multiple evaluations with carers: Appraisal of Caregiving Scale (ACS) (2 evaluations), Bakas Caregiver Outcomes Scale (BCOS) (3), Caregiver Burden Inventory (CBI) (3), Caregiving Appraisal Scale (CAS) (2), Caregiving Impact Scale (CIS) (2), CSI (9), Caregiver Well-Being Scale (3) and Zarit Burden Interview (ZBI) (12).

Recommendations

Two generic health status instruments, SF-36 and GHQ, can be used to provide indirect evidence of carer impact. The CSI and ZBI provide more direct evidence of carer impact, with the CSI being somewhat more supported for use in the format of self complete questionnaire.

It is not possible currently to make definite recommendations for instruments to be used to investigate carer impact. Nevertheless the combined use of generic and general carer instruments may be a sensible strategy.

Patient perceptions of quality of healthcare (Chapter 11)

A total of 13 patient-reported measures of health care quality, of relevance to chronic disease were included for more detailed assessment in the review.

Recommendations

No single measure was considered appropriate to recommend. However the following had some desirable features that are described:

- Patient Assessment of Chronic Illness Care (PACIC)
- The Picker Institute questionnaires
- The OutPatient Experience Questionnaire (OPEQ)
- The QUOTE measures
- The Health care System Hassles Questionnaire (HSHQ)
- EORTC IN-PATSAT32. Although specific to the evaluation of health care for in-patients (receiving medical and/or surgical care) with cancer, it is commended for the extensive involvement of patients and health professionals, across a wide range of cultural settings, in item development and subsequent evaluation.

More generally, although no single measure is unambiguously to be recommended, it is clear that there is a growing consensus of topics, domains and methods that should be included in any assessment of the views regarding quality of care of individuals with long-term conditions.

Chapter 1: INTRODUCTION

Several important government policies and initiatives form the context for this set of reviews. The NHS Plan set in place a process of reforms to develop services designed around the patient (NHS Plan, 2000). The NHS Improvement Plan (Secretary of State for Health, 2004) placed special emphasis on the need to develop services more appropriate for individuals with long-term conditions. More recently, the White Paper, Our Health, Our Care, Our Say (DH, 2006) sets out plans to make health and social services more responsive to patients' and users' needs, choices and preferences. An enormous range of developments have been planned to allow individuals with long-term conditions to avoid unnecessary hospitalisation, reduce dependence on the acute care model for services, maintain independence in the community, promote self care and control over their lives and services and reduce disabilities and disadvantages arising from chronic illnesses. All of these ambitious plans require evidence directly from patients and the public that services are having a positive impact in relation to their experience of long-term conditions and of services. With around 6 in 10 adults reporting some form of chronic condition and these individuals making greater use than others of health and social services, there is enormous scope for evidence of patients' and users' experiences to make a difference to the quality of care and to the quality of lives of those with long-term conditions.

At the same time there has been growing recognition of the real absence of evidence outputs and outcomes of public services generally and the NHS specifically. Traditionally the productivity of services has been measured by indicators that might be better thought of as inputs, numbers of procedures carried out, numbers of consultations and admissions etc. The Office of National Statistics commissioned a review of public service performance and productivity that highlighted this lack of evidence and called for the development and use of better measures of outcome to inform decisions about the productivity of public services (Atkinson, 2005). The scope for patients and users directly to report judgements of outcome in relation to health services was emphasised.

The enormous array of patient-reported outcome measures that have been developed over the last thirty years offers clear opportunities to involve patients and users directly in judgements of the outcomes of services. Variously termed measures of 'health status', 'health-related quality of life', 'functional status', 'patient-reported outcome' or often just 'outcome', the common element is an attempt directly to capture the patient's experience of important aspects of health through questionnaire or interview. Considerable resources and effort have been invested to make such 'instruments' valid measures for use in relation to a wide range of decisions and policies in health. One principle problem is that there are large numbers of such instruments from which to choose for any given health problem or context and insufficient guidance to inform choice (Garratt et al., 2002).

Patient-reported health instruments usually take the form of questionnaires containing several items reflecting the broad nature of health status, disease, or injury, which are most often summed to give a total score. The term 'patient-reported health instrument' will be used throughout this review to refer to patient-completed instruments.

There are two broad categories of patient-reported health instrument: generic and specific. Generic instruments are not age-, disease-, or treatment-specific and contain multiple concepts intended to be relevant to a wide range of patients and the general population. Specific instruments may be specific to a particular disease (for example, diabetes), a patient population (for example, older people), a specific problem (for example, pain), or a described function (for example, activities of daily living). Disease-specific instruments may have greater clinical appeal due to their specificity of content, and associated increased responsiveness to specific changes in condition.

The broad content of generic instruments enables the identification of co-morbid features and treatment side-effects that may not be captured by specific instruments, which suggests they may be useful in assessing the impact of new health-care technologies where the therapeutic effects are uncertain. However, the broad content may reduce responsiveness to small but important changes. It has therefore been recommended that a combination of generic and specific measures be used in the assessment of health outcomes.

Patient-reported health instruments have been increasingly applied in a range of settings including routine patient care, clinical research, audit and quality assurance, population surveys, and resource allocation. However, consensus is often lacking as to which instrument to use; this has important implications for the evaluation of clinical effectiveness. Structured reviews of measurement properties are a prerequisite for instrument selection and standardisation, and instruments with measurement properties that support their application in specific populations and across a range of evaluation settings need to be identified.

Selection criteria have been defined for assessing the quality of patient-reported health instruments (Streiner and Norman, 1995; McDowell and Newell, 1996; Fitzpatrick et al., 1998). These include measurement issues, such as reliability, validity, responsiveness, and precision, as well as practical issues, such as acceptability and feasibility. These criteria are briefly summarised since they directly inform the reviews reported here.

Criteria for assessing patient-reported health instruments

Reliability is concerned with whether measurement is accurate over time and, for multi-item instruments, whether they are internally consistent. Test-retest reliability usually involves instrument self-completion on two occasions separated by a suitable time-period and, assuming no change in the underlying health state, measures the temporal stability of the score (Fitzpatrick et al., 1998). A test-retest period of between two days and two weeks has been recommended for most conditions (Streiner and Norman, 1995). Too short a period may be associated with patient recall of answers, which may artificially inflate reliability (Nunnally and Bernstein, 1994; Streiner and Norman, 1995); too long a period may be associated with actual change in health.

Health transition questions, which invite patients to indicate whether their general or specific health has changed between instrument administrations, are often included in evaluations. The correlation coefficient is the most frequently used method for calculating estimates of test-retest reliability; the intra-class correlation coefficient

(ICC) is used to identify group shift over time as a measure of reliability (Streiner and Norman, 1995). For group comparisons, levels of reliability over 0.70 are required (Streiner and Norman, 1995; Fitzpatrick et al., 1998). For the evaluation of individuals, levels above 0.90 have been recommended (Nunnally and Bernstein, 1994; Fitzpatrick et al., 1998).

Internal consistency reliability of multi-item instruments that adopt a traditional summated rating scale format is tested following a single application. The relationship between all items, and their ability to measure a single underlying domain is assessed using Cronbach's alpha: alpha levels of between 0.70 and 0.90 have been recommended (Nunnally and Bernstein, 1994; Streiner and Norman, 1995; Garratt et al., 2001). Homogeneity at the item level can be assessed using item-total correlation: levels above 0.40 have been recommended (Ware, 1997).

Validity assesses whether an instrument measures what is intended in the different settings in which it may be applied (McHorney, 1996; Fitzpatrick et al., 1998). Instrument validity is not a fixed property. The process of validity testing is ongoing, informing instrument application and interpretation in different settings and with different populations (McHorney, 1996; Ware, 1997). Hence, new and refined instruments, and those applied in different settings or with different populations require evidence of validity. Both qualitative and quantitative methods can be used to assess validity.

Face and content validity require appraisal of item content, and assessment of its relationship to the instrument's proposed purpose and application (Fitzpatrick et al., 1998). Methods of item generation and instrument development may influence this assessment. Literature reviews, theoretical propositions, and interviews or focus groups with patients or health-care professionals may all inform this process. However, for patient-reported instruments to have content validity and relevance to the recipients of care, patients should be involved in item derivation (Fitzpatrick et al., 1998).

The quantitative assessment of validity requires comparison of the scores produced using patient-reported health instruments with those derived from other measures of health, clinical, and socio-demographic variables. Patient-reported instruments measure hypothetical constructs which are by definition non-observable, for example, HRQL and pain, and address a more general hypothesis than that supported by a specific behaviour (Nunnally and Bernstein, 1994). However, by reference to established evidence and the instrument's underlying theoretical base and item content, quantifiable relationships with a range of other instruments and clinical and socio-demographic variables can be expected (Ware, 1997; Fitzpatrick et al., 1998).

Expected correlations between variables should be presented to allow validity to be disproved (McDowell and Jenkinson, 1996). The strength of correlation between variables, be they small (less than 0.30), moderate (less than 0.50), or large (greater than 0.70), indicates that the instrument measures the construct in a manner founded on theory or established evidence (McHorney et al., 1993). For example, two patient-reported measures of functional disability with similar content would be expected to correlate strongly. Construct validity may also be assessed using 'extreme groups', which theorises that one group will possess more or less of a construct (Streiner and

Norman, 1995). For example, compared to the general older population, older people who are hospitalised following a hip fracture may be expected to report greater pain and worse HRQL.

The dimensionality or internal construct validity of a multi-item instrument can be assessed using factor analysis or principal component analysis. Principal component analysis can be used to assess the underlying structure of a multi-item instrument through the identification of components, or domains, into which items may group (McDowell and Newell, 1996). This form of analysis adds empirical weight to a hypothesised domain structure. For example, principal component analysis has supported the hypothesised eight-domain structure of the SF-36 (McHorney et al., 1993).

Responsiveness is considered a necessary measurement property of instruments intended for application in evaluative studies measuring longitudinal changes in health (Beaton et al., 2001; Liang et al., 2002). The numerous approaches to evaluating responsiveness have recently been reviewed by a number of authors (Liang, 1995; Wyrwich et al., 2000; Beaton et al., 2001; Liang et al., 2002; Terwee et al., 2003).

Responsiveness has been described as the ability of an instrument to measure clinically important change over time, when change is present (Deyo et al., 1991; Fitzpatrick et al., 1998). It has also been argued that responsiveness can be viewed as longitudinal validity or as a measure of treatment effect (Terwee et al., 2003). Patient-reported health instruments have had by far the greatest application in clinical trials and most of the literature on responsiveness relates to the measurement of change in health for groups of patients (Fitzpatrick et al., 1998).

There are two broad approaches to assessing responsiveness: distribution-based and anchor-based (Wyrwich et al., 2000; Norman et al., 2001). Distribution-based approaches relate changes in instrument scores to some measure of variability, the most common method being the effect size statistic. The three widely-reported effect size statistics use the mean score change in the numerator, but have different denominators (Fitzpatrick et al., 1998). The effect size (ES) statistic uses the standard deviation of baseline scores (Liang, 1995). The standardised response mean (SRM) uses the standard deviation of the change score to incorporate the response variance in change scores. However, both the ES and SRM may be influenced by natural variance in the underlying state and by measurement error. The modified standardised response mean (MSRM), or responsiveness index, addresses the inherent natural variance that may occur in patients who otherwise report their health as unchanged, and non-specific score change by using the standard deviation of change in patients who are defined as stable (Deyo et al., 1991). In demonstrating responsiveness to clinically important change, instruments should detect change above the non-specific change incorporated in the MSRM (Deyo et al., 1991).

It has been suggested that statistical measures of responsiveness are an insufficient basis for assessing responsiveness and that patients' views on the importance of the change should inform testing (Liang et al., 2002; Terwee et al., 2003). Anchor-based approaches assess the relationship between changes in instrument scores and an external variable (Norman et al., 2001). This includes health transition items or global

judgements of change used to estimate the Minimal Important Difference (MID), the instrument change score corresponding to a small but important change (Jaeschke et al., 1989; Juniper et al., 2002). The MID can inform sample size calculations but consideration must be given to specific groups of patients and specific settings (Terwee et al., 2003). Score interpretation may be improved through the provision of evidence relating to score variation (Terwee et al., 2003) or a score range against which real change may be assessed (Streiner and Norman, 1995; Beaton et al., 2001).

External variables including transition ratings have also been compared to instrument score changes using correlation. This form of longitudinal validity (Kirshner and Guyatt, 1985; Terwee et al., 2003) assesses the extent to which changes in instrument scores concord with an accepted measure of change in patient health (Deyo et al., 1991; Fitzpatrick et al., 1998).

The ability of an instrument to distinguish clearly and precisely between respondents in relation to reported health or illness is referred to as **precision** (Fitzpatrick et al., 1998). Ideally, items within an instrument should capture the full range of health states to be measured, supporting discrimination between respondents at clinically important levels of health (Fitzpatrick et al., 1998). Precision is influenced by several factors including response categories and item coverage of the defined concept of health purportedly measured by the instrument. Limited response categories lack precision and detail, whereas increased gradations of response increase measurement precision (Streiner and Norman, 1995; Fitzpatrick et al., 1998).

Modern psychometric methods, including Rasch analysis, are also used to assess item distribution. Where there is an uneven distribution of items across the proposed hierarchy of health, for example, item grouping in the middle range of functional ability, score change may be influenced by baseline scores and should be considered when interpreting changes in health.

Item content and response format will inevitably influence data quality and scaling, in which floor and ceiling effects are key features. Where more than 20% of responders score at the maximum level of good or bad health, score distribution generally suggests ceiling or floor effects, respectively (Streiner and Norman, 1995; Fitzpatrick et al., 1998). The greater concern is for respondents with already poor health who score at the floor of the instrument range and are consequently unable to report further deterioration in health. Evidence suggests that floor effects are more common with instrument completion by older, sick, or disadvantaged respondents (McHorney, 1996).

Instrument **acceptability** addresses the willingness or ability of patients' to complete an instrument (Fitzpatrick et al., 1998). Although difficult to evaluate directly, this is most readily assessed through instrument completion, response rates, and missing values. Where items within an instrument are consistently omitted, or difficulty is encountered in providing an answer, perhaps due to perceived irrelevance, this would suggest poor acceptability (McHorney, 1996). The font style and size used in questionnaires may also influence completion. Ideally, patients' should be interviewed for their views on instrument completion, content relevance and format during the pre-testing stage of instrument development (Fitzpatrick et al., 1998).

Reading ability is a further consideration regarding instrument acceptability (Streiner and Norman, 1995). A reading level equivalent to that of a 12 year-old has been recommended for questionnaires applicable to the general population (Streiner and Norman, 1995). However, many instruments, including the widely used Nottingham Health Profile (NHP) and the SF-36 have higher reading level requirements (McHorney, 1996; Sharples et al., 2000). It must also be remembered that reading ability may decrease with age (McHorney, 1996). Lack of familiarity with a questionnaire may further reduce response rates in older people (McHorney, 1996).

Instrument completion will also be influenced by mode of administration. Although cheaper than interview or telephone administration, postal administration often results in higher levels of missing values (McHorney, 1996; McColl et al., 2001). Evidence suggests that respondents are more willing to report less favourable health states when completing an instrument themselves than when the instrument is administered by interview (Fitzpatrick et al., 1998; Smeeth et al., 2001). Furthermore, response rates may be influenced by specific item content, for example, items relating to physical or emotional issues; the associated item relevance and appropriateness to the specific population (Bowling, 1998); and response formats, for example, visual analogue scales or Likert scaling (Fitzpatrick et al., 1998). The burden imposed by instrument length and time needed for completion is an important consideration for both respondent and clinician or researcher.

The **feasibility** of instrument administration refers to the time and cost of administration, scoring, and interpretation for clinicians, researchers, and other staff (Fitzpatrick et al., 1998).

REFERENCES

Atkinson T. Atkinson Review: Final Report – Measurement of Government Output and Productivity for the National Accounts. TSO, London 2005.

Beaton DE, Bombardier C, Katz JN, Wright JG. (2001) A taxonomy for responsiveness. *Journal of Clinical Epidemiology.* 54: 1204-1217.

Bowling A. (1995) Measuring Disease. Open University Press, Buckingham.

Bowling A. (1997) Measuring Health. Open University Press, Buckingham.

Department of Health. Our Health, Our Care, Our Say: A New Direction for Community Services. White Paper, London 2006.

Deyo RA, Diehr P, Patrick DL. (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials;* 12: 142S-158S.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998) Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment;* 2(14).

Garratt AM, Hutchinson A, Russell I. (2001) The UK version of the Seattle Angina Questionnaire (SAQ-UK): reliability, validity and responsiveness. *Journal of Clinical Epidemiology;* 54: 907-915.

Garratt AM, Schmidt L, Mackintosh A, Fitzpatrick R. (2002) Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal;* 324 (7351): 1417-1421.

Jaeschke R, Singer J, Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials;* 10: 407-415.

Juniper EF, Price DB, Stampone PA, Creemers JP, Mol SJ, Fireman P. (2002) Clinically important improvements in asthma-specific quality of life, but no difference in conventional clinical indexes in patients changed from conventional beclomethasone dipropionate to approximately half the dose of extrafine beclomethasone dipropionate. *Chest;* 121(6): 1824-32.

Kirshner B, Guyatt G. (1985) A methodological framework for assessing health indices. *Journal of Chronic Diseases;* 38: 27-36.

Liang MH. (1995) Evaluating measurement responsiveness. *The Journal of Rheumatology;* 22(6): 1191-1192.

Liang MH, Lew RA, Stucki G, Fortin PR, Daltroy L. (2002) Measuring clinically important changes with patient-oriented questionnaires. *Medical Care.* 40(4): Supplement: II45-II51.

McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, Thomas R, Harvey E, Garratt A, Bond J. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess.* 2001;**5**(31):1-256

McDowell I, Jenkinson C. (1996) Development standards for health measures. *Journal of Health Service Research Policy;* October. 1(4): 238-246.

McDowell I, Newell C. (1996) Measuring Health: a guide to rating scales and questionnaires. Oxford University Press, New York.

McHorney CA, Ware JE, Raczek AE. (1993) The MOS-36-item Short-Form Health Survey (SF-36): II. Psychometric and clinic tests of validity in measuring physical and mental health constructs. *Medical Care;* Mar.31(3): 247-63.

McHorney CA. (1996) Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 Health Survey. *Gerontologist;* 36: 571-583.

Norman GR, Sridhar FG, Guyatt GH, Walter SD. (2001) Relation of distribution and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care;* 39(10): 1039-1047.

Nunnally JC, Bernstein IH. (1994) Psychometric Theory. McGraw-Hill Series in Psychology, McGraw-Hill, Inc. Third Edition.

Secretary of State for Health The NHS Improvement Plan 2004. London, HMSO, 2004.

Sharples LD, Todd CJ, Caine N, Tait S. (2000) Measurement properties of the Nottingham Health Profile and Short Form 36 health status measures in a population sample of elderly people living at home: results from ELPHS. *British Journal of Health Psychology;* 5: 217-233.

Smeeth L, Fletcher AE, Stirling S, Nunes M, Breeze E, Ng E, Bulpitt CJ, Jones D. (2001) Randomised comparison of three methods of administering a screening questionnaire to elderly people: findings from the MRC trial of the assessment and management of older people in the community. *British Medical Journal;* 323: 1403-1407.

Streiner DL, Norman GR. (1995) Health Measurement Scales. A practical guide to their development and use. Oxford Medical Publications, Inc. Second Edition.

Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. (2003) On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research;* 12: 349-362.

The NHS Plan. HMSO Stationery Office, London 2000.

Ware JE. (1997) SF-36 Health Survey. Manual and Interpretation Guide. The Health Institute, New England Medical Centre. Boston, MA. Nimrod Press. Second Edition.

Wyrwich KW, Wolinsky FD. (2000) Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice;* 6(1): 39-49.

Chapter 2: METHODS

This chapter briefly summarises the methods used to find the evidence used to inform the various reviews included in this report, together with principles for inclusion and exclusion of evidence. The ways in which evidence was judged is also briefly described.

a) Search strategy

The search strategy was designed to retrieve references relating to patient-reported health instruments for each of the disease groups in this review: that is, Asthma (Chapter 4), COPD (Chapter 5), Diabetes (Chapter 6), Epilepsy (Chapter 7), Heart failure (Chapter 8) and Stroke (Chapter 9). Chapters 10 and 11 report reviews of Carer impact health instruments and Patients perceptions of quality.

Hosted by the National Centre for Health Outcomes Development (NCHOD) at the University of Oxford, the Patient-reported Health Instruments (PHI) website (http://phi.uhce.ox.ac.uk/) includes a bibliography of over 12,000 records relating to published instrument evaluations found on the following electronic databases: Allied and Alternative Medicine (AMED), Biological Abstracts, British Nursing Index, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Econlit, EMBASE, Medline, PAIS International, PsycInfo, System for Information on Grey Literature in Europe (SIGLE), and Sociological Abstracts. At the time of this review, the bibliography comprised references up to June 2005. Details of the search strategy for the bibliography are available on request. The primary search of the bibliography used the terms specific to each disease group, as detailed in each review chapter secondary search of the database used the names of identified instruments.

Additional searching included:

The reference lists of included records were reviewed for additional articles. Hand searching of the following journals was carried out:
-Quality of Life Research
-Health and Quality of Life Outcomes
-Medical Care
Other journals specific to the review disease groups were also hand searched.

Where available, websites designated to the included instruments were identified. Listed references were assessed for inclusion and supplementary information summarised.

b) Inclusion criteria

Titles and abstracts of all articles were assessed for inclusion/exclusion by two independent reviewers and agreement was checked. Included articles were retrieved in full. Published articles were included if they provided evidence of measurement and/or practical properties (Fitzpatrick et al., 1998) for multi-item instruments assessing aspects of health status or quality of life in patients with asthma, COPD, diabetes, epilepsy, heart failure and stroke.

Specific inclusion criteria for generic and disease-specific instruments

- The instrument is patient-reported
- There is published evidence of measurement reliability, validity or responsiveness following completion in the specified patient population
- The instrument has been recommended for use with patients with asthma, COPD, diabetes, heart failure, epilepsy or stroke
- The instrument provides English-language versions for use among adult patients from UK, North America and Australia.
- Evidence is available from English language publications, and instrument evaluations conducted in populations within UK, North America, Australasia.

Exclusion criteria

Clinician-assessed instruments,

- Very narrowly focused or single-item instruments
- Instruments only measuring symptoms
- Instruments without empirical evidence of measurement properties.

c) Data extraction

Data extraction followed pre-defined criteria and included both study-specific issues, such as study design and respondent characteristics, and instrument-specific issues, for example, type and description of instrument, including the domains of health status covered response format, length, and evidence of measurement and practical properties (McDowell and Newell., 1996; Fitzpatrick et al., 1998; Garratt et al., 2002).

d) Format of the reviews

The summary of evidence follows that of previous reviews (McDowell and Newell., 1996; Fitzpatrick et al., 1998; Haywood et al., 2004). Detailed reviews of generic and disease-specific instruments are found in Chapters 4 to 11. The following information is provided for each instrument:

Title

The instrument title as given by the original developer. Instrument developers, year of original publication, and subsequent revision.

Description

The purpose and proposed application of each instrument as defined by the developers.

Instrument development, including item derivation, is summarised where available. Instrument content, the domains of health status covered, for example, pain and social well-being, the number of items, response options, and method of scoring are reported. Instrument modifications are described.

Study specific information

Measurement properties are specific to the population and setting in which an instrument is used (Streiner and Norman., 1995; Fitzpatrick et al., 1998). Study-specific information relating to study design and setting, for example, whether the assessment was carried out in a primary care setting, out-patients or in-patients, population characteristics including inclusion/exclusion criteria, intervention(s), duration of the study and follow-up, and mode of questionnaire administration, informs the interpretation of instrument performance and clinical usefulness. Study-specific information summarises population and study characteristics, for example, age, sex, etc.

Measurement properties

For all instruments published evidence of measurement properties is summarised under the following sub-headings:

- reliability
- validity: i. socio-demographic variables and health-service use
  ii. construct validity: other instruments
  iii. other types of validity
- responsiveness
- precision

Practical properties

Where available, published evidence of acceptability and feasibility is summarised.

Tables summarizing the studies that provide evidence for each included instrument take the following form. A tick (✓) is used to indicate that some minimal level of positive evidence was reported within the study supporting the relevant instrument.

e) Review summaries

Evidence for measurement and practical properties was assessed using accepted criteria (Streiner and Norman, 1995; McDowell and Newell, 1996; Fitzpatrick et al., 1998) (detailed in Chapter 1).

Fitzpatrick et al., (1998) list the domains of health status most frequently identified in the literature as relevant to patient-reported health instruments, as shown in Table 2.1. To support comparison between instruments, instrument content was reviewed against this general classification.

The number of studies in which the instrument has been evaluated is provided; where several publications relate to the same study population, this is indicated.

Although there are relatively clear cut and widely agreed criteria available to assess measurement properties of instruments, there are no clear-cut explicit criteria for how to weigh the balance of evidence or weigh the balance of evidence for instruments comparatively. The reviews reported here are based on weighing up for each of the instruments considered in detail: the volume of available evidence, the quality of studies and, ultimately, the overall extent of positive and supportive evidence of measurement properties and feasibility. To some extent the reviews should be considered as based on a form of 'rapid appraisal'. They were written to inform current and pressing policy initiatives in a prompt and timely fashion. Although we are confident that we have a reasonably up-to-date and representative body of evidence to inform recommendations, in the time available it was not feasible exhaustively to search more inaccessible evidence. Nor was there time or resource to test recommendations against a consensus process of relevant user, professional and scientific judgements.

Table 2.1 Domains of health most commonly assessed in patient-reported health instruments.

I Physical Function Mobility, dexterity, range of movement, physical activity Activities of daily living: ability to eat, wash, dress	
II Symptoms Pain Nausea Appetite	Energy, vitality, fatigue Sleep and rest
III Global judgements of health	
IV Psychological well-being Psychological illness: anxiety, depression Coping, positive well-being and adjustment, sense of control, self-esteem	
V Social well-being Family and intimate relations Social contact, integration, and social opportunities Leisure activities Sexual activity and satisfaction	
VI Cognitive functioning Cognition Alertness Concentration	Memory Confusion Ability to communicate
VII Role activities Employment Household management	Financial concerns
VIII Personal constructs Satisfaction with bodily appearance Stigma and stigmatising conditions Life satisfaction Spirituality	
IX Satisfaction with care	

REFERENCES

Garratt AM, Schmidt L, Mackintosh A, Fitzpatrick R. (2002) Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal;* 324 (7351): 1417-1421.

Haywood KL, Garratt AM, Schmidt L, Mackintosh AE, Fitzpatrick R. *Health Status and Quality of Life in Older People: a Structured Review of Patient-assessed Health Instruments* Report from the Patient-assessed Health Instruments Group to the Department of Health, April 2004.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998) Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment;* 2(14).

McDowell I, Newell C. (1996) Measuring Health: a guide to rating scales and questionnaires. Oxford University Press, New York.

Streiner DL, Norman GR. (1995) Health Measurement Scales. A practical guide to their development and use. Oxford Medical Publications, Inc. Second Edition

Chapter 3: GENERIC INSTRUMENTS

In order to avoid unnecessary repetition, this chapter provides a brief description of the twelve generic health status instruments that appear in one or more of the six chapters reviewing patient-reported instruments for specific chronic conditions. Their origins, development and content are briefly summarized. Content and format are further summarized in table 3.1 at the end of this chapter. Evidence for their use in relation to any given chronic illness is reviewed in the relevant chapters.

a) COOP Charts for Primary Care Practice (Nelson et al., 1987)

The Dartmouth Primary Care Cooperative Information Project developed the COOP charts in the late 1980s to provide a screening tool for use by doctors in routine practice (Nelson et al., 1987). The charts support the assessment of patient health status and functioning.

The original instrument, developed in the USA, has nine charts, each containing a single question about health, functioning, or quality of life during the previous month (Table 3.1). Eight charts assess bodily pain (BP), daily activities (DA), emotional condition/feelings (EC), physical fitness (PF), quality of life (QoL), social activities (SA), social support (SS), and current overall health (OH) perceptions. An additional chart assesses change in overall health. Literature reviews, existing instruments, and discussion with practicing physicians and experts in health status measurement informed item derivation (Nelson et al., 1990).

Following a multinational feasibility study, item content was revised to seven charts, omitting quality of life and social support, with a reduced recall period of two weeks (World Organisation of National Colleges, Academies and Academic Associations of General Practitioners and Family Physicians [WONCA]: WONCA/COOP Health Assessment Charts. Froom, 1988; Langraf and Nelson, 1992). Each chart within the WONCA/COOP includes a descriptive title, a question, and a pictorially illustrated five-point response scale, where five is the most severe limitation. Each represents a separate domain; an overall score is not calculated (McDowell and Newell, 1996). The charts can be self or interview-administered.

b) EuroQol-EQ-5D (The EuroQol Group, 1990; revised 1993)

The European Quality of Life instrument (EuroQol) was developed by researchers in five European countries to provide an instrument with a core set of generic health status items (The EuroQol Group, 1990; Brazier et al., 1993). Although providing a limited and standardized reflection of HRQL, it was intended that use of the EuroQol would be supplemented by disease-specific instruments. The developers recommend the EuroQol for use in evaluative studies and policy research; given that health states incorporate preferences, it can also be used for economic evaluation. It can be self or interview-administered.

Existing instruments, including the Nottingham Health Profile, Quality of Well-Being Scale, Rosser Index, and Sickness Impact Profile were reviewed to inform item content (The EuroQol Group, 1990). There are two sections to the EuroQol: the EQ-5D and the EQ thermometer. The EQ-5D assesses health across five domains: anxiety/depression (AD), mobility (M), pain/discomfort (PD), self-care (SC), and usual activities (UA), as shown in Table 3.1. Each domain has one item and a three-point categorical response scale; health 'today' is assessed. Weights based upon societal valuations of health states are used to calculate an index score of –0.59 to 1.00, where –0.59 is a state worse than death and 1.00 is maximum well-being. A score profile can be reported. The EQ thermometer is a single 20 cm vertical visual analogue scale with a range of 0 to 100, where 0 is the worst and 100 the best imaginable health.

c) Health Utilities Index

The Health Utilities Index (HUI) was designed as a comprehensive measure of health status and health related quality of life. The Health Utilities Index (Mark 3) is a system composed of a health status classification which defines 972,000 discrete health states, and a preference, or utility, function which can be used to calculate the desirability for each health state. The HUI3 health status classification was developed by Feeny et al., (1995) to assess capacity on eight dimensions: vision, hearing, speech, ambulation, dexterity, emotion, cognition and pain/discomfort. The utility function reflects community preferences and scores each unique health state on a scale ranging from 0 (death) to 1 (perfect health). An excellent summary of the development of the HUI measures can be found in Feeny et al., (1996). The HUI3 is a development of the Health Utilities Index containing a sub-set of items which constituted the HUI2. This report summarises data for the most recent version of the HUI (i.e. the HUI3).

d) Nottingham Health Profile (Hunt et al., 1980)

The Nottingham Health Profile (NHP) was developed in the UK during the 1970s for use in the evaluation of medical or social interventions (Hunt et al., 1980). Instrument content was derived from over 2000 statements given by 768 patients with a variety of chronic ailments and other lay people.

Part I of the instrument has 38 items across six domains: bodily pain (BP), emotional reactions (ER), energy (E), physical mobility (PM), sleep (S), and social isolation (SI), as shown in Table 1. All items are statements that refer to departures from normal functioning, and relate to feelings and emotional state rather than change in behaviour. Respondents answer 'yes' or 'no' according to whether or not they feel the item applies to them in general. Positive responses are weighted and summed to give six domain scores between 0 and 100, where 100 denotes maximum limitation.

Part II of the NHP is less widely used and provides a brief indicator of handicap. The instrument may be self- interview-, or telephone-administered.

e) Quality of Life Index (Ferrans and Powers, 1985; Ferrans and Ferrell, 1990)

The Quality of Life Index (QLI) was developed in the USA during the 1980s as a measure of morbidity for application in both normal and unwell populations (Ferrans and Powers, 1985; Bowling, 1995).

Instrument content was informed by literature reviews, which considered quality of life across all age-groups and different illnesses (Kleinpell and Ferrans, 2002). Quality of life was defined as a multidimensional construct with four key domains: family, health and function, psychological and spiritual, and social and economic. The instrument comprises two sections assessing respondent satisfaction and relative importance of each domain, respectively. Each section has 32 items, with eight items per domain. Six-point ordinal response scales range from 'very dissatisfied' or 'very unimportant' (1), to 'very satisfied' or 'very important' (6). Scoring is complicated and the developers recommend a computer programme. In summary, importance scores are used to weight satisfaction scores. Index or domain scores range from 0 to 30, where higher scores indicate better quality of life (Bowling, 1995, p54). The instrument has been self-completed by an older population.

The original instrument was developed and tested in patients receiving haemodialysis, and several dialysis-specific items are available (Bowling, 1995). Factor analysis confirmed instrument construction. The QLI has been modified for use with cancer patients (Bowling, 1995).

f) Quality of Well-Being Scale (formerly the Index of Well-Being) (Kaplan et al., 1976; Kaplan et al., 1984; Kaplan et al., 1993)

The Index of Well-Being was modified and renamed the Quality of Well-Being scale (QWB) to emphasize the focus on quality of life evaluation (Kaplan et al., 1993; McDowell and Newell, 1996).

The QWB uses a three-component model of health (Kaplan and Anderson, 1988, cited by McDowell and Newell, 1996) comprising: 1) functional assessment, 2) a value reflecting the utility or desirability of each functional level, and 3) an assessment of illness prognosis to anticipate future health-care need, which may describe positive health. The QWB is interview-administered.

Completion corresponds to the three-component model. First, three domains of self-reported function are assessed, namely mobility and confinement (MOB: three categories), physical activity (PAC: three categories), and social activity (SAC: five categories). Respondents select the most appropriate category to describe their perceived functional level. Domain categories give 45 possible combinations (3 x 3 x 5); with the inclusion of death, 46 function levels are defined for the second stage of completion (McDowell and Newell, 1996). In addition, respondents select from a list of 27 items symptoms or medical problems experienced over the previous eight days.

Social preference weights for each possible health state have been derived from empirical studies. At the second stage, the assignment of an appropriate weight, or utility, to a health state or functional level gives the QWB index score from 0 to 1, where 0 equates to death and 1 to complete well-being. A negative score may be generated, representing a state 'worse than death'. QWB index scores can be converted into Quality-Adjusted Life-Years (QALYs), supporting their application in economic and policy analysis.

Stage three of the QWB addresses issues of prognosis to produce well-life expectancy score (McDowell and Newell, 1996). This stage is not necessary for calculating the QWB index.

A self-administered version has been developed: the QWB-SA (Andersen et al., 1995). Following a review of QWB items, five items were added to a mental health section and three self-rated health items were included. The QWB-SA has five domains: symptoms and problem complexes (58 acute and chronic items), self-care (two items), mobility, physical functioning (11 items for these two), and performance of usual activity (three items). For the first domain, respondents indicate the presence or absence ('yes' or 'no') of chronic (18), acute physical (25), and mental health symptoms (11) over the previous three days. The remaining four domains all use a three-day recall response option. The total number of items is inconsistent, ranging from 71 to 74. Symptom/problem weights for the QWB-SA are based on the original QWB weighting system. The focus of the original QWB is utility measurement and quality of life; the focus of the QWB-SA is symptoms and assessment of function. The QWB-SA has been recommended for self-completion by older adults (Andersen et al., 1995).

g) Reintegration to Normal Living Index (RNLI)

The Reintegration to Normal Living Index (RNLI) was designed to measure the impact of disease or disability on the individual's ability to resume normal patterns of daily living (Wood-Dauphinee and Williams 1987). The RNLI contains 11 items, on which respondents rate their satisfaction of their physical, emotional and social lives on 100 mm visual analogue scales, where '0' means 'does not describe my situation' and '100' means 'describes my situation'. For the total scores, items scores are summed and averaged. Higher scores represent better reintegration. The RLNI is more limited in focus than other quality of life measures, but it includes similar domains and has been recommended as proxy measure of individual quality of life (Wood-Dauphinee and Williams 1987).

h) SF-36: Medical Outcomes Study 36-item Short Form Health Survey (Ware and Sherbourne, 1992; Ware et al., 1994; Ware, 1997)

The Medical Outcomes Study (MOS) Short Form 36-item Health Survey (SF-36) is derived from the work of the Rand Corporation during the 1970s (Ware and Sherbourne, 1992; Ware et al., 1994; Ware, 1997). It was published in 1990 after criticism that the SF-20 was too brief and insensitive. The SF-36 is intended for application in a wide range of

conditions and with the general population. Ware et al., (1994; 1997) proposed that the instrument should capture both mental and physical aspects of health. International interest in this instrument is increasing, and it is by far the most widely evaluated measure of health status (Garratt et al., 2002a).

Items were derived from several sources, including extensive literature reviews and existing instruments (Ware and Sherbourne, 1992; Ware and Gandek, 1998; Jenkinson and McGee 1998). The original Rand MOS Questionnaire (245 items) was the primary source, and several items were retained from the SF-20. The 36 items assess health across eight domains (Ware, 1997), namely bodily pain (BP: two items), general health perceptions (GH: five items), mental health (MH: five items), physical functioning (PF: ten items), role limitations due to emotional health problems (RE : three items), role limitations due to physical health problems (RP: four items), social functioning (SF: two items), and vitality (V: four items), as shown in Table 3.1. An additional health transition item, not included in the final score, assesses change in health. All items use categorical response options (range: 2-6 options). Scoring uses a weighted scoring algorithm and a computer-based programme is recommended. Eight domain scores give a health profile; scores are transformed into a scale from 0 to 100 scale, where 100 denotes the best health. Scores can be calculated when up to half of the items are omitted. Two component summary scores for physical and mental health (MPS and MCS, respectively) can also be calculated. A version of the SF-36 plus three depression questions has been developed and is variously called the Health Status Questionnaire (HSQ) or SF-36-D.

The SF-36 can be self-, interview-, or telephone-administered.

i) SF-20: Medical Outcomes Study 20-item Short Form Health Survey (Stewart et al., 1988; Ware, Sherbourne and Davies, 1992)

The Medical Outcomes Study (MOS) 20-item Short Form Health Survey (SF-20) is a 20-item abbreviation of the same Rand instrument from which the SF-36 originates (Stewart et al., 1988; Ware et al., 1992; McDowell and Newell, 1996). The abridged instrument was intended to reduce respondent burden, whilst comprehensively addressing important issues in health status measurement.

The SF-20 assesses health across six domains, namely bodily pain (BP: one item), general health perception (GH: five items), physical function (PF: six items), mental health (MH: five items), social function (SF: one item), and role function (RF: two items), as shown in Table1. Items have categorical response options (range: 3-6 options); several items have reversed scoring. Domain item summation scores are transformed into a scale from 0 to 100, where higher values denote better health. The instrument may be self-, interview-, or telephone-administered. Instrument self-administration takes approximately four minutes (McDowell and Newell, 1996), but longer completion times have been reported for older people (Siu et al., 1993a, b).

j) SF-12: Medical Outcomes Study 12-item Short Form Health Survey (Ware et al., 1995)

In response to the need to produce a shorter instrument that could be completed more rapidly, the developers of the Medical Outcomes Study (MOS) 36-item Short Form Health Survey (SF-36) produced the 12-item Short Form Health Survey (SF-12) (Ware et al., 1995).

Using regression analysis, 12 items were selected that reproduced 90% of the variance in the overall Physical and Mental Health components of the SF-36 (Table 3.1). The same eight domains as the SF-36 are assessed and categorical response scales are used. A computer-based scoring algorithm is used to calculate scores: Physical Component Summary (PCS) and Mental (MCS) Component Summary scales are generated using norm-based methods. Scores are transformed to have a mean value of 50, standard deviation (SD) 10, where scores above or below 50 are above or below average physical or mental well-being, respectively. Completion by UK city-dwellers reporting the absence of health problems yielded a mean PCS score of 50.0 (SD 7.6) and MCS of 55.5 (SD 6.1) (Pettit et al., 2001). Although not recommended by the developers, Schofield and Mishra (1998) report eight domain scores and two summary scores. The SF-12 may be self-, interview-, or telephone-administered.

Several authors have proposed simplification of the scoring process and revision of the SF-12 summary score structure, where norm-based weighting is replaced by item summation to facilitate score interpretation (Resnick and Nahm, 2001; Resnick and Parker, 2001).

k) SF-6D

The SF-6D was designed to be used in health economic analyses. It is a classification for describing health derived from a selection of SF-36 items. It is composed of six multi-level dimensions.It is a preference based algorithm based on a sub-set of items from the SF36, developed by Brazier et al., (2002). The SF-6D comes with a set of preference weights obtained from a sample of the general population. Using the valuation technique of standard gamble, members of the general population were asked to value a selection of health states from which a model has been estimated to predict all the health states described by the SF-6D.

l) Sickness Impact Profile (Bergner et al., 1976; revised: Bergner et al., 1981)

The Sickness Impact Profile (SIP) was developed in the USA to provide a broad measure of self-assessed health-related behaviour (Bergner et al., 1976; Bergner et al., 1981). It was intended for a variety of applications, including programme-planning and assessment of patients, and to inform policy decision-making (Bergner et al., 1976; Bergner et al., 1981; McDowell and Newell, 1996).

Instrument content was informed by the concept of 'sickness', which was defined as reflecting the change in an individual's activities of daily life, emotional status, and attitude as a result of ill-health (McDowell and Newell, 1996). Item derivation was based

on literature reviews and statements from health professionals, carers, patient groups, and healthy subjects describing change in behaviour as a result of illness. The SIP has 136 items across 12 domains: alertness behaviour (AB: ten items), ambulation (A: 12 items), body care and movement (BCM: 23 items), communication (C: nine items), eating (E: nine items), emotional behaviour (EB: nine items), home management (HM: ten items), mobility (M: ten items), recreation and pastimes (RP: eight items), sleep and rest (SR: seven items), social interaction (SI: 20 items) and work (W: nine items).

Each item is a statement. Statements that best describe a respondent's perceived health state on the day the instrument is completed are ticked. Items are weighted, with higher weights representing increased impairment. The SIP percentage score can be calculated for the total SIP (index) or for each domain, where 0 is better health and 100 is worse health. Two summary scores are calculated: Physical function (SIP-PhysF), a summation of A, BCM, and M, and psychosocial function (SIP-PsychF), a summation of AB, C, EB, and SI. The five remaining categories are scored independently. The instrument may be self or interview-administered.

The Functional Limitation Profile (FLP) is an Anglicized version of the SIP (Patrick and Peach, 1989; McDowell and Newell, 1996). Wording and some weightings have been altered, and summary scores are calculated using different dimensions to those used in the SIP (i.e. FLP Physical summary calculated by summing A, BCM, M and HM; FLP Psychosocial summary calculated by summing RP, EB, AB, SI and SR. Several abbreviated versions of the SIP have been developed, including a 68-item version (De Bruin et al., 1992; Post et al, 1996).

GENERIC INSTRUMENTS

Table 3.1 Generic patient-reported health instruments:

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Completion (time in minutes)*
COOP Charts for Primary Care Practice (COOP) (8+1)	Bodily pain (BP) (1), Daily activities (ADL) (1), Emotional condition (EC) (1), Physical fitness (PF) (1), Quality of life (QL) (1), Social activities (SA) (1), Social support (SS) (1), Overall health perception (OH) (1), Change in health status (1)	Categorical: 1-5 (illustrated) 2-week recall	Chart profile (1-5, 5 no limitations)	Interview or self
European Quality of Life Questionnaire (EuroQol-EQ5D) (5+1)	EQ-5D Anxiety/depression (1), Mobility (1), Pain/discomfort (1), Self-care (1), Usual activities (1) EQ-thermometer Global health (1)	EQ-5D Categorical: 3 options *EQ-thermometer* VAS Current health	EQ-5D Summation: domain profile Utility index (–0.59 to 1.00) *Thermometer* VAS (0-100)	Interview or self
Health Utility Index 3 (Feeny et al, 1995) (8)	Vision, Hearing, Speech, Ambulation, Dexterity, Emotion, Cognition, Pain	Four domains have five response options and five have six response options	Global Utility index and single attribute utility scores for the eight separate dimensions	Self report, face to face and telephone interview
Nottingham Health Profile (NHP) (38)	Bodily pain (BP) (8), Emotional reactions (ER) (9), Energy (E) (3), Physical mobility (PM) (8), Sleep (S) (5), Social isolation (SI) (5)	Yes/no; positive responses weighted Recall 'general' health	Algorithm Domain profile 0-100, 100 is maximum limitation	Interview Self (10-15)
Quality of Life Index (QLI) (64)	*Satisfaction (S) and Importance (I) of each domain:* Family (S 8, I 8) Health and functioning (S 8, I 8) Psychological / spiritual (S 8, I 8) Social and economic (S 8, I 8)	Likert scale 1-6 for satisfaction, importance	Algorithm: satisfaction score weighted by importance score Domain profile (0-30, 30 best HRQL) Index (0-30)	Self
Quality of Well-being Scale (QWB) (30)	Mobility and confinement (MOB) (3 categories) Physical activity (PAC) (3 categories) Social activity (SAC) (5 categories) Symptoms and medical problems (27)	Categorical: yes/no Recall 6 days Symptoms 8 days	Algorithm Index 0-1, 1 complete well-being	Interview Telephone (mean 17.4, range 6-30)
Quality of Well-being - Self-administered (QWB-SA) (71-74)	Mobility and Physical functioning (11) Self-care (2), Usual activity (3) Symptoms (58): acute physical (25), chronic (18), mental health (11)	Categorical: yes/no Recall 3 days	Algorithm Index 0-1, 1 complete well-being)	Self (mean 14.2)
Reintegration to Normal Living Index (RNLI)	Satisfaction with 11 aspects of physical, emotional and social lives	Visual analogue scale	Summation and averaging of responses to individual items	Self
SF-36: MOS 36-item Short Form Health Survey (36)	Bodily pain (BP) (2), General health (GH) (5) Mental health (MH) (5), Physical functioning (PF) (10) Role limitation-emotional (RE) (3), Role limitation-physical (RP) (4), Social functioning (SF) (2), Vitality (V) (4)	Categorical: 2-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Domain profile (0-100, 100 best health) Summary: Physical (PCS), Mental (MCS) (mean 50, sd 10)	Interview (mean values 14-15) Self (mean 12.6)

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Completion (time in minutes)*
SF-20: MOS 20-item Short Form Health Survey (20)	Bodily pain (BP) (1), General health (GH) (5) Mental health (MH) (5), Physical functioning (PF) (6) Role functioning (RF), Social functioning (SF) (1)	Categorical: 3-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Summation Domain profile (0-100, 100 best health)	Self (5-7)
SF-12: MOS 12-item Short Form Health Survey (12)	Bodily pain (BP) (1), Energy/Vitality (V) (1), General health (GH) (1), Mental health (MH) (2), Physical functioning (PF) (2), Role limitation-emotional (RE) (2), Role limitation-physical (RP) (2), Social functioning (SF) (1)	Categorical: 2-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Domain profile (0-100, 100 best health) Summary: Physical (PCS), Mental (MCS) (mean 50, sd 10)	Interview or self
SF-6D: MOS 6-item Short Form Health Survey (12)	Bodily pain (BP) (1), Energy/Vitality (V) (1), Mental health (MH) (1), Physical functioning (PF) (1), Role limitation (1), Social functioning (SF) (1)	Categorical: 3 options	Algorithm Domain profile (0-100, 100 best health)	Interview or self
Sickness Impact Profile (136)	Alertness behaviour (AB) (10), Ambulation (A) (12) Body care and movement (BCM) (23), Communication (C) (9) Eating (E) (9), Emotional behaviour (EB) (9) Home management (HM) (10), Mobility (M) (10) Recreation and pastimes (RP) (8), Sleep and rest (SR) (7) Social interaction (SI) (20), Work (W) (9)	Check applicable statements. Items weighted: higher weights indicate increased impairment Recall current health	Algorithm Domain profile (0-100%, 100 worst health); Index (0-100%) Summary: Physical (A, BCM, M), Psychosocial function (AB, C, EB, SI)	Interview (range: 21-33) Telephone: PF only (11.5) Self (19.7)

Table 3.2 Summary of generic instruments: health status domains ***(after Fitzpatrick et al., 1998)***

Instrument	***Instrument domains***							
	Physical function	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**
COOP (9) WONCA (6)	x	x	x	x	x		x	
EQ-5D (5+1)	x	x	x	x	x		x	
HUI	x			x		x		
NHP (38)	x	x		x	x			
QLI (64)	x			x	x		x	x
QWB (11)	x				x			
RNLI	x			x	x			
SF-12 (12)	x	x	x	x	x		x	
SF-20 (20)	x	x	x	x	x		x	
SF-36 (36)	x	x	x	x	x		x	
SIP (136)	x	x		x	x	x	x	

REFERENCES

Andresen EM, Rothenberg BM, Kaplan RM. (1998b) Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. *Medical Care;* 36: 1349-1360.

Bergner M, Bobbitt RA, Kressel S, et al. (1976) The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. *International Journal of Health Service;* 6: 393-415.

Bergner M, Bobbitt RA, Carter WB et al. (1981) The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care;* 19:787-805.

Bowling A. (1995) Measuring Disease. Open University Press, Buckingham.

Bowling A, Windsor J. (1997) Discriminative power of the health status questionnaire 12 in relation to age, sex, and long-standing illness: findings from a survey of households in Great Britain. *Journal of Epidemiology and Community Health;* 51: 564-573.

Brazier JE, Jones N, Kind P. (1993) Testing the validity of the EuroQol and comparing it with the SF-36 Health Survey questionnaire. *Quality of Life Research;* 2: 169-180.

Brazier JE, Walters SJ, Nicholl JP, Kohler B. (1996) Using the SF-36 and EuroQol on an elderly population. *Quality of Life Research;* 5: 195-204.

Brazier JE, Roberts j, Deverill M: The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* 2002; 21 :271-292.

De Bruin AF, De Witte LP, Stvene F, Diederiks JPM. (1992) Sickness Impact Profile: the state of art of a generic functional status measure. *Social Science and Medicine;* 35: 1003-14.

Feeny, David, William Furlong, Michael Boyle, and George W. Torrance, "Multi-Attribute Health Status Classification Systems: Health Utilities Index." *PharmacoEconomics*, Vol 7, No 6, June, 1995, pp 490-502

Ferrans CE, Ferrell BR. (1990) Development of a quality of life index for patients with cancer. *Oncology Nursing Forum;* 17: 15-19(supplement).

Ferrans CE, Powers MJ. (1985) Quality of Life Index: development and psychometric properties. *Advances in Nursing Science;* 8: 15-24.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998) Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment;* 2(14).

Froom J. (1988) Preface: WONCA committee on international classification. Statement on functional status assessment. Calgery, October. In: Lipkin M (editor).

Garratt AM, Schmidt L, Mackintosh A, Fitzpatrick R. (2002a) Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal;* 324 (7351): 1417-1421.

Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. (1980) A quantitative approach to perceived health status: a validation study. *Journal of Epidemiology and Community Health;* 34: 281-286.

Jenkinson C, McGee H. (1998). Health Status Measurement, a Brief but Critical Introduction. Radcliffe Medical Press Ltd, Oxford.

Kaplan RM, Bush JW, Berry CC. (1976) Health status: Types of validity and the index of well-being. *Health Service Research;* 11:478.

Kaplan RM, Anderson JP. (1988) A General Health Policy Model: update and applications. *Health Service Research;* 23: 203-235.

Kaplan RM, Anderson JP, Ganiats TJ. (1993) The Quality of Well-Being Scale: rationale for a single quality of life index. In: Walker SR, Rosser RM (Eds). Quality of Life Assessment: key issues in the 1990s. Dordrecht, Netherlands: Kluwer Academic Publishers: 65-94.

Kleinpell RM, Ferrans CE. (2002) Quality of life of elderly patients after treatment in the ICU. *Research in Nursing and Health;* Jun. 25(3): 212-21.

Landgraf JM, Nelson EC. (1992) Dartmouth COOP primary care network. Summary of the WONCA/COOP international health assessment field trial. *Australian Family Physician;* 21: 255-269.

McDowell I, Jenkinson C. (1996) Development standards for health measures. *Journal of Health Service Research Policy;* October. 1(4): 238-246.

McDowell I, Newell C. (1996) Measuring Health: a guide to rating scales and questionnaires. Oxford University Press, New York.

Nelson EC, Wasson J, Kirk J, et al. (1987) Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. *Journal of Chronic Disease;* 40(suppl 1): 55S-63S.

Nelson E, Landgraf J, Hays R, Wasson J, Kirk J. (1990) The functional status of patients. How can it be measured in physicians' offices? *Medical Care;* 28: 1111-1126.

Patrick D and Peach H (eds) (1989) Disablement in the Community: A Sociomedical Press Perspective. Oxford, U.K.: Oxford University Press

Pettit T, Livingston G, Manela M, Kitchen G, Katona C, Bowling A. (2001) Validation and normative data of health status measures in older people: the Islington study. *International Journal of Geriatric Psychiatry;* 16: 1061-1070.

Post MWM, de Bruin AF, de Witte et al (1996) The SIP68: a measure of health related functional status in rehabilitation medicine Archives of Physical Medicine and Rehabilitation 77, 440

Resnick B, Nahm ES. (2001) Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. *Journal of Nursing Measurement;* Fall. 9(2): 151-61.

Resnick B, Parker R. (2001) Simplified scoring and psychometrics of the revised 12-item Short-Form Health Survey. *Outcomes Management for Nursing Practice;* Oct-Dec. 5(4): 161-6.

Rosser RM, Allison R, Butler C, et al. (1993) Chapter 8: The Index of Health-Related Quality of Life (IHQL): a new tool for audit and cost-per QALY analysis. In: Walker RS, Rosser RM. (Eds) Quality of Life Assessment: key issues in the 1990s. Dordrecht: Kluwer Academic.

Schofield M, Mishra G. (1998) Validity of the SF-12 compared with the SF-36 health survey in pilot studies of the Australian longitudinal study on women's health. *Journal of Health Psychology;* 3: 259-271.

Siu AL, Reuben DB, Ouslander JG, Osterweil D. (1993a) Using multidimensional health measures in older persons to identify risk of hospitalisation and skilled nursing placement. *Quality of Life Research;* 2: 253-261.

Stewart AL, Hays RD, Ware JE. (1988) The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Medical Care;* 26: 724-735.

Streiner DL, Norman GR. (1995) Health Measurement Scales. A practical guide to their development and use. Oxford Medical Publications, Inc. Second Edition.

The EuroQol Group. (1990) EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy;* 16:199-208.

Ware JE, Sherbourne CD, Davies AR. (1992) Developing and testing the MOS 20-Item Short-Form Health Survey: A general population application. In: Stewart AL, Ware JE. (editors). Measuring functioning and well-being: the Medical Outcomes Study approach (pp.277-290). Durham, NC: Duke University Press.

Ware JE, Sherbourne CD. (1992) The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care;* 30: 473-483.

Ware J, Kosinski M, Keller SD. (1994) SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute, New England Medical Centre.

Ware JE, Kosinski M, Keller SD. (1995) SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. The Health Institute, New England Medical Center. Boston, MA. Second Edition.

Ware JE. (1997) SF-36 Health Survey. Manual and Interpretation Guide. The Health Institute, New England Medical Centre. Boston, MA. Nimrod Press. Second Edition.

Ware JE, Gandek B. (1998) Methods for testing data quality, scaling assumptions and reliability: the IQOLA project approach. *Journal of Clinical Epidemiology;* 51(11): 945-952.

Wood-Dauphinee S Williams JI. Reintegration to normal living as a proxy to quality of life. J Chron Dis. 1987; 40: 491-499.

Chapter 4: Patient-reported Health Instruments used for people with Asthma

Asthma is a chronic inflammatory disorder of the airways associated with variable airflow limitation. Obstruction of the airways can be reversible either spontaneously or with treatment. Diagnosis is often clinical with observed changes in lung function during periods of exacerbation. Whilst symptoms include wheeze, shortness of breath, chest tightness and cough these are not specific to asthma. However, symptoms tend to be intermittent and provoked by triggers such as pollens, dust, exercise, chemicals, smoke and infections. Other associated atopic disorders include eczema and allergic rhinitis (SIGN 2005).

The impact of asthma will be dependant on the severity of the disease and triggers specific to individuals. Exercise and activity limitations particularly during the pollen season can result in social isolation which may also limit employment opportunities. The physical burden of the disease can result in fatigue and sleep disturbance. Understanding the specific impact on a patient's life can contribute to successful management of the condition.

The following review provides current information available of the patient-reported health questionnaires used to measure health-related quality of life with patients with asthma.

Search terms and results: identification of articles
At the time of the review, the PHI database contained 12,000+ records (up to June 2005). The primary search strategy, using the terms 'asthma' and 'respiratory' keyword searching generated 468.records, as shown in Table 4.1. All abstracts were reviewed. When assessed against the review inclusion criteria, 220 articles were retrieved and reviewed in full. Of these, 50 articles were included in the review.

Table 4.1 Number of articles identified by the literature review

Source	***Results of search***	***No. of articles considered eligible***	***Number of articles included in review***
PHI database: original search (up to June 2005) **Total number= 12,562**	468	220	44
Supplementary search	-	-	6
TOTAL	-	-	**50**

Supplementary searches which included hand searching of titles from 2004 to 2006 of the following key journals:
-Chest
-Health and Quality of Life Outcomes
-Medical Care
-Quality of Life Research
-Respiratory Medicine
-Thorax

Further searches were conducted within the bibliography and using Pub Med per instrument up to September 2006.

Identification of patient-reported health instruments

Five generic and 10 asthma-specific instruments were included in the review. The developmental and evaluative studies relating to the generic instruments reviewed are listed in Tables 4.2 and 4.3. Those relating to asthma-specific instruments are shown in Tables 4.4 to 4.10.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Seven generic instruments were identified which were evaluated with patients with COPD. For full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36
b) SF-12
c) EQ-5D
d) Sickness Impact Profile
e) Health Utilities Index

a) SF-36

Twelve studies describe the evaluation of the SF-36 following completion by patients with diagnosed asthma, as shown in Table 4.2. Two studies evaluated the SF-36 as the principal instrument with patients with asthma (Keller et al., 1997; Caro et al., 2001). Nine studies describe the concurrent evaluation of the SF-36 alongside other asthma-specific instruments (Blumenschien et al., 1998; Juniper et al., 2000, 2001; Lee et al., 2003; Mancuso et al., 2001; Mancuso and Peterson 2004; McColl et al., 1995, 2003; Ware et al., 1998). One publication describes an expert consensus conference in which recommendations are made for clinically important differences for the SF-36 (Wywrich et al., 2003).

Keller et al., (1997) evaluated both standard and acute forms of the SF-36, the acute form with a one week recall period and the standard version of 4 week recall. Different order of administration of patient-reported health instruments was evaluated in McColl et al., (2003) with version 1 containing asthma-specific instruments first and in version 2 generic instruments presented first to examine order effects.

Studies were carried out in a primary care or out-patient setting. Two studies describe evaluation of the SF-36 following clinical trials of the effectiveness of different medications. One study was conducted in the UK (McColl et al., 1995). The average age of the patients was forty years. One study used a postal survey as the method of administration and the other study used interviews.

Reliability
Four studies reported evidence of reliability (Keller et al., 1997; Ware et al., 1998; Juniper et al., 2001, McColl et al., 2003).

Test-retest reliability was reported in a concurrent evaluation of the SF-36, Asthma Quality of Life Questionnaire (AQLQ)-Juniper, Standard Gamble (SG) and Rating Scale (RS) (Juniper et al., 2001). Moderate levels of reliability were reported for the SF-36 Physical and Mental component scores (ICC 0.65 PCS; 0.68 MCS); reliability levels were greater for the comparator instruments (>0.70), with the exception of the SG.

With the exception of the domains RE (St. 0.79; Ac 0.59) and MH (Ac 0.64), similar moderate to high levels of internal consistency reliability have been reported for both standard and acute forms of the SF-36 across all domains (all greater than 0.70); the authors attribute the low levels of reliability for the RE domain to low variability in group scores (Keller et al., 1997). Lower levels of internal reliability were also reported for the RP, RE, SF and PI domains (range 0.63 to 0.65) (Ware et al., 1998).

Different order of administration of patient-reported health instruments was evaluated in McColl et al., (2003) with version 1 containing asthma-specific instruments first and in version 2 generic instruments presented first. Internal consistencies were in excess of 0.80 for the SF-36 with PF and BP greater than 0.90 for version 1 and PF, RP and BP for version 2. Generally, alphas were slightly higher for version 2 where the SF-36 was presented first.

Item level analysis
Low levels of item-total correlation (less than 0.40) have been reported for the RE and MH domains (Keller et al., 1997).

High scaling success rates across all domains, where the percentage of scaling successes (positive correlations with hypothesised domains) is reported relative to the total number of scaling tests with other domains, was reported for both standard and acute forms (Keller et al., 1997). Response consistency (proposed by the SF-36 developers as an internal consistency check on 15 item pairs) for the standard form was comparable to US population norms (91.2% and 90.3%); the acute form was lower (86.5%), with the greatest inconsistencies for the MH and GH domains.

High levels of item-discriminant validity (the percentage of times that items correlate higher in the hypothesised domain than other domains) were reported (greater than 0.4) (Keller at al., 1997).

Validity
Seven studies reported evidence of internal and/or construct validity (Blumenschien 1998; Juniper et al., 2000, 2001; Keller et al., 1997; Mancuso et al., 2001; McColl et al., 1995, 2003).

Internal validity
Principal components analysis supported the two factor high-order solution of MCS and PCS, and the eight-domain structure proposed by the instrument developers (Keller et al., 1997).

Health status

Viramontes and O'Brien (1994) evaluated the discriminative validity of the SF-36 with patients with chronic lung diseases including asthma, emphysema and chronic bronchitis and reported significantly different domain scores between disease severity subgroups based on the UK Medical Research Council symptoms classification. Lower SF-36 scores were associated with higher dyspnoea scores as expected and moderate to large correlation was reported for activity threshold and SF-36 PF, GHP and EG. There was no relationship between disease severity and SF-36 ER, SF, BP and MH.

The SF-36 discriminated between asthma and angina patients in McColl et al., (2003) with higher scores for all domains with the exception of MCS and MH domain indicating less impairment for the asthma patients which was in accordance with hypotheses. Version effect was observed for six domains with the exception of PF, BP and MCS with higher scores than predicted for version 2 (SF-36 presented first) (McColl et al., 2003).

Asthma-specific measures of health-related quality of life

Moderate levels of correlation have been were reported between SF-36 domains and several asthma-specific measures of health status (Blumenschien and Johannesson 1998; Juniper et al., 2000, 2001; Mancuso et al., 2001; McColl et al., 1995; Ware et al., 1998).

Correlation between the SF-36 and the Asthma TyPE ranged from -0.32 to -0.58. Most correlations were in the hypothesised direction; the association between the SF-36 PCS, MCS and the Allergy Index component of the Asthma TyPE did not have hypothesised correlations (Blumenschien and Johannesson 1998).

Correlations between the MCS and PCS scores of the SF-36 and the Asthma Control Questionnaire and Asthma Control Diary were as follows: MCS: 0.19 ACQ; 0.31 ACD: PCS 0.53 ACQ; 0.55 ACD (Juniper et al., 2000)) which indicates greater strength of correlation between the ACQ and ACD and the PCS of the SF-36.

Small to moderate levels of correlation respectively were reported between the SF-36 MCS and the PCS with the AQLQ-Juniper (Mancuso et al., 2001; Lee et al., 2003); levels of correlation with the MCS were smaller than hypothesised (Mancuso et al., 2001). Further hypothesises correlations (greater or less than 0.60) were reported for related domains in a study by McColl et al., (2003). This study examined order effects with either asthma-specific measures presented first in the questionnaire package or generic. There was with a slight trend for stronger correlations between the SF-36 and the AQLQ when specific measures were administered first (McColl et al., 2003).

Correlations between related physical function domains on the SF-36 (PF, RP) and the LWAQ (PF) were 0.70 to 0.80. Moderate correlations between related instrument domains assessing elements of emotional health were reported (0.45 to 0.54). Similar results were reported for the social functioning domains (0.54 to 0.64) (McColl et al., 1995).

Moderate levels of correlation were reported between several SF-36 domains (PF, RP, E, SF) and Asthma symptom frequency (range 0.22 MH to 0.59 PF (McColl et al., 1995).

Health utilities
The relationship between a range of methods for obtaining health utilities (Heath State Utilities (Rating Scale (RS), Time Trade Off (TTO), Standard Gamble (SG)), Willingness to Pay (Bid Game; Dichotomous Choice (DC)) and the SF-36 have been explored (Blumenschien and Johannesson 1998). SF-36 domains have small to moderate levels of correlation with the RS (ranging from SF-36 RF 0.28 to SF-36 PF 0.63); these are generally greater levels of correlation than with the TTO (SF-36 range 0.01 to 0.34) and SG (SF-36 range -0.01 to 0.30).

Small levels of correlation between the SF-36, the SG (PCS 0.19; MCS 0.38) and the RS (PCS 0.36; MCS 0.52 were reported in Juniper et al., (2001).

The SF-36 (PF, BP, GHP, VT and SF) had hypothesised correlations (greater than 0.60) for the EQ-5D score (McColl et al., 2003).

The SF-36 has been applied in other studies with patients with asthma where the principal instrument undergoing evaluation is an asthma-specific questionnaire and hypotheses stated about those instruments validity and relationship with the SF-36 (Adams et al., 2000; Juniper et al., 1999a, b, c; Leidy 1998b; Katz et al., 2002 and Reid et al., 1999).

Respiratory function
Small levels of correlation were reported between the SF-36 and respiratory function and medication use (Juniper et al., 2001).

Responsiveness
Responsiveness was reported in six concurrent evaluations (Juniper et al., 2001; Keller et al., 1997; Lee et al., 2003; Mancuso 2001; Mancuso and Peterson 2004; Ware et al., 1998); although responsive to change, the SF-36 was less responsive than asthma-specific instruments in these evaluations.

Following completion by asthmatics taking part in trial of asthma medication, the PCS was able to detect change in physical health both within and between groups of patients; the MCS did not detect change (responsiveness index -0.06) Juniper et al., (2001).

Receiver Operating Characteristics (ROC) curves were used to assess the sensitivity and specificity of the SF-36 and AQLQ-Juniper to patient perceived change in current disease activity (external criteria classified as cases or non-cases (active or non active disease (Mancuso et al., 2001). The SF-36 PCS discriminated between patient's perceptions of disease activity; however ROC curves ranked lower than the AQLQ-Juniper. These results were confirmed in a later study (Mancuso and Peterson 2004); although different analyses of longitudinal data were utilised, all results were in the same direction with lower ROC curves for SF-36 MCS and PCS.

Keller et al., (1997) used patient self-report of change in health (improved, stayed the same, declined) as external criteria for the assessment of instrument responsiveness; the SF-36 acute form was more responsive to change in clinical status over the past week than the standard form.

Ware et al., (2001) also adopted patient-reported, and clinician-reported, assessment of change in health as external criteria. Small to moderate significant relative validity coefficients were observed for all domains (range 0.11 to 0.52) for patient perceived change. Moderate correlation was reported for the RF domain and clinician-assessed change (Treatment Impact; Cough and Wheeze) and RP for Treatment Impact and Cough. All other domains had small correlations with the external criteria. The SF-36 did not perform as well as the MAQLQ with lower relative validity coefficients.

Correlations were reported between changes in SF-36 domain scores and AQLQ score changes with a range of 0.20 for RE to PF 0.62 (Lee 2003). Furthermore, the SF-36 was not as responsive as the AQLQ with ES per domain lower than for the AQLQ (Domains: Large ES: RP, PF; Medium ES GH, VT, SF and small BP, MH and RE Lee 2003). SRM's for the SF-36 domains were lower than for the AQLQ (0.92 to 0.29 vs. 1.17).

Interpretation

Expert consensus

Wyrwich et al., (2003) report on an expert consensus process with the aim to generate recommendations for clinically important differences for the SF-36 and AQLQ.
A modified RAND method was adopted to inform a consensus agreement:

- systematic review of the literature;
- recruitment of healthcare professional experts and researchers for consultation;
- Delphi consensus technique and a subsequent meeting to achieve consensus and formulate recommendations.

Both the SF 36 and AQLQ-Juniper were assessed by the expert group. Recommendations for the interpretation of clinically important differences for SF-36 domains were made: small change in score equates 10 to 16 points; moderate change in score equates 20 to 33 points; large change in score equates 30 to 37 points.

Precision

Three studies reported evidence of precision (Keller et al., 1997; Mancuso et al., 2001; Ware et al., 1998).

Mean scores on the RE scale for the Acute form were significantly higher than scores on the Standard form (Keller et al., 1997); RP and SF Acute form scores were also higher (non-significant). Ceiling effects have been reported for RP, SF and RE (Standard form) and RP, BP, SF and RE (Acute form) (50 to 77%). Further ceiling effects were reported for the SF 36 (Mancuso et al., 2001; Ware et al., 1998).

The Acute and Standard forms had no floor effects (Keller et al., 1997).

Acceptability
Three studies report different aspects of patient acceptability (Caro et al., 2001; Keller et al., 1997; Ware et al., 1998).

Four (6%) of patients indicated that they had no preference for different versions of the instrument in the evaluative study by Caro et al., (2001), 49 (77%) expressed a preference for the electronic version and found it easy to use (this includes preferences for AQLQ combined). A total of 43 spoiled responses were recorded for the paper version of the SF-36 (the electronic version does not permit multiple responses).

The concordance of responses on electronic versus paper versions was also compared (Caro et al., 2001). Patients completed instruments two hours apart; the order of presentation was alternated. A high degree of concordance for patient's scores across either completion format was reported (range 0.83 to 0.96).

Patients preferred the Acute form to the Standard form in Keller et al., (1997) (15/18).

Ware et al., (1998) reported that 94% of responses were logically consistent. McColl (2003) hypothesised that responses would be higher and quicker when asthma-specific instruments (AQLQ, NASQ) were presented before generic instruments (EQ-5D, SF-36). No order effect was found for versions for response rates or response speed.

Feasibility
Completion times for paper or electronic versions of the SF-36 instrument were compared (Caro et al., 2001). A statistically significant difference was reported: 11.21 minutes (electronic) vs. 9.47 minutes (paper).

b) SF-12

The SF-12 has been evaluated in three studies (Franic et al., 2005; Garratt et al., 2000; Magid et al., 2004), one of which was in the UK (Garratt et al., 2000). Two studies used a postal survey as the method of administration. The average age of the patients was forty years.

Reliability
No evidence reported.

Validity
Healthcare utilisation
The SF-12 PCS was predictive of asthma related Emergency department utilisation (Magid et al., 2004). A 10 point (1 SD) decrement was found to be associated with a 72% increased risk of hospital admission / ED admission (OR 1.72; 95% CI 1.46 to 2.02). The SF-12 MCS was not predictive of ED utilisation (OR 1.17; 95% CI 0.96 to 1.44). Scores for patients with asthma were significantly lower than the US norms for the PCS but not for the MCS (Franic et al., 2005).

Asthma-specific patient-reported health instruments
Hypothesised correlations were moderate between the SF-12 Physical component and NASQ (PCS 0.58; MCS 0.36), AQLQ(S) Juniper (PCS 0.58; MCS 0.34). Further similar results were observed for the SF-12 and ACQ with correlations -0.76 for PCS and 0.03 MCS with corresponding correlations for similar domains (Franic et al., 2005)**.**

Generic patient-reported health instruments
Moderate levels of correlation were reported between the SF-12 and EQ-5D (MCS 0.37, PCS 0.49) (Garratt et al., 2000).

Responsiveness
One study evaluated responsiveness of the SF-12 in a concurrent evaluation (Garratt et al., 2000) and reported moderate levels of responsiveness for the PCS (SRM 0.35) which was higher than the EQ-5D. The smallest SRM was reported for the MCS (0.03) suggesting little or no responsiveness.

Precision
No evidence identified.

Acceptability
High levels of completion rates were reported for the SF-12 (94%) in a study by Garratt et al., (2000).

Feasibility
No evidence identified.

a) SF-36 and b) SF12

Table 4.2: Evaluative studies relating to the SF-36 and SF-12 when completed by patients with asthma

Study/ Country	**Population (N) Age (years) Method of administration Setting**	**Measurement and Practical properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Blumenschien and Johannesson (1998) USA	Asthma (69) Age: mean 40 Interview administered Out-patients		Construct ✓				
Caro et al., 2001 Canada	Asthma (68) Age: range 16-75 Interview but patient completed Out-patients					✓	
Juniper et al., 2000 Canada	Asthma (50) Age: mean 37 Self completed Out-patients	Test re-test ✓	Construct ✓	✓			
Juniper et al., 2001 Canada	Symptomatic asthma (40) Age: mean 38 Interview administered Out-patients	Test re-test ✓	Construct ✓	✓			
Keller et al., 1997 USA	Participants in a RCT of asthma medication (142) Age: mean 39 Self report Out patients	Internal consistency ✓	Construct ✓ Internal validity ✓				
Lee et al., 2003 USA	Participants in a RCT of asthma medication (241) Age: mean 38 Self report-hand held electronic device recording patients responses to the instruments) Out patients		Construct ✓	✓			

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Mancuso and Peterson (2004) USA	Asthmatics identified for healthcare plan (185) Age: mean 41 Postal			✓			
Mancuso et al., 2001 USA	Moderate asthma (requiring medications daily) (230) Age: mean 41 Interview Primary care		Construct ✓	✓	✓		
McColl et al., 1995 UK	Asthma (650) Age: over 18 Self-report and postal response Primary care		Construct ✓ Internal ✓				
McColl et al., 2003 UK	Asthma (4751) Age: mean 48 Postal Primary care	Internal consistency ✓	Construct ✓			✓	
Viramontes and O'Brien (1994) Canada	Patients with asthma, chronic bronchitis and emphysema (102) Age: mean 62 Self-reported but interview administered in patient's own homes		Construct ✓				
Ware et al., 1998 USA	Participants in RCT of asthma medication (142) Age: mean 39 Self report Out patients	Test re-test ✓	Construct ✓	✓	✓	✓	
SF-12		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Franic et al., 2005 USA	Asthma (46) Age: mean 46 Self-report Primary care (pharmacies)		Construct ✓	✓			
Garratt et al., 2000 UK	Patients with asthma (394) Age: adults Postal Primary care	Internal consistency ✓	Construct ✓	✓	✓	✓	
Magid et al., 2004 USA	Patients with asthma (1406) Age: mean 35.9 Postal		✓				

c) EuroQol-EQ-5D

One evaluation was identified where the EQ-5D was the principal instrument (Hazell et al., 2003) and three concurrent evaluations (Franic et al., 2005; Garratt et al., 2000; McColl et al., 2003).

Reliability
No evidence reported.

Validity
Age
The EQ-5D index and VAS scores decreased significantly with age with moderate correlations (-0.41; -0.34) (Hazell et al., 2003).

Health status
Hazell et al., (2003) reported a study of the ability of the EQ-5D to discriminate patients with respiratory disease. A postal survey including the ED-5D and a respiratory questionnaire identifying patient with symptoms associated with obstructive airways disease. The survey was posted to all patents identified from a primary care practice in the UK (10,471) and those with self-reported respiratory symptoms were included in the analysis (6828, with 5944 questionnaire computable). The EQ-5D index and VAS scores were significantly lower for those with respiratory symptoms compared to those without.

Patient–reported health instruments
The validity of the EuroQol has been evaluated in a concurrent evaluation with the Newcastle Asthma Symptoms Questionnaire, AQLQ and SF-12 (Garratt et al., 2000). Correlations between the EuroQol and NASQ, SF-12 and AQLQ were moderate and according to hypotheses. The EQ-5D and SF-12 correlations were of similar magnitude to the other instruments in this evaluation. The EQ-5D correlated strongly with the ACQ index -0.72, VAS -0.56 (Franic et al., 2005).

Generic health status
The EQ-5D score had hypothesised correlations (greater than 0.60) for the SF-36 PF, BP, GHP, VT and SF domains (McColl et al., 2003).

Responsiveness
Responsiveness was examined using a patient-reported health transition question and results expressed with SRM's in Garratt et al., (2000). The EQ-5D was responsive with a small SRM but other instruments (AQLQ, NASQ and SF-12 PCS had larger SRM's in this evaluation.

Acceptability
Garratt et al., (2000) compared the number of missing data for different instruments and reported the EQ-5D to have 96% of the scale score computable. 87% of responses for the EQ-5D were computable in a postal survey (6828) with the highest proportion of missing values for the self-care domain (5.7%); anxiety/depression (4.4%); usual activities (4.3%); pain (4.1%); mobility (3.9%). The VAS though had a greater proportion of missing responses (6.3%) (Hazell et al., 2003).

McColl et al., (2003) hypothesised that responses would be higher and quicker when asthma- specific instruments (AQLQ, NASQ) were presented before generic instruments (EQ-5D, SF-36). No order effect was found for versions for response rates or response speed.

Feasibility
No evidence reported.

Table 4.3: Evaluative studies relating to the EQ-5D when completed by patients with asthma

Study/ County	Population	Measurement properties					
EQ-5D		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Franic et al., 2005 USA	Asthma (46) Age: mean 46 Self-report Primary care (pharmacies)		Construct ✓	✓			
Garratt et al., 2000 UK	Patients with asthma (394) Age: adults Postal Primary care		Construct ✓	✓	✓	✓	
Hazell et al., 2003 UK	Asthma related symptoms (5944) Age: mean 48 Postal survey Primary care practice		Construct ✓			✓	
McColl et al., 2003 UK	Asthma (4751) Age: mean 48 Postal Primary care		Construct ✓			✓	

d) Health Utilities Index (HUI)

Validity
Three studies include the HUI in evaluations that focused on the performance of the AQLQ and ACQ (Juniper) following completion by patients with asthma (Franic et al., 2005; Leidy and Coughlin 1998a; Leidy 1998b). As hypothesised, moderate correlations were reported between the AQLQ and the HUI (range 0.40 AQLQ Emotional to 0.60 AQLQ Activities). The item content of the HUI emphasises functional aspects of quality of life, and stronger correlations with the AQLQ Activity limitations domain were as expected (Leidy and Coughlin1998; Leidy 1999b).

Correlations between the ACQ and HUI total was -0.50, with correlations greater than 0.60 for Ambulation, Pain; small correlations for Speech, Dexterity and Cognition and

no correlation for Emotion, Hearing and Vision as would be expected (Franic et al., 2005)

e) Sickness Impact Profile (SIP)

Validity

Three studies include the SIP in evaluations that focus on the performance of the AQLQ-Juniper (Juniper et al., 1993; Rowe and Oxman 1993) or the MAQLQ (Marks et al., 1993) following completion by patients with asthma. This evidence is therefore detailed in chapter 4.2: Asthma-specific instruments.

In summary, the MAQLQ had small correlations with the SIP Total (0.18) and no correlation with the SIP Psychosocial component (-0.01) (Marks et al., 1993). The AQLQ had lower than predicted correlations with the SIP Psychosocial (Rowe 1993) and AQLQ and the SIP correlations were lower than the RAND (Juniper et al., 1993).

RESULTS: ASTHMA-SPECIFIC PATIENT REPORTED HEALTH INSTRUMENTS:

Nine asthma-specific instruments were included in the review. Full details of the development, domains and scoring methods are detailed in Tables 4.4 and 4.5.

The following instruments measurement properties are reported:

a) Asthma Quality of Life Questionnaire (AQLQ)
b) MiniAQLQ
c) AQLQ(S)
d) Acute AQLQ
e) Asthma Control Questionnaire
f) Asthma Control Diary
g) Marks Asthma Quality of Life Questionnaire (MAQLQ)
h) Living With Asthma Questionnaire
i) St. Georges Respiratory Questionnaire

Asthma Quality of Life Questionnaire(s) (Juniper)

The conceptual underpinning of the Asthma Quality of Life Questionnaire(s) developed by Juniper et al., (1993) adopts a functional impairment approach to measurement.

a) Asthma Quality of Life Questionnaire (AQLQ)

The Asthma Quality of Life Questionnaire (AQLQ) was developed in Canada for evaluating health-related impairment of quality of life in adults with asthma in clinical trials (Juniper et al., 1992). The instrument addresses symptoms, emotional function, activity limitations and environmental stimuli.

Instrument content was derived from existing generic instruments; literature review; experiences of patients with chronic airflow limitation; expert consensus; and unstructured interviews with six patients with asthma. From this, 152 items were considered important and an item reduction questionnaire was developed and interview administered to 150 patients (18-70 years) with asthma. Patients were asked which of the 152 items were they affected by in the past year and indicate the importance on a five point Likert scale from ‘not very important’ to ‘extremely important’. The items chosen most frequently and labelled most important were included in the questionnaire. A total of 32 items were included within four domains of symptoms (12 items), emotional function (5 items), exposure to environmental stimuli (4 items) and activity limitations (11 items) were included. For the activity domain, there was a wide range of activities reported by patients during the item reduction phase and the final version thus included five individualised questions relating to activities which patients identified as being problematic (activities offered to aid recall) and a further 6 questions relating to non-specific activities. The time to recall was suggested as two weeks. A seven point Likert scale (1 indicating maximal impairment and 7 no impairment) was developed for responses and scoring is conducted using the mean score per item and domain and an aggregated overall quality of life score.

The instrument underwent further pre-testing to examine face and content validity and acceptability to patients. Thirty patients were interviewed and time to administer was recorded as well as patient feedback about wording and what they understood each question to be asking. The questionnaire was then modified for self-report and then administered to five other patients and no further modifications were considered necessary. Both the interview administered and self-report format took a maximum of 15 minutes to complete.

b) MiniAQLQ

A shorter version of the AQLQ was developed for greater efficiency (Juniper et al., 1999a). Item-total correlations were examined in previously collected data and correlations greater than 0.70 were considered evidence of similar items and combined resulting in 26 items from the original 32. Further analysis of the original AQLQ item reduction questionnaire (Juniper et al., 1992) resulted in exclusion of those items which had the lowest impact for frequency and importance. The final questionnaire was reduced to 15 items, Symptoms (5 items), Emotions (3 items), Environment (3 items) and for Activities (4 items). Generic items were included for the Activity domain thus removing the individualized questions. Nine patients were involved in the pre-testing of the questionnaire and minor wording and modifications were made. The final version included the seven point Likert scale, 2 week recall and took 7-10 minutes to complete by self-report at baseline and 3-5 minutes at follow-up.

c) Standardised Asthma Quality of Life Questionnaire (AQLQ(S))

A standardised version of the original AQLQ was developed (Juniper et al., 1999b) in which five generic activities replaced the individualised approach used in the AQLQ. The items were selected based on the impact and frequency of reporting activities in the item reduction questionnaire (Juniper et al., 1992) and classified as 'strenuous', 'moderate', 'social', 'work related' and 'sleeping'. The wording of the revised, standardised instrument was pre-tested in ten patients with asthma. Scoring and recall period remained the same as the AQLQ.

The questionnaire was administered to forty patients and the classifications of 'activities' were examined in relation to patients self reported activities (as per original instrument). The classifications of activities were considered to represent the patient-specific activities chosen by the patients.

d) Acute Asthma Quality of Life Questionnaire (Acute AQLQ)

The Acute AQLQ is a modification of the AQLQ with the intention of being specific to patients experiencing an acute severe asthma attack (Juniper et al., 2004). The 32 items from the AQLQ were examined and those considered not relevant or unlikely to change to patients during an acute exacerbation were excluded. The final instrument contains two domains: Symptoms (6 items) and Emotions (5 items) and scoring the same as other AQLQ instruments using a seven point scale. The format was tested with ten patients.

e) Asthma Control Questionnaire (ACQ)

Item generation for this patient-reported symptom focused questionnaire was informed by treatment goals from clinical guidelines, reviewing other asthma questionnaires and a postal survey of asthma clinicians to rank symptoms presented for content. The final instrument includes seven items relating to awakening at night by symptoms; waking in the morning with symptoms; limitations in activities; dyspnoea; wheeze and β_2-agonist use. One item, FEV_1 is clinician assessed (Juniper 1999c). Patient's responses are on a 7 point Likert scale and evaluation for the last 7 days. Scoring of the ACQ is computed as the mean of the 7 items with 0= well controlled and 6= poorly controlled.

f) Asthma Control Diary (ACD)

The Asthma Control Diary is modified form the Asthma Control Questionnaire for daily completion using PEF instead of FEV_1.

g) Asthma Quality of Life Questionnaire (Marks) (MAQLQ)

The initial items for the instrument were derived from analysis of results from a focus group with eight patients with a wide range of asthma severity; from patients participating in an asthma education programme and clinical experience of the developers (Marks et al., 1992). Initial testing was with 283 patients using principal components analysis. Further evaluation of measurement properties was conducted with seventy-seven patients with stable asthma and another sample of patients with unstable asthma (n=42).

The instrument measures the effect of the disease with negative statements (not at all; mildly; moderately; severely; very severely). Conceptually, the AQLQ is underpinned by a limitation and negative approach of the impact of asthma on the individual.

Content validity was examined empirically using principle components analysis. Items were excluded is they had highly skewed distribution; missing values; or low loadings. Principal components analysis gave a six component solution and items most strongly correlated with each component were labelled Breathlessness, Concerns, Mood, Social, Cough and Control. Item-total correlation ranged from 0.13 to 0.72 with correlations less than 0.5 for Cough and Control. These items were deleted based on weak correlation and being considered unrelated to quality of life. The final instrument contained four domains (Breathlessness, Concerns, Mood and Social) and a total of 20 items. Each item contributes to the total scale and domain scores are calculable.

g.i) Modified Marks Asthma Quality of Life Questionnaire

In the original instrument developed by Marks there were two items related to activities which were combined to a single item. Adams et al., (2000) extended the number of items to 22 in the instrument to allow for different responses for this 'activity' question. In addition, a seven point Likert scale was used with the intention of increasing reliability.

h) Living With Asthma Questionnaire (LWAQ)

The Living With Asthma Questionnaire was developed by Hyland (1991, UK) using a comprehensive methodology. Six focus groups were conducted, four with patients and two with the general population (under-graduates). Eleven themes (classified as domains) were identified from content analysis and further items and domains were developed following analysis. The questionnaire was further tested and refined in three phases with a total of 656 patients from primary care. Psychometric testing and item reduction included principal factor analysis, item variability analysis and patient comment. The final questionnaire contained eleven domains and 68 items with a 3 point response format to statements: 'untrue of me', 'slightly true of me', 'very true of me' with an additional option of 'not applicable'. Hyland (1991) attempted to compensate for acquiescence bias by ensuring there were both negative and positive statements. Both negative and positive statements were included in the questionnaire with a third of statements negative. Factor analysis indicated a unifactorial solution.

The final instrument has five constructs: Avoidance, Distress, Preoccupation, Colds and Activities with eleven domains and 68 items. Mean scale scores are obtained with 2 indicating poor quality of life and 0 best.

h.i) ms-LWAQ

Modifications were made to the LWAQ by Reid et al., (1999) for use with Americans. The instrument has twenty-seven items and five subscales: Consequences (10 items); Affect (6 items); Leisure (4 items); Seriousness (5 items) and Drugs (2 items). Scoring is the same as the LWAQ but with different wording of responses.

i) St. George's Respiratory Questionnaire (SGRQ)

The SGRQ was developed in the UK to measure the impact of asthma and chronic obstructive pulmonary disease (COPD) from a patient perspective. There are two parts of the instrument. Part 1 is concerned with symptoms focusing on the severity, frequency and effect of respiratory symptoms over the last year and responses are obtained with a 5 point Likert scale. Part 2 includes two domains: Activity limitations and social and psychological impact and focuses on the patient's current state with True or False responses. Three components scores are calculated and a total score. All items have empirically derived weights and normative data are available. Scoring algorithms and calculators are available from the developers. Scores are expressed as the percentage of overall impairment with 100 equaling to worst possible health and zero the best.

Items were initially derived from studies with adult patients with asthma examining distress ratings relating to symptoms and the impacts of asthma (Quirk and Jones 1990) and the influence of demographic and disease factors with the degree of distress (Quirk et al., 1991). Empirical weights were obtained from one hundred and forty patients with asthma (Quirk 1991). Further analysis of previously derived weights were compared with patients with COPD with thirty-six patients (mean age 66) (Jones et al., 1991) and no significant differences between the item weights from the asthma patients (Quirk et al., 1991) and COPD patients.

ASTHMA-SPECIFIC INSTRUMENTS:

Table 4.4: Asthma-specific patient-reported health instruments

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Asthma Control Questionnaire (ACQ) (Junipers)	*7 Symptoms (1 clinician assessed):* Sleep related (2); breathlessness; wheeze; activity limitations; use of bronchodilators *(FEV_1 % of predicted clinician assessed)*	7 point Likert	Mean score of all items 0=well controlled, 6=extremely poorly controlled	Self- report and clinician assessed (one item)
Asthma Control Diary (ACD) (Junipers') ***Modified ACQ***	*7 Symptoms:* Sleep related (2); breathlessness; wheeze; activity limitations; use of bronchodilators; morning peak expiratory flow rate (PEFR)	7 point Likert	Mean score of all items 0=well controlled, 6=extremely poorly controlled	Self- report
Asthma Quality of Life Questionnaire (AQLQ) (Junipers')	*4 domains/32 items* 1. Symptoms (12) 2. Emotions (5) 3. Environment (4) 4. Activities (11 including 5 individualised questions)	7 point Likert	Summation and domain score Mean score of all items Index: 1 = maximal impairment , 7 = no impairment	Interviewer- and self-administered format 10 minutes to complete at the first visit and 5 minutes at follow-up.
Standardised Asthma Quality of Life Questionnaire (AQLQ(S)) (Junipers')	*4 domains/32 items* 1. Symptoms (12) 2. Emotions (5) 3. Environment (4) 4. Activities (11 including 5 standardised activity classifications)	7 point Likert	Summation and domain score Mean score of all items Index: 1 = maximal impairment , 7 = no impairment	Interviewer- and self-administered format 10 minutes to complete at the first visit and 5 minutes at follow-up.
Mini Asthma Quality of Life Questionnaire (MiniAQLQ) (Junipers')	*4 domains/15 items* 1. Symptoms (5) 2. Emotions (3) 3. Environment (3) 4. Activities (4 all generic)	7 point Likert	Summation and domain score Mean score of all items Index: 1 = maximal impairment , 7 = no impairment	Self administered 7-10 minutes to complete at baseline and 3-5 minutes at follow-up
Acute Asthma Quality of Life Questionnaire (Acute AQLQ) Junipers	*2 domains/11 items* 1. Symptoms (6) 2. Emotions (5)	7 point Likert	Summation and domain score	

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Marks Asthma Quality of Life Questionnaire (MAQLQ)	*4 domains (20 items)* Breathlessness Concerns Mood Social	5 point Likert Modified: 7 point Likert	Total and domain score	
Living with Asthma Questionnaire (LWAQ)	*5 constructs/ 11 domains/68 items* *Constructs:* Avoidance, Distress, Preoccupation, Colds, Activities *Domains:* 1. Social/leisure (6) 2. Sports (3) 3. Holidays (3) 4. Sleep (4) 5. Work and other activities (6) 6. Colds (5) 7. Mobility (6) 8. Effects on others (5) 9. Medication usage (6) 10.Sex (1) 11.Dysphoric states and attitudes (23)	3 point scale with additional option of n/a	Construct and domain scores	Self-administered 10 to 20 minutes completion
St. George's Respiratory Questionnaire (SGRQ)	*Two parts; Domains (3)/17 items* Part 1: Symptom scores (8) Part 2: Activity and Impact (9)	Part 1: 5 point Likert Part 2: True or False	Weighted scoring Total and domain scores Percentage of overall impairment 0=best possible health and 100 worse	Self- report but recommended interview administered 8- 15 minutes to complete

Table 4.5: Summary of asthma-specific instruments: health status domains ***(after Fitzpatrick et al., 1998)***

Instrument	***Instrument domains***								
	Physical function	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**	**Treatment satisfaction**
AQLQ		x		x			x		
MiniAQLQ		x		x			x		
AQLQ(S)		x		x			x		
Acute AQLQ		x		x					
ACQ)		x							
ACD		x							
MAQLQ		x		x	x			x	
LWAQ		x		x	x			x	
SGRQ	x	x		x	x		x		x

RESULTS: AQLQ Junipers

a) Asthma Quality of Life Questionnaire (AQLQ-Juniper)

Seventeen studies were identified which evaluated the AQLQ, five were in concurrent evaluations (Caro et al., 2001; Cook et al., 1993; Juniper et al., 1992, 1993, 1994, 1999a,b, 2000, 2001; Leidy and Coughlin 1998a, Leidy et al., 1998b; Mancuso et al., 2001; Mancuso and Peterson 2004; Orr et al., 2003; Ware et al., 2002; Wywrich et al., 2002). Typically studies included patient samples with an average age around forty years. One study evaluated the instrument with patients participating in a RCT of asthma treatment regimes (Cook 1993) and the others with patients with a diagnosis of asthma. Two studies used a postal survey and the other studies interview administered the self-reported instruments. Two studies were conducted in the UK (McColl et al., 2003; Orr et al., 2003).

Reliability
High levels of test re-test reliability have been reported for the AQLQ with values exceeding 0.90 in five studies for the Summary score and greater than 0.80 for component scores (Juniper et al., 1993, 1999a, 1999b; Rowe et al., 1993; Leidy and Coughlin 1998a). Higher levels of test-retest reliability were reported for the AQLQ in comparison to several generic measures (SF-36, SG, and RS) completed during a comparative evaluation of measurement performance (Juniper et al., 2001).

ICC values greater than 0.70 have been reported across all domains of the AQLQ with the exception of the Environment domain (ICC 0.67) (Leidy and Coughlin 1998a)). Different order of administration of patient-reported health instruments was evaluated in McColl et al., (2003) with version 1 containing asthma-specific instruments first and in version 2 generic instruments presented first. Internal consistencies were in excess of 0.90 for the AQLQ domains with the exception of the Environment domain (0.80) for version 1 and Emotional (0.89) and Environment (0.79) for version 2.

High levels of internal consistency reliability have been reported (Juniper et al., 1999a; Leidy and Coughlin1998a, b; Wyrwich et al., 2002): alpha values greater than 0.90 have been reported for the Summary score.

Interscale correlation coefficients among AQLQ sub-scales and between each sub-scale and the Summary scores were greater than 0.50 (Leidy et al., 1998b). Lower correlations were reported between the Environmental and Emotional domains, but Summary score, Symptoms and Activity were higher.

Validity
Evidence of validity was reported in eight studies (Leidy and Coughlin 1998a, Leidy et al., 1998b; Juniper et al., 1993, 1999a, b, 2000, 2001; Rowe et al., 1993).

Socio-demographic variables
There was no relationship between age, overall score and sub-scales reported in Leidy and Coughlin (1998a). They did however report statistically significant worse HRQoL in women for Overall score, Activity and Environmental domains. Further evidence of the instruments ability to discriminate was reported in Leidy et al., (1998b) with men reporting better HRQoL in the activity and symptom domains. Furthermore, effects

for race were statistically significant for all domains and overall score with African Americans reporting poorer HRQoL (Leidy et al., 1998b). Education effects were statistically significant for Summary score and all domains. (Leidy and Coughlin 1998a)

Health status
Overall and domain scores on the AQLQ (with the exception of the Activity domain) were lower when the AQLQ was presented first (version 1 McColl et al., 2003) but less than 0.5.

Asthma-specific patient-reported health instruments
The Summary score discriminated between asthma severity groups using Physician Severity Rating scheme and Asthma Disease Severity Scale (ADSS) as a reference criterion (Leidy and Coughlin 1998a). The Summary score, Activity and Symptom domains differentiated patients with different disease severity in Rowe (1993). Stronger correlations between AQLQ and patient reported Symptom score and Global Assessment were reported seven to 10 days following treatment in an ED. All correlations were in accordance with hypotheses at baseline and follow-up (Rowe et al., 1993).

Moderate levels of correlation were reported between the AQLQ Summary and the ACQ (0.76) and ACD (0.75) (Juniper et al., 2000). Correlations between the NASQ and the AQLQ were greater than 0.60 for all domains (McColl et al., 2003). The AQLQ was moderately correlated with the ACQ with the lowest coefficient for the Environment domain (0.55) (Franic et al., 2005).

Generic patient-reported heath instruments
Several studies have reported moderate levels of correlation between the AQLQ domain scores and the SF-36 PCS (Juniper et al., 1999a,b; Juniper 2000, 2001; Lee et al., 2003; Mancuso et al., 2001); smaller correlations have been reported between the AQLQ Activity domain and the SF-36 MCS (Juniper et al., 1999a,b) and AQLQ and SF-36 RE (Lee et al., 2003). Further hypothesises correlations (greater or less than 0.60) were reported for similar and different traits with a slight trend for stronger correlations between instruments when specific measures were administered first (McColl et al., 2003) between the AQLQ and SF-36.

Correlations between the AQLQ and RAND physical and emotional domains that had hypothesised associations were moderate For the SIP although statistically significant hypothesised relationships were found; these were lower than for the RAND (Juniper et al., 1993). Correlations between the AQLQ and the SIP domains that had hypothesised associations were moderate but lower correlations were found for the AQLQ domains and SIP-psych (Rowe and Oxman1993).

Moderate to large correlations were reported in accordance with hypotheses with the HUI and instrument domains (Leidy 1998a). Correlations between Summary score and domains and HUI, SF-36 Physical and Mental components and Cantril's Ladder (a global QoL measure) with a population of low income African Americans and Caucasians although in the expected direction, were small to moderate (Leidy et al., 1998b).

Juniper et al., (2001) reported in a concurrent evaluation of the AQLQ, SG, RS and SF-36, higher correlations for the AQLQ and RS with moderate correlation with SF-36 and SG (0.48 to 0.53).

Respiratory Function
Several authors have reported small to moderate levels of correlation between the AQLQ and a range of measures of lung function (Leidy et al., 1998b; Juniper et al., 1999[a, b]; Juniper et al., 2000).

Responsiveness
Eight studies reported evidence of responsiveness (Juniper et al., 1993, 1994, 2001; Mancuso et al., 2001; Mancuso and Peterson 2004; Orr et al., 2003; Rowe and Oxman 1993; Wyrwich et al., 2002).

Following completion in a clinical trial of asthma medications the AQLQ detected within-subject change and group change (responsiveness index 0.64) (Juniper et al., 1993). Moderate to strong levels of correlation between change scores were reported for change in AQLQ score and change in clinical asthma based on changes in respiratory function and frequency of symptoms; change in generic quality of life score (SIP) and Asthma global ratings of change (Juniper et al., 1993).

Orr et al., (2003) evaluated changes in patients quality of life using the AQLQ related to changes in respiratory function during a four week treatment programme. Laboratory measures included FEV_1, Bronchial hyper-responsiveness (BHR) using Methacholine bronchial challenge (Methacholine $PD_{20)}$ and patient administered Peak expiratory flow measurement (PEF). Self-reported measures included AQLQ and a total daily symptom score. Clinically important differences in AQLQ scores (defined a priori by authors as 0.5) were reported post intervention for summary and domains scores, with the exception of the Activity domain (change in score for this domain reached statistical significance only).

All domains were responsive to change in terms of agreement with patient-reported change in condition in patients presenting for assessment and treatment in an Emergency department (0.68 to 0.78) with the exception of the Environmental domain (0.44) (Rowe and Oxman (1993).

Several authors have reported greater levels of responsiveness for the AQLQ when directly compared to the responsiveness of generic instruments in concurrent evaluations (Juniper et al., 2000, 2001; Lee et al., 2003; Mancuso 2001; Mancuso and Peterson 2004).

Mancuso et al., (2001) evaluated the discriminative ability of the AQLQ and SF-36 using ROC curve analysis with the patient's perception of disease activity as the external criterion and report higher ranked curves for AQLQ than the SF-36. In a further study by Mancuso and Peterson (2004), similar results were reported with the AQLQ ranking higher than the SF-36. Juniper et al., (2001) reported the AQLQ had the highest responsiveness index (1.35) in comparison to the SG, RS and SF-36. Further evidence of the AQLQ being more responsive than the SF-36 is reported in Lee et al., (2003) with a larger effect size and SRM (1.26; 1.17) than for all SF-36

domains. Moderate correlation was reported between changes in the AQLQ scores and SF-36 with a range of 0.20 for RE to PF 0.62 (Lee et al., 2003).

Moderate levels of correlation between change scores for the AQLQ, ACQ and ACD were reported at 9 weeks; only small levels of correlation between change scores for the AQLQ and respiratory function, peak expiratory flow and medication usage were reported (Juniper et al., 2002).

Furthermore, the SF-36 was not as responsive as the AQLQ with ES per domain lower than for the AQLQ (Domains: Large ES: RP, PF; Medium ES GH, VT, SF and small BP, MH and RE Lee 2003). SRM's for the SF-36 domains were lower than for the AQLQ (0.92 to 0.29 vs. 1.17).

Interpretation

Several studies have contributed to furthering interpretation of change in score for the AQLQ, and in determining a minimal important difference (improvement or deterioration) in scores (Juniper et al., 1994; Rowe and Oxman 1993; Wyrwich et al., 2002, 2003).

A small study was carried out that identified minimally important change scores on the AQLQ by use of a patient rating scale of change (Juniper et al.,1994). A change in score of 0.5 on the seven point AQLQ indicated minimal important difference and a change of 1.0 suggestive of a moderate change. The authors acknowledge the small sample size and number of patients experiencing a large change to be able to be confident of the change in score for this group.

Rowe and Oxman (1993) using a patient-reported change scale (0=no change; 1-3=minimal change [MID]; 4-5 moderate change; 6-7= substantial change) to determine AQLQ change scores in a group of patients visiting an emergency department for treatment and at a 2 week follow-up., MID for the AQLQ was reported as 0.51 which is consistent with other reports.

Wyrwich et al., (2002) explored the relationship between the MID (AQLQ) and the standard error of measurement (SEM) and reported evidence to support the use of one SEM to identify important individual change in HRQoL measures supported by weighted kappa values (0.88-0.93). Values of one SEM were computed using the baseline SD and reliability estimates with Activity=4.43; Symptoms= 4.18; Emotional= 3.04; Environmental 2.89. Further computation for SEM per item values: Activity= 0.40; Symptoms=0.35; Emotional=0.6

Expert consensus

Wywrich et al., (2003) provide a report of an expert consensus process utilising a modified RAND method. The following procedure was carried out:
-systematic review of the literature;
-recruitment of healthcare professional experts and researchers for consultation;
-Delphi consensus technique and a subsequent meeting to achieve consensus and formulate recommendations.

The panel defined a clinically important change (CID) as '*what the physician found important in the treatment of an individual patient even if the CID did not necessarily*

lead to a change in the patients therapy'. A Small change was recommended between 5 and 12; Moderate change 8 to 24 and Large change as 12 to 33 (overlap accounts for different domain CIDs).

Wyrwich et al., (2003) attributed the differences found between Juniper (1994) study of determining the MID as a result of the approaches used to define MID and CID. Minimal important difference as defined by Juniper refers to patients perceptions of change as opposed to clinically importance changed defined by experts.

Precision

Four studies examined precision (Garratt et al., 2000; Mancuso et al., 2001; McColl et al., 2003; Wyrwich et al., 2002) and reported normal response distributions with no evidence of floor or ceiling effects.

Acceptability and Feasibility

Three studies assessed different aspects of acceptability of the AQLQ (Caro et al., 2001; Cook et al., 1993; Garratt et al., 2000).

The concordance of responses on electronic versus paper versions was also compared in Caro et al., (2001). Patients completed instruments two hours apart; the order of presentation was alternated. A high degree of concordance for patients' scores for overall score (0.99) and domain scores (range 0.97 to 0.98) was reported.

Concordance of responses was examined for interview vs. self-administration of the AQLQ where patients were randomized to receive self-administered questionnaire followed by interview-administered two weeks later (Cook et al., 1993). The self-administered approach produced a higher percentage of item endorsement and impact than the interview method. The ICCs for endorsement 0.84 and for total impact 0.93 indicating that both instrument administrations were similar.

McColl et al., (2003) hypothesised that responses would be higher and quicker when asthma- specific instruments (AQLQ, NASQ) were presented before generic instruments (EQ-5D, SF-36). No order effect was found for versions for response rates or response speed.

In a concurrent evaluation including AQLQ and AQLQ(S) Garratt et al., (2000) reported more missing data for the AQLQ individualized Activity questions.

Feasibility

Completion times for paper and electronic versions of the AQLQ were compared (Caro et al., 2001). Similar administration/completion times were reported: 12 (electronic) vs. 11 minutes (paper). Four (6%) of patients indicated that they had no preference for different versions. 49 (77%) expressed a preference for the electronic version and found it easy to use (this includes preferences for SF-36 combined).

b) MiniAQLQ (Juniper)

Four studies were identified which evaluated the MiniAQLQ (Baghi et al., 2004; Juniper et al., 1999a; Magid et al., 2004; Pinnock et al., 2004). Two studies used postal surveys; one used an online method of administration and one interview administered. One study was conducted with UK participants (Pinnock et al., 2004).

Reliability

Test re-test reliability was reported in two studies (Baghi 2004; Juniper 1999a) with the ICCs for the MiniAQLQ reported as consistently lower than the AQLQ in Juniper et al., (1999a). Levels greater than 0.90 for individual comparison were reported in Baghi et al., (2004).

Two studies (Juniper et al., 1999a; Baghi et al., 2004) reported evidence of internal consistency with alpha levels greater than 0.70 with the exception of the Environment domain (pre-test) with an alpha of 0.65 (Baghi et al., 2004).

Validity

Evidence of internal and construct validity of the MiniAQLQ is supported by four studies (Baghi et al., 2004; Franic et al., 2005; Juniper et al., 1999a: Magid et al., 2004).

Health service utilisation

In a prospective study (Magid et al., 2004), patients with a low baseline score (MiniAQLQ) were more likely to have an ED visit and subsequent asthma related healthcare utilization. Multivariate analysis adjusted for sociodemographic and clinical factors, the MiniAQLQ was predictive of ED visit (OR 1.34; 95% CI 1.18 to 1.52).

Internal validity

Baghi et al., (2004) investigated internal validity using principal components analysis pre and post-test of the effectiveness of a web-based intervention for self management. The analysis of the fifteen items extracted 4 factors which accounted for 69% of the variance for pre-test and 76% for post-test providing further empirical evidence that the 15 items are measuring the 4 conceptual constructs within the instrument.

Asthma-specific patient-reported health instruments

The MiniAQLQ scores were similar for the AQLQ for Symptoms and Emotional function but slightly lower for the Environmental and Activity domains. Overall correlation between the two instruments and Symptoms, Environment and Emotional domains were high (0.80) as hypothesised. There was moderate correlation between the Activity domains of both instruments which reflects the differing methods of identifying important activities (Juniper et al., 1999a).

The MiniAQLQ Environment was moderately correlated with the ACQ (-0.55) and Symptoms -0.83 (Franic et al., 2005).

Responsiveness

Two studies evaluated responsiveness (Baghi et al., 2004; Juniper et al., 1999a).

The MiniAQLQ did not correlate as well as the AQLQ with changes overtime with the ACQ and SF-36 physical domain suggesting that the MiniAQLQ may not be as responsive as the AQLQ (Juniper et al., 1999a).

Baghi et al., (2004) reported statistically significant change in scores pre and post testing of the MiniAQLQ following evaluation of a web-based intervention for asthma self management.

Precision

No evidence reported.

Acceptability

Completion errors and response rates of postal MiniAQLQ were compared to interview administered in patients recruited from a primary care practice in the UK (Pinnock et al., 2004). Instruction sheets were provided for guidance. Ninety-eight percent response-rates for the postal questionnaire were reported and of these 10% contained one or more missing responses. Question 15 (work related activities) was the most common error where non-workers considered this question not applicable to them. There were no completion errors for the interview administered method.

Feasibility

No evidence reported.

c) Standardised Asthma Quality of Life Questionnaire (AQLQ(S)) (Junipers')

Two studies were identified which evaluated the AQLQ(S) (Garratt et al., 2000; Juniper et al., 1999b). Garratt et al., (2000) evaluated the AQLQ(S) and other instruments (SF-12; NASQ; EQ-5D) with participants of a RCT of the effectiveness of evidenced based guidelines by means of a postal survey; Juniper et al., (1999b) administered the questionnaire in an out-patient setting.

Reliability

High levels of internal consistency were reported in one study (Garratt et al., 2000) with alphas ranging from 0.81 for Environment to 0.96 for Symptoms.

High levels of test re-test reliability are reported for the Summary score and Activities and are similar to the AQLQ (Juniper et al., 1999b).

Validity

Two studies provide evidence of validity (Garratt et al., 2000; Juniper et al., 1999b).

Health status
The AQLQ(S) discriminated between smokers and non-smokers with the exception of Activities and Environment domains in Garratt et al., (2000).

Asthma-specific patient-reported health instruments

The overall correlation between the AQLQ(S) and AQLQ was high as hypothesised but the Activity domain, correlation was weaker than hypothesised. The mean difference in the Activity domain for the AQLQ(S) was higher than for the AQLQ and the overall score slightly higher for the AQLQ(S). These results are indicative of the difference between the instruments in the items within the activity domain. The AQLQ(S) has standardized activities and the AQLQ adopts an individualised approach (Juniper et al., 1999b).

In a concurrent evaluation, Garratt et al., (2000) reported large correlation with the NASQ and AQLQ(S) as hypothesised.

Generic patient- heath instruments
Moderate correlation was reported for the AQLQ(S) and SF-12 PCS and only weak correlation with the MCS (range 0.27 to 0.36). Moderate correlation was reported for AQLQ(S) and EQ-5D (Garratt et al., 2000).

Respiratory function
The AQLQ(S) correlated moderately with clinical indicators of respiratory function.

Responsiveness
Responsiveness was evaluated in a concurrent evaluation reporting SRMs ranging from 0.32 for Environmental exposure to 0.77 for Symptoms which ranked higher in magnitude than for other instruments in this study (NASQ, SF-12, and EQ-5 D) (Garratt et al., 2000).

Precision
No evidence reported

Acceptability
Garratt et al., (2000) reported less missing data with the AQLQ(S) than the AQLQ.

Feasibility
No evidence reported.

d)Acute Asthma Quality of Life Questionnaire (Acute AQLQ)

One study was identified which evaluated the Acute AQLQ (Juniper et al., 2004) in an emergency department setting with patients with acute broncho-constriction.

Reliability

High levels of internal consistency were reported with alphas 0.82 to 0.90 (Juniper et al., 2004).

Validity

There was no correlation with respiratory function (FEV_1, % predicted) and the AQLQ reported in Juniper et al., (2004); moderate correlation was reported with Asthma Symptom Severity (patient-reported).

Responsiveness

The responsiveness of the Acute AQLQ was evaluated in a Randomised Controlled Trial (RCT) of asthma medications and reported a responsiveness index of 2.5 (Juniper et al., 2004). The instrument was able to detect improvement seventy-five minutes after treatment.

Moderate correlation was reported in Juniper et al., (2004) in longitudinal correlation between respiratory function changes (FEV_1, % predicted) and patient-reported Asthma Symptom Severity.

Precision
No evidence reported

Acceptability
No evidence reported

Feasibility
No evidence reported

e) Asthma Control Questionnaire (ACQ)
Two studies were identified which evaluated the ACQ as the principal instrument under study (Juniper et al., 1999c, 2005) and Juniper et al., (2000) used the ACQ in a concurrent evaluation with the ACD.

Reliability
High level of test re-test reliability was reported in Juniper et al., (1999c) (ICC $\geq$0.90).

The ACQ was internally consistent in (Juniper et al., 2005) with levels greater than 0.90.

Concordance of responses between the ACQ and ACD was high with ICC= 0.87 (Juniper 2000) and in the same study higher levels of test-retest reliability (0.90) was reported for the ACQ than the ACD (Juniper et al., 2000).

Validity
Three studies reported evidence of validity (Franic et al., 2005; Juniper et al., 1999c, Juniper et al., 2000).

Asthma severity and medication use
The Global Initiative for Asthma (GINA) guidelines uses four variable to classify asthma severity based on frequency of symptoms, lung function and medication regime. Correlations between the ACQ with a patient-reported version of item 7

relating to peak flow recordings and the GINA guidelines were moderate (0.57) and accorded with hypotheses. The ACQ discriminated between patients with increased usage of quick-relief medication with less asthma control (correlation 0.84) (Franic et al., 2005).

Asthma-specific patient-reported health instruments
The ACQ had hypothesised correlations with AQLQ total score and sub-domains and there was also moderate correlation as predicted with other patient-reported asthma symptoms (Juniper et al., 1999c; 2000).

.
As hypothesised, there was strong association between the scores for the ACQ and ACD but only moderate correlation between the ACQ and respiratory function in Juniper 2000). The ACQ was moderately correlated with the AQLQ with the lowest coefficient for the Environment domain (0.55) with similar results for the MiniAQLQ: range -0.55 for Environment to -0.83 Symptoms (Franic et al., 2005).

Generic patient-reported health instruments
Moderate correlations as hypothesised were reported for the ACQ with SF-36 Physical (0.55), and weak for SF-36 Mental component (0.19) (Juniper et al., 1999c; 2000). Further similar results were observed for the ACQ and SF-12 with correlations -0.76 for PCS and 0.03 MCS with corresponding correlations for similar domains (Franic 2005)**.** Correlations between the ACQ and HUI total was -0.50, with correlations greater than 0.60 for Ambulation, Pain; small correlations for Speech, Dexterity and Cognition and no correlation for Emotion, Hearing and Vision as would be expected. The ACQ correlated strongly with the EQ-5D index -0.72, Vas -0.56 (Franic et al., 2005).

Responsiveness
The ACQ detected change reporting a responsiveness index of 1.35 and moderate correlations between changes in the ACQ and other instruments (AQLQ; ACD) as hypothesised (Juniper et al., 1999c).

Precision
No evidence reported.

Acceptability
No evidence reported.

Feasibility
No evidence reported.

e.i) Shortened version of the ACQ

Reliability
High levels of internal consistency and test-re-test reliability are reported for all shortened versions of the ACQ (Juniper et al., 2005). A high level of concordance was reported for all shortened versions with data from the original ACQ.

Validity
Evidence of construct validity is reported with hypothesised correlations between shortened versions and the MiniAQLQ.

Responsiveness
All versions of the ACQ had hypothesised correlations with the MiniAQLQ in longitudinal analysis of change and all versions detected similar change scores between baseline and 26 weeks (Juniper et al., 2005).

Interpretation
The change in ACQ that was equivalent to a change in MiniAQLQ score of 0.5 was calculated by regressing model (Juniper et al., 2005). The MID for all versions was close to 0.5. Furthermore, changes in all versions were associated with changes in lung function and β_2-agonist use

Precision
No evidence reported.

Acceptability and Feasibility
No evidence reported

f) Asthma Control Diary (ACD)
One study was identified which evaluated the ACD in a concurrent evaluation with the ACQ (Juniper et al., 2000).

Reliability
Concordance of responses between the ACD and ACD was high (ICC=0.87). Reliability was high for test re-test within a four week period and although an acceptable level of ICC (0.86) was achieved it was not as reliable as the ACQ for individual assessment (0.90).

Validity
The ACD had similar correlations with the AQLQ as the ACQ (range 0.52 (Environment) to 0.75 for other AQLQ domains).

Weak to moderate correlation was reported for the ACD and SF-36 components with the Mental component having a weaker correlation (0.31).

Responsiveness
Juniper et al., (2000) compared the responsiveness of the ACD and ACQ and reported responsiveness indexes of similar value but the ACD was less in magnitude than the ACQ.

Table 4.6: Developmental and evaluation studies relating to the AQLQ (Junipers) instruments:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
AQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Caro et al., 2001 Canada	Asthma (68) Age:16-75 Interview but patient completed Out-patients					✓	
Cook et al., 1993 Canada	Asthma participating in a RCT of different methods of administration (Interview administered vs. self-administered (150) Age: mean 39					✓	
Juniper et al., 1992 Canada	Developmental study		Content ✓				
Juniper et al., 1993 Canada	Further development study Patients who were symptomatic or required treatment at least once a week (150) Age: 39-77 Interview Out-patients		Content ✓				
Juniper et al., 1993, 1994 Canada	Patients who reported asthma symptoms at least once per week and hyperesponsiveness to methacholine[o] (39). Age: 16-60 Interview-self reported Out-patients	Test re-test ✓	Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
AQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Juniper et al., 1999a Canada	Development/ adaptation study Patients with symptomatic asthma (40) Age: 18-65 Interview Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			
Juniper et al., 1999b Canada	Patients with current symptoms of asthma (40) Age: 18-65 Self-report Out-patients	Test re-test ✓	Construct ✓	✓			
Juniper et al., 2000 Canada	ACQ score>0.5 (50) Age: Mean 37 Self-report Completion during one week before follow-up appointment	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			
Juniper et al., 2001 Canada	Patients with symptomatic asthma (40) Age: mean 38 Interview administered Out-patients	Test re-test ✓	Construct ✓	✓			
Lee et al., 2003 USA	Participants in a RCT of asthma medication (241) Age: mean 38 Self report-hand held electronic device recording patients responses to the instruments) Out patients		Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
AQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Leidy and Coughlin 1998a USA	Patients attending asthma clinic (161) Age: mean 34.7 (data derived from a study testing the ASUI) Interview-self reported Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓				
Leidy et al., 1998b USA	Patients with self reported diagnosis of asthma; asthma symptoms; low income (112: n=46 African American/AA, n=66 Caucasian/C) Age: mean 33.4 Interview-self reported Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓				
Mancuso et al., 2001 USA	Patients with moderate asthma (requiring medications daily) (230) Age: mean 41 Interview Primary care		Construct ✓	✓	✓		
Mancuso and Peterson 2004 USA	Asthmatics identified for healthcare plan (185) Age: 41 Postal			✓			
McColl et al., 2003 UK	Asthma (4751) Age: mean 48 Postal Primary care	Internal consistency ✓	Construct ✓			✓	
Orr et al., 2003 UK	Uncontrolled asthma patients participating in treatment programme Age: mean 44 Out patients Self report		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
AQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Rowe and Oxman (1993) Canada	Patients who met the American Thoracic Society diagnosis (52) Age: 18-64 Interview administered Emergency department	Test re-test ✓	Construct ✓	✓			
Ware et al., 1998 USA *Only activities questions*	Patients participating in a RCT of asthma medications (142) Age: mean 39 Self reported Out patients	Test re-test ✓	Construct ✓	✓			
Wyrwich et al., 2002 USA	Diagnosis of asthma and/or prescription for asthma medication in the last 2 years (198) Age: mean 37 Interview Out-patients	Internal consistency ✓	Construct	✓			
Mini Asthma Quality of Life Questionnaire (MiniAQLQ) (Junipers')							
Baghi et al., 2004 USA	Patients participating in web based management tools (307) Age: mean 36 Self-report Online	Internal consistency ✓ Test re-test ✓	Construct ✓				
Franic et al., 2005 USA	Asthma (46) Age: mean 46 Self-report Primary care (pharmacies)		Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
MiniAQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Juniper et al., 1999a Canada	Development/ adaptation study Patients with symptomatic asthma (40) Age: 18-65 Interview Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			
Magid et al., 2004 USA	Asthmatics identified from health plan (1406) Age: mean 35.9 Postal		Construct ✓				
Pinnock et al., 2004 UK	Asthma (96) Age: mean 58.5 Postal vs. Interview administered Primary care		Construct ✓			✓	
Standardised Asthma Quality of Life Questionnaire (AQLQ(S)) (Junipers')							
Garratt et al., 2000 UK	Patients participating in a randomised trial assessing the affects of evidence based guidelines (235) Age: 18-60 Postal Primary care	Internal consistency ✓	Construct ✓	✓	✓		
Juniper et al., 1999b Canada	Patients with current symptoms of asthma (40) Age: 18-65 Self-report Out-patients	Test re-test ✓	Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Acute Asthma Quality of Life Questionnaire (Acute AQLQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Juniper et al., 2004 Canada	Patients with acute broncho-constriction (88) Age: 18-70 (RCT) Interview Emergency department	Internal consistency ✓	Construct ✓				
Asthma Control Questionnaire (ACQ) (Junipers')							
Franic et al., 2005 USA	Asthma (46) Age: mean 46 Self-report Primary care (pharmacies)		Construct ✓	✓			
Juniper et al., 1999c Canada	Developmental study ACQ score>0.5 (50) Age: Mean 37 Self-report (one item clinician assessed) Out-patients	Test re-test ✓	Construct ✓	✓			
Juniper et al., 2000 Canada	ACQ score>0.5 (50) Age: Mean 37 Self-report Completion during one week before follow-up appointment	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			
Juniper et al., 2005 Canada	Patients requiring inhaled steroids participating in a RCT comparing different treatments (552) Age: mean 44.7 Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			
Asthma Control Diary							
Juniper et al., 2000 Canada	ACQ score>0.5 (50) Age: Mean 37 Self-report Completion one week before follow-up	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			

g) Marks AQLQ (MAQLQ)

Four studies provide evidence of the measurement performance of the MAQLQ (Gupchup et al., 1997; Katz et al., 1999; Marks et al., 1992, 1993; Ware et al., 1998). Patients with a range of asthma severities were included in these studies and one study evaluated the performance with patients recruited in a RCT of asthma medications (Ware et al., 1998). The average age of the participants was forty years. Two studies used a telephone survey (Gupchup et al., 1997; Katz et al., 1999) and others obtained responses from patients during an out-patient appointment.

Reliability

Five studies reported reliability evidence (Hyland et al., 1992; Gupchup et al., 1997; Katz et al., 1999; Marks et al., 1992, Ware et al., 1998).

High levels of internal consistency (greater than 0.90) were reported for stable and unstable patients in Marks et al., (1992) developmental study. Similarly positive internal consistency was observed for the Total score; Total and domain in Gupchup et al., (1997); Total in Katz et al., (1999). Ware et al., (1998) reported alphas greater than 0.80 for Breathlessness, Social, Concerns and Overall with 0.71 for Moods.

Test re-test reliability results exceeded the recommended 0.70 for group comparison with exception of the Breathlessness domain (ICC 0.61).

Item level analysis

Item-total correlations were greater than 0.40 for each domain (Marks et al., 1992; Gupchup et al., 1997; Katz et al., 1999).

Validity

Four studies reported evidence of validity (Gupchup et al., 1997; Katz et al., 1999; Marks et al., 1992; 1993).

Internal validity

The conceptual framework of the MAQLQ with a four domain structure was empirically supported in factor analysis (Katz et al., 1999) although the overlapping items of two domains were eliminated in these analyses.

Asthma-specific patient-reported health instruments

No evidence was found comparing the MAQLQ with other asthma-specific measures.

Generic patient-reported health instruments

Moderate, hypothesised correlation were reported between the MAQLQ scales and the SF-36 PCS range 0.43 (Emotional impact) to 0.66 (Total); MCS range 0.22 (Physical) to 0.60 (Emotional) (Katz et al., 1999).

Lung function

Moderate correlations were reported for number of medications and MAQLQ total and domain scores (Marks 1992) in patients with unstable asthma. Weak correlations were found for MAQLQ scores and baseline FEV_1 and degree of bronchial hyperesponsiveness. The authors attribute the weak correlation with the physiological measures due to the variability of airflow obstruction over a period of time. Gupchup et al., (1997) also reported significant but weak to moderate correlation between

number of medications and MAQLQ Total and domain scores. Weak correlations were reported for the MAQLQ total and domain scores with FEV_1 (Range 0.06 to -0.17).

Responsiveness
Evidence of responsiveness of the MAQLQ is supported by three studies (Katz et al., 1999; Marks et al., 1993; Ware et al., 1998).

Marks et al., (1993) evaluated the longitudinal validity of the MAQLQ hypothesising moderate correlation with changes in the SIP, patient-reported symptoms, peak flow variability and degree of bronchial hyperesponsiveness at baseline and at follow-up (3/4 months). There were moderate correlations as hypothesised between the MAQLQ total score and Symptoms and degree of bronchial hyperesponsiveness (BR), but weak (non-significant) correlation for SIP and Peak flow variability. Significant moderate correlation was found for MAQLQ Social domain and Symptoms, Peak flow variability and BR. There was no to weak correlation with MAQLQ and SIP (Psychosocial). The MAQLQ was able to detect change in improved patients and identify those who had remained stable and the magnitude of the responsiveness index greater than for other measures.

Responsiveness was evaluated by comparing changes in MAQLQ to external criteria of changed defined for the SF-36 components and Asthma Severity scale by calculating the better and worse group as one standard deviation above or below the mean for the entire group (Katz et al., 1999). All differences were statistically significant demonstrating responsiveness to change in this group of patients.

Ware et al., (1998) evaluated the responsiveness of the MAQLQ compared with the SF-36 expressing results as relative validity (RV) coefficients. The best measure with RV estimates of 1.0 were reported for MAQLQ Breathlessness and Clinician assessed pulmonary function, Chest tightness, Wheeze, Shortness of breath and Overall condition. In this concurrent evaluation, the MAQLQ was more responsive than the SF-36.

Precision
No ceiling or floor effects were reported in Ware et al., (1998).

Acceptability
Gupchup et al., (1997) assessed acceptability of the MAQLQ during a telephone survey by offering a ‘don’t know’ option for each item. No participants chose this option. Ware (1998) reported that 99% of all items were completed by patients.

Feasibility
No evidence reported.

g.i) Modified Asthma Quality of Life Questionnaire-Marks (MAQLQ-M)

One study was identified which modified the MAQLQ and evaluated its performance (Adams et al., 2000).

Reliability

Internal consistency reliability values for the MAQLQ-M for the Total scale exceeded 0.90 (Adams et al., 2000). High levels of Test-retest reliability were found in a two week re-test period in the same study population.

Validity

Internal validity

Factor analysis of the structure of the modified instrument refuted the results from a previous study of the original instrument by Katz et al., (1999) in Adams et al., (2000). A three component solution was reported for Breathlessness, Mood but loadings on one factor only for the Social/ Concerns domain. The authors suggest that the alteration of the response scale from a 5-point Likert to 7 may have accounted for these results.

Healthcare utilisation

Patients who did not have repeated hospital admissions or visits to the ER reported better quality of life as hypothesized.

Generic patient-reported health instruments

Moderate to large correlations with the MAQLQ-M and SF-36 PCS were reported (0.71) and MCS (0.62) (Adams et al., 2000).

Lung function

Several disease reference measures were employed to assess the correlation of MAQLQ-M scores. Stronger associations were reported between patient-reported symptoms (range 0.35 to 0.56) than lung function (range -0.29 to 0.30).

Responsiveness

Small to moderate correlations with changes in MAQLQ-M scores and respiratory function but with stronger correlations with self-reported measures of symptoms were reported in (Adams et al., 2000).

Precision

Adams et al., (2000) reported no floor or ceiling effects of the MAQLQ-M and the distribution of scores was normal.

Acceptability

No evidence reported.

Feasibility

No evidence reported.

Table 4.7: Developmental and evaluation studies relating to the MAQLQ instruments:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
MAQLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Marks et al., 1992 Australia Developmental study	Focus group Patients with a wide range of asthma severity (8) Patients with asthma (283) Age: mean 39 Out-patients		Internal ✓				
	Patients with stable asthma (77) Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓				
	Patients identified from population survey with unstable asthma (87) Population survey	Internal consistency ✓	Construct ✓				
Marks et al., 1993 Australia	Patients attending asthma clinic (44) Age: mean 33 Out-patients			✓			
Gupchup et al., 1997 USA	Patients taking medications for asthma (106) Age: range 18 and over Community-telephone survey	Internal consistency ✓	Construct ✓			✓	
Ware et al., 1998 USA	Patients enrolled in a RCT of asthma medication (142) Age: mean 39.5 Out patients	Internal consistency ✓	Construct ✓	✓	✓	✓	
Katz et al., 1999 USA	Patients selected from physician records (539) Age: mean 39.4 Community-telephone survey	Internal consistency ✓	Construct ✓ Internal ✓	✓			
Modified Marks Asthma Quality of Life Questionnaire (MAQLQ-M)							
Adams et al., 2000 Australia	Patients selected with physiological evidence of asthma (293) Age: mean 42 Out-patients and postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓ Internal ✓	✓			

h) Living With Asthma Questionnaire (LWAQ)

Three studies were identified which evaluated the LWAQ (Hyland 1991; 96; McColl et al., 1995).

Reliability

High ICC levels for test re-test reliability were reported in Hyland (1991) (≥ 0.90).

Validity

Internal validity

Hyland et al., (1996) conducted exploratory factor analysis and two cognitive factors (activities and avoidance) and two emotional (distress and pre-occupational) with a general factor of disease severity were reported.

Generic patient-reported health instruments

The LWAQ and the SF-36 were evaluated concurrently with patients with asthma (McColl et al., 1995) and reported hypothesised correlations with related domains for example Physical functioning (0.70 to 0.80). The Emotional and Mental domains for both instruments were only moderately correlated (0.45 to 0.54). Similar results were reported for the Social functioning domains (0.54 to 0.64)

Responsiveness

No evidence reported.

Precision

No evidence reported.

Acceptability

No evidence reported.

Feasiblity

No evidence reported.

-ms-LWAQ

Reliability

Adequate levels of internal consistency with the exception of 'Drugs construct' with an alpha of 0.40 reported in Reid et al., (1999).

Validity

Healthcare utilisation

All of the sub-scales of the ms-LWAQ were associated with the level of healthcare utilisation as measured by visiting physicians; emergency room visit and hospital in-patient (Reid et al., 1999).

Correlations with SF-36 domains were moderate as hypothesised in Reid et al., (1999) although small correlation was reported for Seriousness and Role Emotional and Pain; and Affect and Pain

Responsiveness

No evidence reported.

Precision
No evidence reported.

Acceptability
No evidence reported.

Feasiblity
No evidence reported.

Table 4.8: Developmental and evaluation studies relating to the Living With Asthma Questionnaire instruments:

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
LWAQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hyland (1991) UK. Developmental study	Patients with a wide range of asthma severity and general population sample Six focus groups		Content ✓				
	Asthma 101, 150, 405, 282 Primary care		Internal ✓				
	Asthma (81) Aged over 18 years Primary care	Test re-test ✓	Construct Internal ✓				
Hyland et al., 1996 UK	Asthma (810) Primary care		Internal ✓				
	Participants of a RCT of two different asthma medications (149) Primary care			✓			
McColl et al., 1995 UK	Asthma (650) Age: over 18 Primary care Self-report and postal response		Construct ✓ Internal ✓				
Ms-LWAQ							
Reid et al., 1999 USA	Asthma (250) Age: 19-83 Primary care	Internal consistency ✓	Construct ✓				

i) St. George's Respiratory Questionnaire

Reliability
Reproducibility was examined with asthmatic patients (40) and patients with COPD (20) with a two week recall period with high levels of test-retest reliability (Jones et al., 1992) (0.91).

Validity
Health status
The SGRQ Symptoms domain discriminated between patients with respiratory symptoms and those without but was weakly correlated with physiological measures, dyspnoea grade, mood state and SIP scores (Jones et al., 1991). Moderate correlations were reported as hypothesised for the Activity and Impact domains and MRC dyspnoea grade, physical function test, psychological functioning and general health. Stronger correlations were reported for the Impact domain and anxiety and depression (Jones et al., 1991).

Generic patient-reported health instruments
For the Total score the highest but moderate correlation was reported for SIP total and dyspnoea, followed by anxiety and depression.

Responsiveness
Smaller than hypothesised but significant correlation was reported in longitudinal analysis (one year) of changes in SGRQ scores and other measures: (SIP; Respiratory function) (Jones et al., 1991).

Precision
No evidence reported.

Acceptability
No evidence reported.

Feasiblity
No evidence reported

Table 4.9: Developmental and evaluation studies relating to the St Georges Respiratory Questionnaire evaluations in asthma

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
SGRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Quirk and Jones (1990) UK	Asthma (40) Age:16-75 Interview but patient completed Out-patients Development study		Content ✓				
Quirk et al., 1991 UK	Asthma (140) Age: mean 44 Interview but patient completed Out-patients Development study: empirical weights		Content ✓				
Jones et al., 1991 UK	Asthma (40); COPD (20) Age: mean 45; 66 Self-completed	Test re-test ✓	Construct ✓	✓			

Other asthma-specific instruments identified from the review.

The following table provides an overview of other records of asthma-specific instruments identified of either newly developed instruments or single study reporting of measurement properties and/or evaluation.

Table 4.10

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
AQ30 and AQ20 Barley et al., 1998 UK	Patients with asthma (90) Age: mean 46 Self report Out-patients		✓	✓	✓		✓	No advantage over the AQ30 over the AQ20. Correlations reported with respiratory function, SGRQ and AQLQ Junipers. AQ30 and AQ20 evaluated in patients with COPD (Quirk 1994)
AQ18 Barley and Jones, 2006 UK	Asthma (144) Self report UK	✓			✓			Rasch analysis of the AQ20. Highlights the usefulness of multiple repeat assessments over time allowing for testing of differential item functioning (DIF)
Asthma Therapy Assessment Questionnaire Vollmer et al., 2002 USA	Asthma		✓					Problems based questionnaire to generate an index of asthma control and relationship with healthcare utilisation
Asthma TYPE Blumenschien and Johannesson (1998) USA	Patients with Asthma (69) Age: mean 40 Out-patients Interview administered		✓					Concurrent evaluation with SF-36 (have put this in concurrent evaluation table and SF-36 table) Moderate correlation for all domains except Allergy index

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Integrated Therapeutics Group Asthma Short Form (ITG-ASF Bayliss et al., 2000 USA	Development study: Three groups of patients (Total: 584) Age: Over 14 years	Internal consistency ✓	✓	✓	✓		✓	High level of internal consistency Acceptable ceiling and floor effects Moderate correlation with MAQLQ Correlation with changes in asthma severity and lung function
Life Activities Questionnaire for Adult Asthma Creer et al., 1992 USA	Developmental study including different groups of patients with asthma	Internal consistency ✓ Test re-test✓	✓					High levels of internal consistency and test re-test reliability. Content validity established by patients judgement
Life Quality (LQ) Test Winder et al., 2000 USA	Patients with asthma (239) out-patients, and people without asthma (46) from a dental practice		✓					Higher scores indicate worse asthma specific quality of life. Patients diagnosed with asthma had statistically significantly higher scores than those without asthma.
Perceived Control of Asthma Questionnaire (PCAQ) Katz et al., 2002 USA	Patients with asthma (374) Age: over 18 years Telephone survey	Internal consistency ✓	✓	✓	✓			High level of internal consistency Small to moderate correlation with clinical variables and perceived asthma severity scores; SF-36 and MAQLQ both cross sectional and longitudinal analysis One other record identified: Chinese

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Quality of Life Diary Hyland et al., 1995 UK	Patients participating in a RCT (426) Age: 16 years and over		✓	✓	✓	✓		Correlation with respiratory function and LWAQ 75% compliance with diary for 20 days Moderate correlation with LWAQ and respiratory function cross sectional and longitudinal
University of Alabama at Birmingham (UAB) Functional Impairment Scale Player et al., 1994 USA	Total of 382 patients with asthma Self-report Out-patients	Internal consistency ✓	✓		✓			High level of internal consistency No floor or ceiling effects Moderate correlation with other asthma measures: Bother scale, symptom survey and asthma opinion survey

SUMMARY - GENERIC INSTRUMENTS

A total of twenty-one articles were included in the review which reported results from evaluation studies of generic instruments evaluated with patients with asthma.

Six generic instruments were identified in the review which had been evaluated with people with asthma. Only two though, the SF-36 and SF-12 were the principal instrument undergoing evaluation. The others included the EuroQol which was evaluated concurrently with other asthma-specific instruments (Garratt 2000); HUI, RAND and SIP as a reference instrument (Leidy and Coughlin1998a, Leidy et al., b; Juniper et al., 1993; Rowe and Oxman 1993; Marks et al., 1993).

The included instruments were evaluated with a wide range of patients with different asthma severities and classifications defined as patient-reported symptom prevalence and severity, use of medications and physiological lung function. The overall age of participants was forty years and study sizes ranged from 40 to 1406.

Only two studies were conducted in the UK (Garratt et al., 2000; McColl et al., (1995) using the SF-12 and SF-36. Other studies were from Canada and USA.

Details of the instruments domains, items and scoring procedures are detailed in Chapter 3, Tables 1 and 2. All instruments are multi-dimensional with an average of six domains similar in construct. The SIP though does not include a Global judgement question but has a Cognitive functioning domain. Item content ranges from five (EQ-5D) to 136 for the SIP. All instruments have a scoring algorithm and the SF-36, SF-12 and SIP have domain scoring and component scores (Physical and Mental). The SIP items are weighted. All are self-administered or interview. One study evaluated and compared results of two versions of the SF-36 (Standard vs. Acute form) (Keller et al., 1997). Completion times range from five (EQ-5D) to twenty minutes (SIP).
.
The most frequently reported instrument evaluated was the SF-36 with evidence provided for most measurement selection criteria. The overall evidence supports the use of the SF-36 as a generic instrument for the assessment of health-related quality of life in people with asthma. The studies included in the review report acceptable internal consistency and reproducibility for group comparison. The validity of the SF-36 has been comprehensively examined in concurrent evaluation with asthma-specific instruments and also the HUI and provides evidence of a relationship of measuring similar constructs. Empirical evidence supports the internal structure and proposed health domains of the SF-36. The SF-36 has evidence of responsiveness but does not perform as well as asthma-specific instruments. The evidence available suggests that it is acceptable to patients in the studies included. The SF-12 has also been evaluated with patients with asthma in two UK studies and evidence although limited, supports the hypothetical construct underpinning the instrument and that it is responsive to change.

Limited evidence is reported for other instruments identified in the review (EQ-5D, HUI, RAND and SIP). The EQ-5D performed equally as well as the SF-12 in a concurrent evaluation but was less responsive than asthma-specific instruments. The HUI, RAND and SIP have not been the principal instrument under study nor evaluated concurrently. Evidence of performance therefore can only be extrapolated

from studies which have used these instruments as a reference measure of construct validity. Results from these studies suggest that there is moderate correlation between specific and generic instruments.

Overall, the SF-36 is the most rigorously evaluated generic instrument and provides evidence to support its application with patients with asthma. The SF-12 has some evidence to support application but further evaluations are needed to be confident in recommending it. There is limited evidence available to support or refute the use of other generic instruments included in this review.

Limited evidence is available for the comparative performance of generic instruments. The lack of this evidence is limiting as this would give a clear indication of which instrument performs the best with patients with asthma. Concurrent evaluations and principal instrument evaluations are necessary for other available generic instruments to provide evidence of the measurement and practical properties before recommendations can be made.

Table 4.11: Summary of generic instruments: measurement properties

Instrument	**Measurement properties**	**Availability**: Royalty; Scoring methods and interpretation guide	**Acceptability/Feasibility**: Patient acceptability Staff acceptability
SF-36	Two UK evaluations Several studies evaluating most measurement criteria. Good evidence of reliability, validity and responsiveness supporting application	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self- report 10 to 15 minutes to complete Some difficulties experiences with completion
SF-12	One UK evaluation	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self-report 5 minutes completion Acceptable to patients
SIP	Limited evaluations and evidence of measurement properties. It has only been used as a reference measure.		
HUI	Limited evaluations and evidence of measurement properties. It has only been used as a reference measure.		
EuroQol	Three /four evaluations in the UK Some evidence of measurement properties	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self-report 5 minutes completion VAS higher proportion of missing responses

SUMMARY - ASTHMA-SPECIFIC INSTRUMENTS

Thirty-five studies provided some evidence of measurement and/or practical properties for the asthma-specific measures included in the review.

Nine instruments were reviewed: ACQ, ACD, AQLQ, AQLQ(S), MiniAQLQ, Acute AQLQ MAQLQ, and LWAQ. The SGRQ, a general respiratory specific instrument, was also included. A further eight instruments were identified which had undergone one evaluation and are illustrated in Table 4.10.

The number of participants included in studies ranged from 40 to 1406 and average age was 40 years. Four of the 35 studies were conducted in the UK. Most studies were conducted in an out-patient primary care setting. Two studies were conducted in emergency departments. Several studies administered questionnaires via interviews; six studies adopted postal surveys, and two telephone surveys. One study used an online version of the AQLQ (Baghi et al., 2004). Four studies evaluated instrument measurement properties following clinical trials of medication effectiveness (Cook et al., 1993, Garratt et al., 2000, Ware et al., 1998, Juniper et al., 2005).

All instruments included the assessment of symptoms, with the exception of the ACQ and ACD. Psychological well-being was the next most frequently assessed domain. Several instruments assessed role activities (AQLQ, MiniAQLQ, AQLQ(S), and SGRQ). Social well-being was assessed in two instruments (MAQLQ; SGRQ). One instrument assessed personal construct (MAQLQ) and one assessed treatment satisfaction (SGRQ). The number of assessed domains ranged from three to six; total number of items ranged was from seven (ACD/ACQ) to 68 (LWAQ). The SGRQ has a total and domain weighted scoring system. All are available as self completion although interview methods are recommended by the developers for the SGRQ. The ACQ has self report responses and one clinician assessed item.

The most comprehensively evaluated instruments were the Juniper Asthma Quality of Life Questionnaires (AQLQ, AQLQ(S), MiniAQLQ and AcAQLQ). An extensive and thorough synthesis of evidence in support of a wide range of measurement and practical properties provides favourable support for the collection of instruments developed by Juniper et al. Modest and promising evidence of both measurement and practical properties are presented for the LWAQ, SGRQ and MAQLQ.

All instruments have been developed in collaboration with patients with asthma and have undergone evaluation of face and content validity.

All instruments included in the review have evidence of reliability supporting application in studies involving groups of patients; two instruments have higher levels of reliability supporting application in individual analysis (AQLQ, MAQLQ).

Empirical evidence supports the proposed domain structure for the MiniAQLQ, MAQLQ and LWAQ.

Most instruments were assessed for validity through comparison with other instruments. All instruments have evidence for validity through comparison with

instruments that measure similar or related constructs. This is most extensive for the Juniper instruments.

Evidence of responsiveness was reported for all instruments except the LWAQ. In concurrent evaluations, the instruments included in the review performed better than generic comparator instruments.

Instrument patient acceptability is reported for the AQLQ and MAQLQ; some patients experience difficulty completing the individualised questions included in the AQLQ. The AQLQ(S) does not include the individualised section, but has good evidence of measurement and practical properties.

There is good evidence of the measurement and practical properties for the AQLQ, AQLQ(S), and MiniAQLQ and the MAQLQ. Limited evidence for the LWAQ and SGRQ was reviewed; further evaluations are required.

Some instruments have been identified in this review which report only one evaluation. There is insufficient evidence therefore to make firm recommendations about these at present.

Concurrent evaluations of asthma-specific and generic instruments provide further good evidence of performance. As expected, results indicate that asthma-specific instrument generally perform better that generic instrument particularly with reference to responsiveness. This may reflect the specific domain structure of the instruments and greater relevance to health concerns of patients with asthma.

There is limited evidence available for the comparative performance of asthma-specific instruments with the exception of comparative performance of different versions of instruments. The lack of this evidence is limiting as this would give a clear indication of which instrument performs the best with patients with asthma.

Table 4.12: Summary of asthma-specific instruments: measurement properties

Instrument	**Measurement properties**	**Availability**: Royalty; Scoring methods and interpretation guide	**Acceptability/Feasibility**: Patient acceptability Staff acceptability
Asthma-specific			
ACQ) ACD AQLQ MiniAQLQ AQLQ(S) Acute AQLQ	Several studies evaluating most measurement criteria. Good evidence of reliability, validity and responsiveness supporting application (AQLQ, AQLQ(S), MiniAQLQ) 4/33 evaluations in the UK. Others in Canada and USA	All questionnaires and the translations are copyrighted. They must not be altered in any way, sold, translated or adapted for another medium (*e.g.,* computer) without the written permission of Professor Elizabeth Juniper. Scoring methods illustrated	Acceptable to patients particularly the AQLQ(S) and the MiniAQLQ. Patients experienced some difficulty with the individualised Activity domain of the AQLQ. Questionnaires suitable for self-completion and interview administration Maximum of 10 minutes completion No details of cost
MAQLQ	Developed and evaluated in Australia Six evaluations with comprehensive testing of measurement properties Good evidence of reliability, validity and responsiveness supporting application No UK evaluations	No details of licensing or permission for use. Contact details provided.	Acceptable to patients Questionnaires suitable for self-completion and interview administration No details of cost, completion time

Instrument	**Measurement properties**	**Availability**: Royalty; Scoring methods and interpretation guide	**Acceptability/Feasibility**: Patient acceptability Staff acceptability
LWAQ	Developed in the UK Some evidence of reliability and validity	Permission required and contact details provided	Suitable for self-completion but 68 items 10 to 20 minutes completion No details of patient acceptability or feasibility
SGRQ	Developed in the UK Three evaluations with people with Asthma. Has been used with patients with COPD Some evidence of reliability and validity	Contribution to the St. George's Research Fund is requested from commercial organizations using the instrument. Permission should be obtained from the authors Scoring algorithms and calculators are available from the developers	Self- report but recommended interview administered 8- 15 minutes to complete No details of patient acceptability

DISCUSSION AND RECOMMENDATIONS

Many evaluations have been identified in this review of both generic and asthma-specific instruments with patients with different disease severities. The evaluations were conducted mainly in an out-patient setting and although all instruments were completed by the patients, some were administered during interviews. There are limited UK evaluations. Most have been applied in the USA or Canada.

The SF-36 is the most widely evaluated generic instrument and the Juniper collections of instruments have extensive evidence of measurement properties. There is also promising evidence for several additional asthma-specific instruments - MAQLQ, LWAQ, SGRQ.

The generic instruments chiefly the SF-36, included in the review adopt a multi-dimensional perspective to the measurement of patient-reported health.

All reviewed asthma-specific instruments address multi-dimensional aspects of health-related quality of life. All include the assessment of symptoms and most also include psychosocial well-being. Other frequently assessed dimensions include the impact of asthma on role activities and personal constructs.

The lack of studies evaluating the performance of different generic instruments is disappointing. The SF-12 and EQ-5D performed equally well in one concurrent evaluation where moderate levels of correlation were reported between the SF-36 and HUI.

Concurrent evaluations of asthma-specific instruments were dominated by evaluations of modifications of the AQLQ (Juniper): modifications of the AQLQ performed as well as the original version in most populations and settings.

However, several studies report the concurrent evaluation of generic and asthma-specific instruments. Good evidence supports the reliability and validity of both generic (SF-36) and asthma-specific (AQLQ collection) measures, supporting their combined use in people with asthma. However, and as expected, consistently higher levels of responsiveness were reported for the asthma-specific instruments.

Recommendations

Synthesising the available primary evidence reported in this review and extrapolating evidence from concurrent evaluations supports the use of both generic and asthma-specific patient-reported health instruments for people with asthma. The SF-36 is recommended as a generic instrument for the broad evaluation of health-related quality of life for people with asthma. Further evaluations are required, particularly concurrent evaluations of different generic instruments to inform further recommendations and for the UK population.

Asthma-specific instruments particularly the AQLQ Juniper collection and the MAQLQ are recommended and different versions of the AQLQ instruments selected for particular purposes. For example, although the AQLQ original version of the

instrument has been widely evaluated and demonstrates good performance, some patients experience some difficulty when completing the individualised activity questions. The AQLQ(S), which does not include the individualised questions, may therefore be preferable. Furthermore, the MiniAQLQ may be more acceptable to patients and administrators. It may not be as responsive to changes in health as the other, more comprehensive versions (AQLQ and AQLQ(S)). Further evaluations are needed to support the use of these instruments specifically in the UK and further evaluations are required for the SGRQ and LWAQ. The SGRQ has though been evaluated extensively with patients with COPD (Chapter 5).

REFERENCES

Adams RJ, Ruffin RE, Smith BJ. Validity of a modified version of the Marks Asthma Quality of Life Questionnaire. *Journal of Asthma* 2000; **37**:131-43.

Baghi, H., & Atherton, M. Construct Validity and Reliability of Scores on Scales to Measure the Impairment of Health-Related Quality of Life in Persons with Asthma. *Journal of Nursing Measurement* 2004; **12**(1), 21-31.

Barley EA, Quirk FH, Jones PW. Asthma health status measurement in clinical practice: validity of a new short and simple instrument. *Respiratory Medicine* 1998; **92:**1207-14.

Barley EA, Jones PW. Repeatability of a Rasch model of the AQ20 over five assessments. *Quality of Life Research* 2006; **15**: 801-809

Bayliss MS, Espindle DM, Buchner DA, Blaiss MS, Ware JEJ. A new tool for monitoring asthma outcomes: the ITG Asthma Short Form. *Quality of Life Research* 2000; **9**:451-66.

Blumenschein K, .Johannesson M. Relationship between quality of life instruments, health state utilities, and Willingness To Pay in patients with asthma. *Annals of Allergy, Asthma and Immunology* 1998; **80**:189-94.

Caro JJ, Caro I, Caro J, Wouters F, Juniper EF. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research* 2001; **10**:683-91.

Cook DJ, Guyatt GH, Juniper EF, Griffith LE, McIlroy W, Willan AR *et al*. Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. *Journal of Clinical Epidemiology* 1993; **46**:529-34.

Creer TL, Wigal JK, Kotses H, McConnaughy K, Winder JA. A life activities questionnaire for adult asthma. *Journal of Asthma* 1992; **29**:393-9.

Franic, Duska M; Aull, Larry; Grauer, Dennis; Oyelowo, Olatoye. Adherence, asthma control, generic and disease-specific quality-of-life instruments in asthma. *Expert Review of Pharmacoeconomics and Outcomes Research* 2005; 5, 4: 411-421(11)

Garratt AM, Hutchinson A, Russell IT. Patient-assessed measures of health outcome in asthma: a comparison of four approaches. *Respiratory Medicine* 2000; **94**:597-606.

Gupchup GV, Wolfgang AP, Thomas JI. Reliability and validity of the Asthma Quality of Life Questionnaire-Marks in a sample of adult asthmatic patients in the United States. *Clinical Therapeutics* 1997; **19**:1116-25.

Hazell M, Frank T, Frank P. Health related quality of life in individuals with asthma related symptoms. Respir Med. 2003;97(**11**):1211-8.

Hyland ME, Finnis S, Irvine SH. A scale for assessing quality of life in adult asthma sufferers. *Journal of Psychosomatic Research* 1991; **35**:99-110.

Hyland ME. The living with asthma questionnaire. *Respiratory Medicine* 1991; **85**:13-6.

Hyland ME. Measuring quality of life of adult asthmatics: a patient-centred approach. In Christie MJ, French DJ, eds. *Assessment of quality of life in childhood asthma. 189 pp*, pp 147-55. Langhorne, PA, USA: Harwood Academic Publishers/Gordon and Breach Science Publishers, 1994.

Hyland ME, .Crocker GR. Validation of an asthma quality of life diary in a clinical trial. *Thorax* 1995; **50**:724-30.

Hyland ME, Bellesis M, Thompson PJ, Kenyon CAP. The constructs of asthma quality of life: psychometric, experimental, and correlational evidence. *Psychology and Health* 1996; **12**:101-21.

Jones PW, Quirk FH, Baveystock CM. The St George's Respiratory Questionnaire. *Respiratory Medicine* 1991; **85**:25-31.

Jones PW. Testing health status (quality of life) questionnaires for asthma and COPD. *European Respiratory Journal* 1998; **11**:5-6.

Juniper EF, Guyatt GH, Epstein RS, Ferrie PJ, Jaeschke R, Hillers TK. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. *Thorax* 1992; **47**:76-83.

Juniper EF, Guyatt GH, Ferrie PJ, Griffith LE. Measuring quality of life in asthma. *American Review of Respiratory Disease* 1993; **147**:832-8.

Juniper EF, Guyatt GH, Willan AR, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *Journal of Clinical Epidemiology* 1994; **47**:81-7.

Juniper EF. Assessing health-related quality of life in asthma. *Canadian Respiratory Journal* 1997; **4**:145-51.

Juniper EF, Buist AS, Cox FM, Ferrie PJ, King DR. Validation of a standardized version of the Asthma Quality of Life Questionnaire. *Chest: the Cardiopulmonary Journal* 1999; **115**:1265-70.

Juniper EF, Guyatt GH, Cox FM, Ferrie PJ, King DR. Development and validation of the Mini-Asthma Quality of Life Questionnaire. *European Respiratory Journal* 1999; **14**:32-8.

Juniper EF, O'Byrne PM, Guyatt GH, Ferrie PJ, King DR. Development and validation of a questionnaire to measure asthma control. *European Respiratory Journal* 1999; **14**:902-7.

Juniper EF, O'Byrne PM, Ferrie PJ, King DR, Roberts JN. Measuring asthma control: clinic questionnaire or daily diary? *American Journal of Respiratory and Critical Care Medicine* 2000; **162**:1330-4.

Juniper EF, Norman GR, Cox FM, Roberts JN. Comparison of the Standard Gamble, Rating Scale, AQLQ, and SF-36 for measuring quality of life in asthma. *European Respiratory Journal* 2001; **18**:38-44.

Juniper EF, Svensson K, Mork AC, Stahl E. Measuring health-related quality of life in adults during an acute asthma exacerbation. *Chest* 2004; **125**:93-7.

Juniper EF, Svensson K, Mork AC, Stahl E. Measurement properties and interpretation of three shortened versions of the asthma control questionnaire. *Respir.Med.* 2005;**99**:553-8.

Katz PP, Eisner MD, Henke J, Shiboski S, Yelin EH, Blanc PD. The Marks Asthma Quality of Life Questionnaire: further validation and examination of responsiveness to change. *Journal of Clinical Epidemiology* 1999; **52**:667-75.

Katz PP, Yelin EH, Eisner MD, Blanc PD. Perceived control of asthma and quality of life among adults with asthma. *Annals of Allergy, Asthma and Immunology* 2002; **89**:251-8.

Keller SD, Bayliss MS, Ware JEJ, Hsu MA, Damiano AM, Goss TF. Comparison of responses to SF-36 health survey questions with one-week and four-week recall periods. *Health Services Research* 1997; **32**:367-84.

Lee TA, Hollingworth W, Sullivan SD. Comparison of directly elicited preferences to preferences derived from the SF-36 in adults with asthma. *Medical Decision-Making* 2003; **23**:323-34.

Leidy NK, .Coughlin C. Psychometric performance of the Asthma Quality of Life Questionnaire in a US sample. *Quality of Life Research* 1998a; **7**:127-34.

Leidy NK, Chan KS, Coughlin C. Is the Asthma Quality of Life Questionnaire a useful measure for low-income asthmatics? *American Journal of Respiratory and Critical Care Medicine* 1998b; **158**:1082-90.

Magid DJ, Houry D, Ellis J, Lyons E, Rumsfield JS. Health-related quality of life predicts emergency department utilization for patients with asthma. *Annals of Emergency Medicine* 2004; **43**:551-7.

Maille AR, Koning CJM, Zwinderman AH, Willems LNA, Dijkman JH, Kaptein AA. The development of the Quality of Life for Respiratory Illness Questionnaire/QOL-RIQ: a disease-specific quality of life questionnaire for patients with mild to moderate chronic non-specific lung disease. *Respiratory Medicine* 1997; **91**:297-309.

Mancuso CA, Peterson MGE, Charlson ME. Comparing discriminative validity between a disease-specific and a general health scale in patients with moderate asthma. *Journal of Clinical Epidemiology* 2001; **54**:263-74.

Mancuso CA, .Peterson MGE. Different methods to assess quality of life from multiple follow-ups in a longitudinal asthma study. *Journal of Clinical Epidemiology* 2004; **57**:45-54.

Marks GB, Dunn SM, Woolcock AJ. A scale for the measurement of quality of life in adults with asthma. *Journal of Clinical Epidemiology* 1992;**45**:461-72.

Marks GB, Dunn SM, Woolcock AJ. An evaluation of an asthma quality of life questionnaire as a measure of change in adults with asthma. *Journal of Clinical Epidemiology* 1993;**46**:1103-11

McColl E, Steen N, Meadows KA, Hutchinson A, Eccles MP, Hewison J *et al*. Developing outcome measures for ambulatory care: an application to asthma and diabetes. *Social Science and Medicine* 1995; **41**:1339-48.

McColl E, Eccles MP, Rousseau NS, Steen IN, Parkin DW, Grimshaw JM. From the generic to the condition-specific?: Instrument order effects in Quality of Life Assessment. *Med Care.* 2003;**417**:777-90.

Orr LC, Fowler SJ, Lipworth BJ. Relationship between changes in quality of life and measures of lung function and bronchial hyper-responsiveness during high-dose inhaled corticosteroid treatment in uncontrolled asthma. *American Journal of Respiratory Medicine* 2003; **2**:433-8.

Pinnock H, Sheikh A, Juniper EF. Evaluation of an intervention to improve successful completion of the Mini-AQLQ: Comparison of postal and supervised completion. *Primary-Care-Respiratory-Journal* 2004; **13**:36-41.

Player R, Richards JMJ, Kohler CL, Woodby LL, Brooks CM, Bailey WC. Scale for assessing functional impairment in adults with asthma. *Journal of Asthma* 1994; **31**:437-44.

Quirk FH, Jones PW. Patients' perception of distress due to symptoms and effects of asthma on daily living and an investigation of possible influential factors. *Clin Sci* 1990; **79**: 17-21.

Quirk FH, Baveystock CM, Wilson RC, Jones PW. Influence of demographic and disease related factors on the degree of distress associated with symptoms and restrictions on daily living due to asthma in six countries. *Eur Respir J* 1991; **4**: 167-71.

Quirk FH, .Jones PW. Back to basics: how many items can adequately represent health-related quality of life in airways disease? *European Respiratory Review* 1997; **7**:50-2.

Ried LD, Nau DP, Grainger-Rousseau TJ: Evaluation of patient's Health-Related Quality of Life using a modified and shortened version of the Living With Asthma Questionnaire (ms-LWAQ) and the medical outcomes study, Short-Form 36 (SF-36). *Quality of Life Research* 1999;**8[6**]:491-9.

Rowe BH, .Oxman AD. Performance of an asthma quality of life questionnaire in an outpatient setting. *American Review of Respiratory Disease* 1993; **148**:675-81.

SIGN British Guideline on the management of Asthma 2005: Scottish Intercollegiate Guideline Network

Viramontes JL, .O'Brien BJ. Relationship between symptoms and health-related quality of life in chronic lung disease. *Journal of General Internal Medicine* 1994; **9**:46-8.

Vollmer M, Markson L E, O'Conner E, Sanocki L L, Fitterman L, Berger M, and Buist S. Association of Asthma Control with Health Care Utilization and Quality of Life 1999. *Am. J. Respir. Crit. Care Med*. 160, **5**,:1647-1652

Ware JEJ, Kemp JP, Buchner DA, Singer AE, Nolop KB, Goss TF. The responsiveness of disease-specific and generic health measures to changes in the severity of asthma among adults. *Quality of Life Research* 1998; **7**:235-44.

Winder JA, Nash K, Brunn JW. Validation of a life quality (LQ) test for asthma. *Annals of Allergy, Asthma and Immunology* 2000; **85**:467-72.

Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Quality of Life Research* 2002; **11**:1-7.

Wyrwich KW, Nelson HS, Tierney WM, Babu AN, Kroenke K, Wolinsky FD. Clinically important differences in health-related quality of life for patients with asthma: an expert consensus panel report. *Annals of Allergy, Asthma and Immunology* 2003; **91**:148-53.

Chapter 5: Patient-reported Health Instruments used for people with Chronic Obstructive Pulmonary Disease (COPD)

Chronic obstructive pulmonary disease (COPD) is a major cause of morbidity and mortality and is characterised by airflow obstruction. It is usually progressive and the result of chronic inflammation resulting in airway and parenchymal damage usually as a result of smoking (NICE (2004). It represents a substantial economic and social burden throughout the UK and a significant contributor to mortality. The exact prevalence of COPD is difficult to determine and define and therefore frequently under diagnosed and under treated, which further compromises morbidity. The burden of COPD to patients and their families and carers is high, both in terms of health-related quality of life and health status affecting physical and emotional functioning (Belza et al., 2005). COPD can lead to feelings of anxiety because of breathlessness. People may reduce their activities to avoid becoming breathless and subsequently become dependant on people for carrying out activities of daily living.

Reducing the burden of COPD requires better evaluation and diagnosis, as well as improved management of chronic symptoms and understanding the effect on health-related quality of life (Halpin and Miravitlles 2006). Understanding the impact of the disease on patients and carers can facilitate targeted interventions thus improving their quality of life.

The following review provides current information available of the patient-reported health questionnaires used to measure health-related quality of life with patients with COPD.

Search terms and results: identification of articles

At the time of the review, the PHI database contained 12,000+ records (up to June 2005). The primary search strategy, using the terms 'chronic obstructive pulmonary disease' and 'respiratory' keyword searching generated 468.records, as shown in Table 4.1. All abstracts were reviewed. When assessed against the review inclusion criteria, 220 articles were retrieved and reviewed in full. Of these, 46 articles were included in the review.

Table 5.1

Source	***Results of search***	***No. of articles considered eligible***	***Number of articles included in review***
PHI database: original search (up to June 2005) **Total number= 12,562**	468	220	41
Supplementary search	-	-	5
TOTAL	-	-	**46**

Supplementary searches which included hand searching of titles from 2004 to 2006 of the following key journals:
-Chest
-Health and Quality of Life Outcomes
-Medical Care
-Quality of Life Research
-Respiratory Medicine
-Thorax

Further searches were conducted within the bibliography and using Pub Med per instrument up to September 2006.

Identification of patient-reported health instruments
Seven generic and 5 COPD-specific instruments were included in the review. The developmental and evaluative studies relating to the instruments reviewed are listed in Tables 5.2 to 5.14. Table 5.15 illustrates instruments where only one publication was identified. Table 5.16 details instruments which were excluded from the review.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Seven generic instruments were identified which were evaluated with patients with COPD. For full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36
b) SF-20
c) SF-12
d) EQ-5D
e) Nottingham Health Profile
f) COOP Charts
g) Sickness Impact Profile

a) SF-36:

Eleven studies were identified which provide evidence of measurement properties for the SF-36. Two studies were conducted in the UK (Harper et al., 1997, Wilson et al., 1997). Three studies evaluated the SF-36 as the principal instrument (Benzo et al., 2000; Ruffin et al., 2000; Sprenkle et al., 2004); and the remaining included the SF-36 in concurrent evaluations.

Six studies reported that the SF-36 was self-completed; the others were either interview administered by telephone or face to face. All patients included in the studies had Chronic Obstructive Pulmonary Disease but two studies provided specific diagnoses for example Viramontes and O'Brien (1994) included patients with asthma, chronic bronchitis and emphysema. Patients were generally over 60 years old in the studies and representative of the disease population.

The number of patients included in the studies was variable and most sample sizes were less than 200. Exceptions were Ruffin et al., (2000), Sprenkle et al., (2004) and Wyrwich et al., (1999) with the number of patients ranging from 329 to 8345. Although most had predicted associations between similar constructs for validity and high levels of reliability, none provided a priori hypotheses for the strength of correlation.

Reliability

Reproducibility for the SF-36 following a six month period of testing was generally poor (lower than the 0.70 threshold) with the exception of the PF (0.86) and the MH (0.74) (Harper et al., 1997). However, this is a long period in which to assess reproducibility.

Internal consistency reliability was generally high for all domains with all exceeding 0.80 with exception of the GHP domain (Desikan et al., (2002; Harper et al., 1997; Wyrwich et al., 1999).

Item total correlations

Thirty-three of 35 item correlations were greater than 0.40 in Harper et al., (1997) but no details were provided of which items.

Validity

Health service use

Patients who had been in hospital during the last six months had worse scores for the Pain and Physical functioning domains of the SF-36 than population norms. No differences were found for the MCS (Harper 1997). The PCS was independently associated with healthcare utilisation with odds ratios of 1.54 ((95% CI 1.26 to 1.87) for high primary care and 1.46 (95%CI, 1.21 to 1.78) specialty medicine utilisation. No association was found for MCS and healthcare utilisation (Sprenkle et al., 2004). In addition, The SF-36 PCS and MCS were independent predictors of mortality with increasing hazard ratios with worsening quartiles compared to the reference population (first quartile) (Sprenkle et al., 2004).

Health status

Viramontes and O'Brien (1994) evaluated the discriminative validity of the SF-36 with patients with chronic lung diseases including asthma, emphysema and chronic bronchitis and reported significantly different domain scores between disease severity subgroups based on the UK Medical Research Council symptoms classification. Lower SF-36 scores were associated with higher dyspnoea scores as expected and moderate to large correlation was reported for activity threshold and SF-36 physical functioning, general health perception and energy. There was no relationship between disease severity and SF-36 ER, SF, Pn and MH. Furthermore, the SF-36 domains discriminated patients with breathing problems in the last four weeks with statistically significantly lower scores than those patients without breathing difficulties (Wyrwich et al., 1999).

Large effect sizes (≥0.80) (where patients were classified into severe and less severe breathlessness cases) for SF-36 Physical functioning; Moderate (<0.80) effect sizes, SF-36 SF, VT and GH and Small effect sizes ≥0.20 to <0.5) for SF-36 Pain, MH and RL emot.) (Harper et al., 1997)

SF-36 PCS was moderately correlated as hypothesised with a symptom severity score (Chronic Lung Disease Index) (Dyspnoea with PF -0.53) with small correlation for the MCS (-0.19) and Dyspnoea. All other correlations were less than -0.60 for all SF-36 domains and Cough and Wheeze (Ruffin et al., 2000).

Generic patient-reported health instruments
Strong correlation was observed for some related domains between the SF-36 and NHP in patients with chronic airflow limitation being assessed for home oxygen therapy. For males SF-36 PF and NHP Energy -0.67; MH with NHP Energy -0.66; Energy/VT with NHP Energy -0.80; PF with NHP Emotional reactions -0.54; MH with NHP Emotion -0.73; MH with NHP social isolation -0.61; PF with NHP Physical mobility -0.66. Scales from the two instruments that assessed different traits were not correlated as expected. For females there was strong correlation between the SF- 36 BP with NHP Pain -0.74; Energy/VT with NHP Energy -0.62. All other correlations were small to moderate for similar domains and no correlation for different domains as would be expected (Crockett et al., (1996).

COPD-specific patient-reported health instruments
Strong correlation (greater than 0.60) was found between the SF-36 and CRQ related domains (PF and Dyspnoea; Vitality and Fatigue; RE, MH and Emotional function). There was moderate correlation with all SF-36 domains and the CRQ Mastery domain (Wyrwich et al., (1999).

Respiratory function
The following domains were strongly correlated with the Baseline Dyspnoea Index: PF 0.91; RP 0.72; VT 0.60 and GHP 0.68. PF was also strongly correlated with FEV (Mahler and Mackowiak 1995).

Responsiveness
Several studies provide evidence of responsiveness of the SF-36. There were significant differences in change scores for the SF-36 PF and SF between sub-groups differing in their views of change on a transition question (Harper et al., 1997). The responsiveness of the SF-36 was reported in Benzo (2000) with small effect size for Pain and RP; and moderate effect sizes for other domains. This was a small study (22 participants of a rehabilitation programme) and the authors attribute the lack of improvement and responsiveness of the RP domain to the variance in the group.

The SF-36 was responsive to change applying a one-SEM criterion with similar percents of change across most domains (Wyrwich et al., 1999). The SF-36 PCS showed similar responsiveness to the CRQ and the Activities and Impact domains of the SGRQ but the MCS was not as responsive as the related emotional domains of these COPD-specific instruments (Puhan et al., 2006).

Clinically important difference
An expert physician panel established small, moderate and large clinically important change levels for the SF-36 as follows: *Small change* 8.3 (RE) to 12.5 (RP, VT, SF); *Moderate change*: 16.7 (RE) to 25 (RP, VT, SF); *Large change*: 25 (RE) to 37.5 (RP, VT, SF) (Wyrwich et al., 2003).

Precision
Floor effects have been reported in two studies (Harper et al., 1997; Wyrwich 1999) for Role Limitations: physical, Role Limitations: emotional. Ceiling effects were found for RL emotional (Harper et al., 1997) and RP and SF in Wyrwich et al., (1999).

Acceptability and Feasibility
The SF-36 is generally acceptable to patients but some evidence exists of missing data (20%) (Harper et al., 1997; Wyrwich et al., 1999).

SF-20
One study provided evidence of the measurement properties of the SF-20 (Mahler and Mackowiak 1994).

Reliability
No evidence reported.

Validity
Lung function
The SF-20 PF domain was strongly correlated with Baseline Dyspnoea Index (BDI) (0.70) which was greater than for other physiological measures. Overall, correlations with the SF-20 domains and other physiological measures (FEV, FVC) were less than 0.60 with no relationship for the Pain domain. The authors in this study hypothesised that the BDI would have greater impact on self-reported health status than other physiological lung function, a hypothesis which was supported in the results (Mahler and Mackowiak 1994).

No other measurement criteria evaluated

SF-12

One study provided evidence of the measurement properties of the SF-12 (Katz et al., 2005).

Health status
Scores on the SF-12 MCS were comparable to population norms and PCS scores were in the lower quartile of scores compared to norms (Katz et al., 2005). The authors hypothesised that poorer physical health (PCS) would be associated with difficulty with self-care and recreational activities. There was strong association with PCS scores and self care and recreational difficulty and consequent psychological distress in regression analyses (Katz et al., 2005).

No other measurement criteria evaluate

Table 5.2: Developmental and evaluation studies relating to the SF-36, SF-20 and SF-36 applied in patients with COPD

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Benzo et al., (2000) USA	COPD (22) participating in a rehabilitation programme Age: mean 64 Self-complete Out-patients			✓			
Crockett et al., (1996) Australia	Chronic Airflow limitation. Patients being assessed for home oxygen therapy (60) Age: mean Females 70; Males 67 Self-completed Out-patients		Construct ✓				
Desikan et al., (2002) USA	COPD (40) Age: range 41 to 71 Telephone interview	Internal consistency ✓	Construct ✓				
Harper et al., (1997) UK	COPD (156) Age: mean 67 Self-completed Out-patients	Internal consistency ✓	Construct ✓	✓	✓	✓	
Mahler and Mackowiak (1995) USA	COPD (50) Age: mean 72 Self-completed Out-patients		Construct ✓				
Puhan et al., (2006) Canada	COPD (177) participating in a rehabilitation programme Age: mean 69 Self completion Out-patients			✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Ruffin et al., (2000) Australia	Chronic Lung Disease (329) Age: mean 44 Interview administered		Construct ✓				
Sprenkle et al., (2004)| USA	Veterans with self-reported diagnosis of COPD (8345) Age: mean 60 Postal survey		Construct ✓				
Viramontes and O'Brien (1994) Canada	Patients with asthma, chronic bronchitis and emphysema (102) Age: mean 62 Self-completed but interview administered in patient's own homes		Construct ✓				
Wilson et al., (1997) UK	Patients with bronchiectasis (111) Age: mean 52 Self-completed Out-patients	Test re-test ✓	Construct ✓	✓			
Wyrwich et al., (1999) USA	COPD (487) Age: mean 58 Telephone interview		Construct ✓	✓	✓	✓	
SF-20							
Mahler and Mackowiak (1992) USA	Symptomatic COPD (110) Age: mean 67 Self completed Out-patients		Construct ✓				
SF-12							
Katz et al., (2005) USA	COPD (334) Age: mean 64 Telephone interview		Construct ✓				

b) EuroQol- EQ-5D

Three UK evaluations provide evidence of the measurement properties of the EQ-5D (Harper et al., 1996; Hazell et al., 2003; Paterson et al., 2000).

Reliability
Test re-test reliability for the EQ-5D, [6 months] 0.67 in Harper et al., (1996).

Validity
Age
The EQ-5D index and VAS scores decreased significantly with age with moderate correlations (-0.41; -0.34) as predicted by the authors (Hazell et al., 2003).

Health status
Hazell et al., (2003) reported the ability of the EQ-5D to discriminate patients with respiratory disease. A postal survey including the ED-5D and a respiratory questionnaire identifying patient with symptoms associated with obstructive airways disease. The survey was posted to all patents identified from a primary care practice in the UK (10,471) and those with self-reported respiratory symptoms were included in the analysis (6828, with 5944 questionnaire computable). The EQ-5D index and VAS scores were significantly lower for those with respiratory symptoms compared to those without. The EQ-5D also discriminated patients with COPD indicating poorer health than pop norms (Harper et al., 1996).

Responsiveness
Responsiveness was examined in a concurrent evaluation of the EQ-5D, MYMOP and MOS-6A (Patterson et al., 2000). Responsiveness for the EQ-5D was variable with the SRM comparable to the MYMOP (0.71) but the VAS not responsive (SRM 0.37).

Precision
EQ-5D no floor or ceiling effects were reported in Harper et al., (1996).

Acceptability
87% of responses for the EQ-5D were computable in a postal survey (6828) with the highest proportion of missing values for the self-care domain (5.7%); anxiety/depression (4.4%); usual activities (4.3%); pain (4.1%); mobility (3.9%). The VAS though had a greater proportion of missing responses (6.3%) (Hazell et al., 2003; Patterson et al., 2000). Completion rates were 92-96% in Harper et al., (1996).

Table 5.3: Developmental and evaluation studies relating to the EQ-5D applied in patients with COPD

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
EQ-5D		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Harper et al., (1997) UK	Patients with COPD (156) Age: mean 67 Self-completed Out-patients	Test re-test ✓	Construct ✓	✓	✓	✓	
Hazell et al., (2003) UK	Asthma related symptoms including COPD (5944) Age: mean 48 Postal survey Primary care practice		Construct ✓			✓	
Paterson et al., (2000) UK	Acute exacerbation of chronic bronchitis (81) Age: mean 61 Self-completed Out-patients		Construct ✓	✓		✓	

c) NHP

Reliability
No evidence found.

Validity
Generic patient-reported health instruments
Strong correlations were observed for some related domains between the SF-36 and NHP in patients with chronic airflow limitation being assessed for home oxygen therapy. For males SF-36 PF and NHP Energy -0.67; MH with NHP Energy -0.66; Energy/VT with NHP Energy -0.80; PF with NHP Emotional reactions -0.54; MH with NHP Emotion -0.73; MH with NHP social isolation -0.61; PF with NHP Physical mobility -0.66. There was no correlation with scales with similar traits as would be expected. For females there was strong correlation between the SF- 36 BP with NHP Pain -0.74; Energy/VT with NHP Energy -0.62. All other correlations were small to moderate for similar scales and no correlation for other domains (Crockett et al., (1996).

No other measurement criteria reported

Table 5.4: Developmental and evaluation studies relating to the NHP applied in patients with COPD

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
NHP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Crockett et al., (1996) Australia	Chronic Airflow limitation. Patients being assessed for home oxygen therapy (60) Age: mean Females 70; Males 67 Self-completed Out-patients		Construct ✓				

d) COOP Charts

One study provided evidence of measurement properties (Eaton et al., 2005).

Reliability

ICC assessed after 2 months were as follows in a study by Eaton et al., (2005): PF 0.45; DA 0.48; Pain 0.61; SA 0.43; SS 0.38; Feelings 0.53; OH 0.51; Change in Health 0.17 (ns/s) QOL 0.36.

Validity

Respiratory specific patient-reported health instruments

Stronger associations were reported for the COOP Feeling domain with CRQ emotional function 0.70 and HAD anxiety 0.70. The COOP Physical with CRQ dyspnoea was moderately correlated (0.40), consistent with expectations of two related but different constructs (Eaton et al., 2004).

Generic patient-reported health instruments

There was moderate correlation with the COOP PF with SF-36 PF 0.4 which was not as high as was expected. Strong correlations were found for similar domains in the two instruments (BP and Pain; SA with SF) (Eaton et al., 2004).

Responsiveness

There was moderate longitudinal correlation between the COOP PF with SF-36 PF 0.5; COOP DA and SF-36 RP 0.5; BP and Pain 0.80; SA with SF 0.60; Feelings with RE 0.50; Overall health with GHP 0.50 (Eaton et al., 2004). Effect size statistics (Standardised means: SE) were moderate for SA and Change on health (-0.51; -0.59) but small for other domains (Eaton et al., 2004).

Precision

No evidence found

Acceptability
The COOP was easy to administer; no patient required assistance to complete and there were no missing values in Eaton et al., (2004).

Feasibility
No evidence found

Table 5.5: Developmental and evaluation studies relating to the COOP Charts applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
COOP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Eaton et al., (2004) New Zealand	COPD (50) patients participating in a RCT of ambulatory oxygen therapy Age: mean 68 Self completion Out-patients	Test re-test ✓	Construct ✓	✓			

e) Sickness Impact Profile

Two studies provide evidence of validity of the SIP (McSweeny et al., 1982; Okubadejo et al., 1996) one of which was conducted in the UK.

Reliability
No evidence found.

Validity
Respiratory function
There was no correlation with the SIP Physical with FEV 0.10; Psychosocial 0.02; Total 0.14. Similar results were reported for PaO2 and PaCo2 (Okubadejo et al., (1996). Small to moderate correlation with SIP Total and maximum workload during exercise and Oxygen transport was reported in McSweeny et al., (1982).

No other measurement criteria reported

Table 5.6: Developmental and evaluation studies relating to the SIP applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Okubadejo et al., (1996) UK	COPD (41) Age: median 70 Self completion with supervision Out-patients		Construct ✓				
McSweeny et al., (1982) USA	COPD (203) Age: mean 65 Self completion Out-patients		Construct ✓				

COPD-SPECIFIC PATIENT- REPORTED HEALTH INSTRUMENTS:

Five COPD-specific instruments were identified which were evaluated with patients with COPD. Full details of the development, domains and scoring methods are detailed in Tables 5.7 and 5.8.

The following instruments measurement properties are reported:

a) Breathing Problems Questionnaire
b) Chronic Respiratory Disease Questionnaire
c) Functional Performance Inventory
d) St. George's Respiratory Questionnaire
e) Seattle Obstructive Lung Disease Questionnaire

a) Breathing Problems Questionnaire (BPQ)

The Breathing Problems Questionnaire (BPQ) items were derived from focus groups with fifteen patients and refined by 89 COPD patients (Hyland 1994). The instruments foundation was based on three constructs of quality of life: Problems, Negative evaluations and Positive evaluations. Following factor analysis two factors emerged with 27 items constituting the BPQ problems score and 6 items the BPQ negative evaluations score. Further modifications have been made with a shortened version of ten items and a single scale score (Hyland et al.,1998).

b) Chronic Respiratory Disease Questionnaire (CRQ)

The Chronic Respiratory Disease Questionnaire (CRQ) was developed following interviews with 100 patients with chronic airflow limitation to identify the impact on their quality of life and how important their symptoms were (Guyatt 1987). The most frequently reported and important items were selected and provided the conceptual framework for the instruments which were categorised into four domains: Dyspnoea, Fatigue, Emotional function and Mastery. The Dyspnoea domain is individualised and related to activities which patients report their degree of dyspnoea. A list is provided to aid selection where needed.

The CRQ has four domains with a total of 20 items and include Dyspnoea (5 items); Fatigue (4 items); Emotional functioning (7 items); Mastery (4 items). Scoring is by domains and uses a seven-point Likert scale with higher scores reflecting no impairment.

c) Functional Performance Inventory (FPI)

The FPI was developed in the USA involving both patients and clinical experts. Focus groups with patients informed an activity profile and content evaluated by clinical experts. Pre-testing of the instruments face validity was evaluated with a group of patients. The FPI is based on a conceptual framework of functional status as a multidimensional concept involving activities carried out to meet basic need, fulfill roles and maintain health and well-being (Leidy., 1999). The FPI has six domains (65items): Body Care (9); Household Maintenance (21); Physical Exercise (7); Recreation (11); Spiritual Activities (5) and Social Activities (12). Response options range from 1 where the activity can be performed easily to 4 where the activity is no longer performed for health reasons. Higher scores reflect high functioning. Domain

and Total scores are computable. Modifications have been made to scoring by Larson et al., (1998).

d) St. George's Respiratory Questionnaire (SGRQ)

The SGRQ was developed in the UK to measure the impact of asthma and chronic obstructive pulmonary disease (COPD) from a patient perspective. There are two parts of the instrument. Part 1 is concerned with symptoms focusing on the severity, frequency and effect of respiratory symptoms over the last year and responses are obtained with a 5 point Likert scale. Part 2 includes two domains: Activity limitations and social and psychological Impact and focuses on the patient's current state with True or False responses. Three components scores are calculated and a total score. All items have empirically derived weights and normative data are available. Scoring algorithms and calculators are available from the developers. Scores are expressed as the percentage of overall impairment with 100 equaling to worst possible health and zero the best.

Items were initially derived from studies with adult patients with asthma examining distress ratings relating to symptoms and the impacts of asthma (Quirk and Jones1990) and the influence of demographic and disease factors with the degree of distress (Quirk et al., 1991). Empirical weights were obtained from one hundred and forty patients with asthma (Quirk et al., 1991). Further analysis of previously derived weights was compared with patients with COPD with thirty-six patients (mean age 66) (Jones 1991) and no significant differences between the item weights from the asthma patients (Quirk et al., 1991) and COPD patients.

The final instrument has 50 items and 76 weighted responses divided into three components Symptoms, Activities and Impacts

e) Seattle Obstructive Lung Disease Questionnaire

The Seattle Obstructive Lung Disease Questionnaire (SOLQ) was developed using the CRQ model of functional status but with the intention of providing a self-reported questionnaire which can be computer scannable and therefore processing and scoring enabling feasibility. Dimensions were selected from patient interviews, literature, and clinical experience of the developers and the CRQ model of COPD specific health-related quality of life.

The instrument comprise of four domains: Physical functioning, Emotional functioning, Coping skills and Treatment satisfaction with 29 items. Scoring is on a simple linear scale with lowest scores indicating poorer functioning. Domain scores are computed. Permission for use is required from the author:

COPD-SPECIFIC INSTRUMENTS:

Table 5.7: COPD-specific patient-reported health instruments

Instrument	***Domains (no. items)***	***Response options***	***Score***	***Administration/ Completion (time)***
Breathing Problems Questionnaire (BPQ)	*Thirteen domains (33),* Two subscales: Problems and Emotional evaluations Walking (3) Bending or reaching (2) Washing and bathing (2) Household chores (3) Social interactions (3) Effects of weather and temperature (4) Effects of smells and fumes (2) Effects of colds (1) Sleeping (2) Medicine (2) Dysphoric states (5) Eating (2) Excretion urgency (2)	4 point Likert	Subscale scores Lower scores better quality of life	
Chronic Respiratory Disease Questionnaire (CRQ)	*Four domains (20 items)* Dyspnoea (5) Fatigue (4) Emotional functioning (7) Mastery (4)	7 point Likert	Domain Higher score indicate no impairment	Maximum 30 minutes
Functional Performance Inventory (FPI)	*Six domains (65 items*): Body Care (9); Household Maintenance (21); Physical Exercise (7); Recreation (11); Spiritual Activities (5) and Social Activities (12).	5 point Likert	Total and Domain Higher scores reflect high functioning	
Seattle Obstructive Lung Disease Questionnaire (SOLQ)	*Four domains (29 items)* Physical functioning Emotional functioning Coping skills Treatment satisfaction	Linear scale	Domain score Lower score indicate lowest function	5-10 minutes completion Computer scannable
St. George's Respiratory Questionnaire (SGRQ)	*Two parts; Domains (3)/17 items* Part 1: Symptom scores (8) Part 2: Activity and Impact (9)	Part 1: 5 point Likert Part 2: True or False	Weighted scoring Total and domain scores Percentage of overall impairment 0=best possible health and 100 worse	Self- report but recommended interview administered 8- 15 minutes to complete

Table 5.8: Summary of COPD-specific instruments: health status domains

	Instrument domains (after Fitzpatrick et al., 1998)								
Instrument	**Physical function**	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**	**Treatment satisfaction**
BPQ	x	x		x	x		x		x
CRQ	x	x		x				x	
FPI	x				x		x	x	
SGRQ	x	x		x	x		x		x
SOLQ	x			x				x	x

COPD- SPECIFIC PATIENT- REPORTED HEALTH INSTRUMENTS

a) Breathing Problems Questionnaire (BPQ)

Two UK studies provide evidence of measurement properties for the BPQ (Hyland et al., 1998; Yohannes et al., 1998). Both studies administered the questionnaire by interview in an out-patients setting.

Reliability
Test re-test reliability was adequate in Hyland et al., (1998): 0.73 Problems and 0.64 for Emotional evaluation.

Internal consistency alphas for all domains were greater than 0.75 pre and post rehabilitation (Hyland et al., 1998).

Validity
Internal validity
Empirical evidence was established in the developmental study with two factors emerging; the BPQ Problems score (27 items) and BPQ Negative evaluations score (6 items) (Hyland 1994). Further analysis of the structure of the instrument was conducted to examine the relationship between personality and measures of problems both negative and positive and thus total life satisfaction. It was predicted that negative evaluations and positive evaluations contribute independent variance to total life satisfaction. The revised Eysenck Personality Questionnaire (EPQ-R), Satisfaction with Life Questionnaire and the Satisfaction with Illness Scale were administered to test these hypotheses. The results were as predicted that negative evaluations were correlated with neuroticism and that positive evaluations correlated with extraversion (Hyland 1994).

Health status
The BPQ scores for patients with chronic airflow limitation were statistically significantly higher (indicating poorer quality of life) than controls with no lung disease (Yohannes et al., 1998).

Respiratory function
A lower value measured by the Shuttle Walk Test which is associated with morbidity was most strongly correlated with the Problems sub-scale (Hyland 1994). Small but significant correlation between changes observed by the Shuttle Walk Test and Treadmill Endurance Test and nine of the items within the instrument indicating an improvement in exercise associated with an improvement in QoL (Hyland et al., 1998).

No other measurement criteria reported

Shortened version
Following analysis of change in BPQ items before and after rehabilitation the magnitude of effect sizes determined the items selected for the shortened version: seven items for the Problems scale and 3 for the emotional evaluations. Scoring was proposed as a single scale (Hyland et al., 1998). Correlations between the full version

of the BPQ and the shortened version ranged from 0.74 for Emotional to 0.91 for Total score.

Table 5.9: Developmental and evaluation studies relating to the BPQ applied in patients with COPD

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
BPQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hyland (1994) UK	Development study COPD patients Interview administered		Construct ✓				
Yohann es et al., (1998) UK	CAL (151) Age: mean 78 Out-patients Interview administered		Construct ✓				

b) Chronic Respiratory Disease Questionnaire (CRQ)

Eighteen studies provide evidence of measurement properties for the CRQ. Five were with a UK population (Brightling et al., 2001; Harper et al., 1997; Singh et al., 2001; Williams et al., 2001, 2003). Four studies used self-completion of the questionnaire. These studies compared both interview-administered and self-report methods with patients (Schunemann et al., 2003; 2005; Williams et al., 2001; 2003).

The participants included in the studies were representative of COPD patients and sample sizes ranged from 24 to 156. Only one study recruited an adequate number of patients (487) (Wyrwich et al., 1999).

Reliability

Internal consistency was high for all domains (greater than 0.70) in studies by Harper et al., (1996) and Wyrwich et al., (1999).

Reproducibility was greater than 0.90 for individual analysis for all domains in Brightling et al., (2001); Desikan et al., (2002) and greater than 0.70 in Aaron et al., (2002).

ICC's were comparable for self-reported and interview administered versions of the instruments in Williams et al., (2001). In a small sample of patients with COPD (15) results were similar with the exception of Fatigue: 0.20 (Martin 1994).

Validity

Health status

The CRQ domains discriminated patients who had breathing problems in the last four weeks with statistically significantly lower scores than those patients who had no breathing problems in a study by Wyrwich et al., (1999).

Respiratory-specific patient-reported health instruments
A small study of 41 patients reported moderate correlation with the CRQ domains and patient- global ratings of similar constructs. The Fatigue domain was the only strongly correlated domain with Patient global rating fatigue 0.62 (Guyatt 1987). Moderate correlation was found for the CRQ and respiratory function and global ratings of change for dyspnoea, fatigue and emotions which were lower than hypothesised in the study by Guyatt et al., (1999). Small correlations were found with, Shuttle walk test 0.33, treadmill 0.29 and no correlation with FEV (Singh et al., 2001).

Internal validity
The four hypothesized factors were supported in the study by Wyrwich et al., (1999).

Validity of different versions and methods of administration
Two RCTs provide measurement properties for different versions and methods of administration (Schunemann et al., 2003, 2005). The CRQ-Self Administered version (CRQ-SA) and Interview administered (CRQ-IA) were further within- randomization to individualized and standardized Dyspnoea ratings. Overall there was greater correlation for the standardised component of IA and SA methods.

Further evidence is reported for self-reported vs. interview administered CRQ completion with no statistical difference in mean scores for Mastery and Fatigue but there was a statistically significant difference for Dyspnoea and Emotion (Williams et al., 2001).

Generic patient-reported health instruments
There was greater correlation for the 'standardised' Dyspnoea domain for both IA and SA methods with the SF-36 PCS. Other domain correlations for both methods were moderate to strong for similar items and domains (Schunemann et al., (2005).

Stronger correlations (greater than 0.60) was found between the SF-36 and CRQ related domains (PF and Dyspnoea; Vitality and Fatigue; RE, MH and Emotional function). There was moderate correlation with all SF-36 domains and the CRQ Mastery domain (Wyrwich et al., (1999). No correlation with SF-36 and the CRQ Mastery domain and only small to moderate for other domains in small study by Martin (1994)

Responsiveness
It was hypothesised that patients participating in a rehabilitation programme would improve following intervention. Two weeks following discharge substantial improvement was observed in scores on all four domains. Higher responsiveness statistics were found for CRQ than for other instruments in groups of patients distinguished in terms of level of change on the Transition Dyspnoea index. As hypothesized, correlations between changes in FEV and CRQ were uniformly higher than RAND and Oxygen Cost Diagram 0.55 vs. 0.28 and 0.43 (Guyatt (1987). In contrast to cross sectional validity there was no trend for higher correlation for the standardised dyspnoea component for either SA or IA. The individualised component was more responsive than the standardised. The SA was more responsive than the IA (Schunemann et al., 2003; 2005).

There were no statistically significant differences in responsiveness or longitudinal validity for the CRQ with the SGRQ according to whether patients were reminded of their baseline scores in a study by Schunemann et al., (2002).

Longitudinal changes were correlated between the CRQ and SGRQ but the strength of correlation less than 0.60 (de Torres et al., 2002). Longitudinal changes between CRQ and FEV and TDI were greater for all domains (greater than 0.60 (Aaron et al., 2002).

The CRQ- self-completion and interview administered versions were equally as responsive to change with large statistical and clinically significant changes in mean scores and no difference was observed in the magnitude of change between the two methods (Williams et al., 2003). However, the SRM's were higher for the CRQ-SA compared to the CRQ-IA in a study by Puhan et al., (2005). The CRQ standardized Dyspnoea domain was more responsive than the Fatigue, Emotions and Mastery domains but also the SGRQ and SF-36. The Emotional and Fatigue domains were also more responsive to change following a rehabilitation programme than the SF-36 similar domains.

Weak longitudinal correlation was found between the CRQ all domains with QWB scale and SIP with no correlation with the SG (Guyatt et al., 1999).

Large effect sizes (≥0.80) were observed for CRQ Dyspnoea (where patients were classified into severe and less severe breathlessness cases); Mastery: Moderate (<0.80) and small effect sizes (≥0.20 to <0.5) for Fatigue (Harper et al., 1997). Effect sizes representing responsiveness were 0.40 for Fatigue to 0.90 for Mastery in Guyatt et al., (1999).

The CRQ was responsive to change applying a one-SEM criterion with similar percents of change across most domains (Wyrwich et al., 1999). The responsiveness statistics for domains (dyspnoea, fatigue, emotion and mastery) were 2.2; 4.1; 2.5; 4.2 with greater than 1.5 indicating responsiveness by the authors (Aaron et al., 2002).

Clinically important difference

An expert physician panel established small, moderate and large clinically important change levels for the CRQ as follows: *Small change*: 2 (Fatigue) to 5 (Emotional function; *Moderate change*: 4 (Fatigue) to 10 (Emotional function); *Large change*: 6 (Fatigue) to 15 (Emotional function) (Wyrwich et al., 2003).

Precision

No floor or ceiling effects have been reported (Harper et al., 1996; Wyrwich et al., 1999)

Acceptability

More items were completed for the individualised version than the standardised version of the Dyspnoea domain (96 vs. 70%) on the SA and 98 vs. 75% on the IA. There was no missing data for the other domains (Schunemann et al., (2005). At least 80% of data were computable in Singh et al., (2001) and Wyrwich et al., (1999).

Feasibility

No evidence reported

Table 5.10: Developmental and evaluation studies relating to the CRQ applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
CRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Aaron et al., (2002) Canada	COPD (70) visiting ED Age: mean 70 ED Interview administered	Test re-test ✓	✓	✓			
Brightling et al., (2001) UK	COPD (61) Age: mean 66 Out-patients Interview administered	Test re-test ✓					
Desikan et al., (2002) USA	COPD patients (40) Age: range 41 to 71 Telephone interview	Test re-test ✓	✓				
de Torres et al., (2002) USA	COPD patients FEV<40% (37) Age: mean 63 Interview administered Out-patients			✓			
Guyatt (1987) Canada	COPD (41) participating in a rehabilitation programme Age no details Interview administered Out-patients		Construct ✓	✓			
Guyatt et al., (1991) Canada	Chronic Airflow Limitation (24) participating in a trial of bronchodilators Age: mean 66 Interview administered Out-patients			✓			

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
CRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Guyatt et al., (1999) Canada	Chronic airflow limitation (89) Age: no details Interview administered		Construct ✓	✓			
Harper et al., (1997) UK	Patients with COPD (156) Age: mean 67 Self-completed Out-patients	Internal consistency ✓ Test re-test ✓	Construct ✓	✓	✓	✓	
Martin (1994) USA	COPD (15) Age: mean 67 Interview administered Out-patients	Test re-test ✓	Construct ✓				
Puhan et al., (2006) Canada	COPD (177) participating in a rehabilitation programme Age: mean 69 Self completion Out-patients			✓			
Schunemann et al., (2002) Canada	COPD (85) participating in a RCT of blind vs. informed response administration of the CRQ and SGRQ Age: mean 66 Interview administered Out-patients			✓			
Schunemann et al., (2003) Canada	COPD (51) Age: mean 67 Interview vs. self administered Out-patients		✓	✓			

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
CRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Singh et al., (2001) UK	Patients with COPD participating in a rehabilitation programme (97) Age: mean 67 Interview administered Out-patients		Construct ✓	✓		✓	
Williams et al., (2001) UK	COPD (52) Age: mean 66 Interview vs. self-completion Out-patients		Construct ✓			✓	
Williams et al., (2005) UK	COPD (35) Age: mean 67 Interview vs. self-completion Out-patients			✓			
Wyrwich et al., (1999) USA	COPD (487) Age: mean 58 Telephone interview	Internal consistency ✓	Construct ✓	✓	✓	✓	

c) Functional Performance Inventory (FPI)

Three studies form the USA provide evidence of measurement properties (Larson et al., 1998; Leidy et al., 1999a, b). Two studies used a postal survey and one self-completion during an out-patients visit. Small numbers of patients were recruited in these studies (23 to 154).

Reliability
The FPI was reproducible in a study by Leidy et al., (1999a) with ICC values as follows: Total 0.87: BC 0.76; HM 0.89; PE 0.66; R 0.85; SpA 0.71; SoA 0.82.

Internal consistency of the FPI was high with alphas all above 0.70 for all domains (Larson et al., 1998; Leidy et al., 1999a) and 0.90 for Maintaining household (Larson et al., 1998; Leidy et al., 1999a) and FPI Total 0.96 (Leidy et al., 1999a).

Less than 3% for the total instrument inter-item correlations exceeded 0.50 and items with weak item-total correlations (less than 0.20) reflected low levels of physical activity from the Body Care domain and were consistent with ceiling effects noted in this scale (Larson et al., 1998).

Validity

Generic patient-reported health instruments
Strong correlation was observed for the FPI Total and SIP Physical (-0.60); PPI Body Care and SIP Physical and Total (-0.64, -0.62).The FPI Total correlated with SF-36 PF (0.69), FPI Body care, Maintaining household and Physical exercise were strongly correlated with the SF-36 PF. There was moderate correlation with other similar domains for the FPI and SF-36 and SIP (Larson et al., 1998).

FPI correlations were strong (0.60 and above) between the FPI and Functional Status Questionnaire for BC, HM, PE and Total. Weak correlation was reported for Spiritual activities. Moderate correlation was reported between the FPI and DASI (Leidy and Kapella 1999a).

Respiratory function
The range of correlations were weak to strong for FPI and respiratory function tests as follows: 0.12 for Spiritual activity to 0.63 for Total with: FEV, Bronchitis Emphysema Symptom Checklist (Leidy and Kapella1999a).The FPI Maintaining household was the only domain which was strongly associated with lung function tests in a small study of 23 patients with COPD (Leidy et al., 1999b). Strong correlation with self-reported functional performance was reported tin this study for all domains with exception of the Spiritual activity domain.

Responsiveness
No evidence reported.

Precision
There were no ceiling and floor effects for the Total FPI but floor effects were observed for the following domains of the FPI in a small study of seventy patients

with COPD: Spiritual activity (26%), Work/school (66%). Ceiling effects were 32% for Body Care and 18% for Spiritual activity (Larson et al., 1998).

Acceptability
Postal survey to 293 patients with COPD resulted in a 66% response rate (Leidy et al., 1999a).

23% missing data for the Work/School domain was reported in Larson et al., (1998). 3% to 6% of data were missing for all other domains except Body care and Spiritual activity with no missing data.

Feasibility
No evidence reported.

Table 5.11: Developmental and evaluation studies relating to the FPI applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
FPI		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Leidy and Kapella (1999a) USA	COPD (154) Age: mean 64 Postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓	✓	✓		
Leidy et al., (1999b) USA	COPD (23) Age: mean 64 Self completion Out-patients		Construct ✓				
Larson et al., (1998) USA	COPD (72) Age: mean 70 Telephone and postal survey		Construct ✓				

d) SGRQ

Ten studies provide evidence of the measurement properties for the SGRQ, five from the UK (Harper 1997; Jones et al., 1991; Okubadejo et al., 1996; Singh et al., 2001; Wilson et al., 1997). Five studies administered the questionnaire during an interview in out-patients, others were self-completed. The number of participants in all studies was less than 160.

Reliability
Four studies provided evidence of reliability (Jones et al., 1991; Desikan et al., 2002; Harper et al., 1997; Wilson et al., 1997).

High levels of test-retest reliability were reported for patients with COPD (Jones 1991) (Total 0.92). Test re-test reliability coefficients were greater than 0.70 but did not reach 0.90 in Harper et al., (1997 and Desikan et al., (2002) with the exception of

the Impact domain (0.46) (Harper et al., 1997). Higher levels of reproducibility were reported in Wilson et al., (1997) with all ICC's >0.90.

Internal consistency alphas were acceptable in Harper et al., (1997) and Wilson et al., (1997) (0.71 to 0.92).

Twenty-seven of 50 (54%) item total correlations for the SGRQ were greater than 0.40 Harper et al., (1997) but no details were provided of items inferring lack of homogeneity.

Validity
Eight studies provide evidence to support the validity of the SGRQ.

Health status
The SGRQ Symptoms domain discriminated between patients with respiratory symptoms and those without but was weakly correlated with physiological measures, dyspnoea grade, mood state and SIP scores (O.07 to 0.12) (Jones 1991). Moderate correlations were reported as hypothesised for the Activity and Impact domains and MRC dyspnoea grade, physical function test, psychological functioning and general health. Stronger correlations were reported for the Impact domain and anxiety and depression (Jones et al., 1991).

The SGRQ Activity, Impact and Total distinguished the presence of co morbidity in Harper et al., (1997) with large effect sizes (≥0.80) (where patients were classified into severe and less severe breathlessness cases) for SGRQ Total, Impact and Activity; small effect sizes ≥0.20 to <0.5) for SGRQ Symptoms (Harper et al., 1997).

The Symptoms score was significantly higher in patients with wheeze on most days compared with those who had occasional wheeze and higher in patients reporting more than three infections in the past year (Wilson et al., 1997).

Healthcare utilisation
Poorer quality of life as indicated by high SGRQ scores were related to a greater likelihood of hospitalisation, ER and primary care visits in (Alemayehu et al., 2002).

COPD-specific patient-reported health instruments
All correlations were strong between the SGRQ domains and the AQ30 and AQ20 (Alemayehu et al., 2002).

The SGRQ Total was moderately correlated with BPQ: -0.59; CRQ -0.39; and Global QOL -0.48. (Singh et al., 2001).

Generic patient-reported health instruments
Strong correlation was reported as predicted between the SGRQ Total and domains with the SF-36 PCS with smaller to moderate correlation for the MCS (Wilson et al., 1997). Similar results were reported for the SIP Physical and Psychological scores. Small to moderate correlations were reported between the SGRQ Total, HAD anxiety and HAD depression (0.20 for Symptoms to 0.58 Impact (Wilson et al., 1997).

Respiratory function
Three studies evaluated the relationship between HRQL with the SGRQ and respiratory function reporting small to moderate correlations which were less than expected (Okubadejo et al., 1996; Singh et al., 2001; Wilson et al., 1997). Respiratory function was measured using FEV, Shuttle walk test, Treadmill, and oxygen tension (PaO2 and PaCO2). No correlation was observed between FEV and Symptoms (0.03) in Okubadejo et al., (1996).

Responsiveness
There was little change in a small group of patients with COPD in a study by Jones et al., (1991) for physiological variables, dyspnoea grade and SIP scores (Jones et al., 1991). Changes in SGRQ scores were most positively correlated with dyspnoea grade (0.22). With small changes in health in a group of patients over six months (Wilson et al., 1997) hypothesised correlations were stronger for the SGRQ and Physical component score (SF-36) than changes in the Mental component. The SGRQ Total score and Impacts domain were more responsive than the SF-36 following a rehabilitation programme but less so than the CRQ in a concurrent evaluation (Puhan et al., 2006). However for the other domains of the SGRQ, the SRM's were similar to the SF-36. The correlations with the SGRQ and MRC dyspnoea scale were also stronger for Impact and Total (0.43; 0.38) than Symptoms and Activity (0.15; 0.27).

The SGRQ did not correlate as strongly as the CRQ with changes Fatigue with Shuttle walk test and Treadmill endurance in a concurrent evaluation (Singh et al., 2001). The SGRQ and the CRQ had moderate correlations of change in a concurrent evaluation by de Torres et al., (2002).

Using a transition question of patient perceived change, statistically significant differences between groups with different levels of change were reported for Symptoms, Activity and Total but not for the Impact domain (Harper et al., 1997). SRM's for [6 months and 12 months] were as follows: SGRQ Total Large (≥0.80), SGRQ Symptoms and Activity moderate (≥0.5 to <0.80) (Harper et al., 1997).

Precision
A normal distribution and no floor or ceiling effects were found with patients with COPD (Jones et al., 1991; Wilson et al., 1997). Floor effects for Activity (25.9%) were reported in Harper et al., (1997).

Acceptability and Feasibility
Higher completion rates were reported when the SGRQ was administered in out-patients but in a concurrent evaluation the SGRQ had the lowest completion rates compared to the EQ-5D and SF-36 (Harper et al., 1997). Further evidence of missing data were reported in Singh (2001) with 68% of data complete and only 50% of questionnaire computable in a postal survey by Alemayehu et al., (2002) although this study included the AQ20 and AQ30 and the authors only included in analysis responses to all questionnaires.

Table 5.12: Developmental and evaluation studies relating to the SGRQ applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
SGRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Jones et al., (1991) UK	Studies including patients with asthma and COPD COPD patients (20) Age: mean 66 Interview but patient completed Out-patients	Test re-test ✓	✓	✓			
	Patients with chronic airflow limitation (141) Age: mean 63		Construct ✓	✓	✓		
Alemayehu et al., (2002) USA	COPD (181) Age: mean 68 Postal survey		Construct ✓				
Desikan et al., (2002) USA	COPD patients (40) Age: range 41 to 71 Telephone interview	Test re-test ✓	✓				
de Torres et al., (2002) USA	COPD patients FEV<40% (37) Age: mean 63 Interview administered Out-patients			✓			
Harper et al., (1997) UK	Patients with COPD (156) Age: mean 67 Self-completed Out-patients	Internal consistency ✓ Test re-test ✓	✓	✓	✓	✓	

Study/ County	Population (N) Age Method of administration Setting	Measurement properties					
SGRQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Puhan et al., (2006) Canada	COPD (177) participating in a rehabilitation programme Age: mean 69 Self completion Out-patients			✓			
Schunemann et al., (2002) Canada	COPD (85) participating in a RCT of blind vs. informed response administration of the CRQ and SGRQ Age: mean 66 Interview administered Out-patients			✓			
Singh et al., (2001) UK	Patients with COPD participating in a rehabilitation programme (97) Age: mean 67 Interview administered Out-patients		Construct ✓	✓			

e) SOLQ

Three studies provided evidence of measurement properties, two of which were large studies by postal survey with over 1000 and 3000 patients with COPD. All patients were representative of the COPD population.

Reliability

High levels of internal consistency were reported for postal administration (0.79 for Emotional function to 0.93 for Physical) (Tu et al., 1997).

Test re-test reliability over a four month period was 0.64 for Treatment satisfaction to 0.87 Physical functioning (Tu et al., 1997).

Validity

Healthcare utilisation

In a large postal survey using the SOLQ, patients Physical function scores which were in the 0-25th percentile were six times more likely to have a COPD related hospitalisation within a year of baseline measurement. For other domains Odds Ratios for hospitalisation were 3.0 for Emotional function and 3.2 for Coping Skills (Fan et al., 2002).

COPD-specific patient-reported health instruments

Hypothesised correlations between the SOLQ and the CRQ were supported in Tu et al., (1997) but the expected strength of correlation not specified a priori. Correlations were small to moderate for all domains. Treatment satisfaction was strongly correlated with overall satisfaction measured by the Patient Satisfaction Questionnaire (0.54) as hypothesised.

Generic patient-reported health instruments

There was large hypothesised correlation between the SOLQ Physical functioning domain and the SF-36 PCS but only moderate correlation between the similar emotional domains for each instrument (Belza et al., 2005).

Respiratory function

Small correlations were reported between the SOLQ and lung function (FEV and 6 min Walk test (Tu et al., 1997, Belza et al., 2005).

Responsiveness

The SOLQ was able to detect change with high responsiveness statistics for each domain (0.78 to 0.87) (Tu et al., 1997). Correlation of change with the SF-36 and the SOLQ were moderate with none greater than 0.60 (Belza et al., 2005).

Clinically important difference

MCID was reported using patient-reported assessment of improved, unchanged or deteriorated and a change of 5 points was observed (Tu et al., 1997) and statistically significant change in scores post rehabilitation (Physical 3.79; Emotional function 9.20; Coping skills 7.26) (Belza et al., 2005).

Precision

No evidence reported

Acceptability
There was a 60% response rate to the SOLQ postal survey (Fan et al., 2002).

Feasibility
No evidence reported

Table 5.13: Developmental and evaluation studies relating to the SOLQ applied in patients with COPD

Study/ County	**Population (N) Age Method of administration Setting**	**Measurement properties**					
SOLQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Tu et al., (1997) USA Development study	COPD: 203; 97; 920 Age: mean 70 Postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓				
Fan et al., (2002) USA	COPD (3282) Age: mean 65 Postal general health survey		Construct ✓			✓	
Belza et al., (2005) USA	COPD (58) participating in a rehabilitation programme Age: mean 66 Self completed Out-patients		Construct ✓	✓			

Other COPD-specific instruments identified from the review.

The following table provides an overview of other records of COPD-specific instruments identified of either newly developed instruments or single study reporting of measurement properties and/or evaluation.
Table 5.14

Instrument/ reference	*Population (N) Age Method of administration Setting*	*Reliability*	*Validity*	*Responsiveness*	*Precision*	*Acceptability*	*Feasibility*	*Comments* *No other records identified unless stated*
Single studies								
AQ30 and AQ20 Chen et al., (2006)	COPD (352)		✓					Validity of the AQ20 established with moderate correlation with the SF-12 No advantage over the AQ30 over the AQ20. Correlations reported with respiratory function, SGRQ and AQLQ Junipers. AQ30 and AQ20 evaluated in patients with Asthma (Quirk 1994) Foreign language evaluations and one UK translation to Bengali (Griffiths 2000)
Manchester Respiratory Activities of Daily Living Questionnaire Yohannes et al., (2000) UK	COPD (188)	✓	✓					Instruments developed from selected items from Nottingham Extended ADL Questionnaire and Breathing Problems Questionnaire. Domains focus on mobility and activity limitations
Pulmonary Functional Status Score Weaver et al., (1998) USA	COPD (365)	✓	✓					Three factors: Daily activities/social functioning; Psychological functioning and sexual functioning

Table 5.15: Instruments excluded from the review

Instruments excluded	***Reason for exclusion***
The Breathless, Cough and Symptoms Scale	Not multidimensional. Measures the severity of symptoms
Baseline Dyspnoea Index/ Transitional Dyspnoea Index	Dimension-specific. Assesses the severity of dyspnoea
COPD Severity Score	Not multidimensional
San Diego Shortness of Breath Questionnaire	Assesses self-reported shortness of breath while performing activities of daily living
Pulmonary Functional Status and Dyspnoea Questionnaire	Assesses intensity of dyspnoea and affect on activities
London Chest ADL Scale	Measures breathlessness when carrying out activities

SUMMARY - GENERIC INSTRUMENTS

A total of seventeen articles were included in this review of evaluation studies of generic instruments for COPD. Seven generic instruments were identified that had been evaluated with people with COPD (SF-36; SF-20; SF-12; EQ-5D; SIP; Dartmouth COOP and NHP). The most frequently reported instrument evaluated was the SF-36 with evidence of all measurement properties.

The number of participants included each was generally small with only five studies recruiting greater than 200 participants: Ruffin et al., (2000), Sprenkle et al., (2004) and Katz et al., (2005) for the SF-36; Hazell et al., (2003) for the EQ-5D and McSweeny et al., (1982) for the SIP. Six studies were conducted in the UK with one using the SIP, 3 using the EQ-5D and 2 studies the SF-36. Although in all studies the patients completed the questionnaires, many were administered during an interview. Two studies used a postal survey as the method of administration and one using a telephone interview. Patients were recruited from primary care practices or out-patient settings. No study was conducted whilst a patient was in hospital. Patients all had a general diagnosis of COPD, but some studies recruited patients with specific lung disease such as emphysema, bronchiectasis and airflow limitation including asthma.

Generic instruments included domains, items and scoring procedures are detailed in Chapter 3, Tables 1 and 2. All instruments are multi-dimensional with an average of six domains similar in construct.

Only two studies reported internal consistency reliability for the SF-36 with high alpha levels. Reproducibility was generally lower than the 0.70 threshold for group testing for the SF-36, EQ-5D and COOP charts. No other generic instruments have reliability evidence in the COPD population. Construct validity is supported for all generic instruments included in the review with extensive evaluation for the SF-36. The SF-36 PCS scores were predictive of healthcare utilisation and discriminated patients with different disease severities. The MCS was not predictive or discriminative. The EQ-5D too discriminated between different health states and respiratory disease severities. The internal structure was supported in one study for the SF-36. Strong correlations between the SF-36 and other generic instruments (NHP and COOP Charts) and the CRQ COPD-specific instrument were reported for scales of similar domains.

Evidence of responsiveness is reported for the SF-36, EQ-5D and COOP charts with the exception of the SF-36 MCS and the EQ-5D VAS which were not responsive to change. Recommended clinically important differences were reported from consensus group proceedings for the SF-36 only. Ceiling and floor effects were reported for the SF-36 but not for the EQ-5D. The other instruments were not evaluated for precision.

The COOP Charts was the most easy to administer instrument with no missing data. Missing data was reported for the SF-36 and EQ-5D particularly the VAS.

SUMMARY – COPD-SPECIFIC INSTRUMENTS

A total of twenty-nine studies were included in the review which reported results from evaluations of COPD-specific instruments. Five instruments were included in the review which had undergone different aspects of measurement performance (BPQ, CRQ, FPI, SOLQ and the SGRQ). A further three instruments were identified which had undergone one evaluation (Table 5.15) and 6 instruments were excluded (Table 5.16). The CRQ was the most frequently and comprehensively evaluated instrument (17) followed by the SGRQ (10).

Most studies recruited less than 200 participants with the exceptions of Wyrwich (1999) for the CRQ (477) and two studies with the SOLQ (Tu 1997) 920 and Fan (2002) 3282 patients. Generally the patients were older than 50 years of age and representative of the COPD population. Of the twenty-nine studies, 11 were conducted in the UK (BPQ, CRQ and SGRQ). The FPI and SOLQ were the only instruments to be completely self reported either by post or during and out-patient visit. The BPQ, CRQ and SGRQ were generally interview administered.

All instruments included scales to assess Physical functioning, psychological, social well-being, and role activities. Personal construct and treatment satisfaction were common domains in most. The number of items ranged from 17 (SGRQ) to 65 (FPI). Total and domain scores are computable for the FPI and SGRQ.

All instruments included in the review have evidence of reproducibility supporting application in studies involving groups of patients. All instruments have reported high levels of internal consistency.

All instruments were assessed for validity through comparison with other instruments with similar constructs and predictions a priori about hypothetical relationships generally supported between similar domains with generic instruments particularly for the CRQ and SGRQ. There was often a poor relationship between COPD-specific measures and respiratory function. Evidence of responsiveness was reported for all instruments except the BPQ. In concurrent evaluations, the instruments included in the review performed better than generic comparator instruments.

Although for all instruments, complete data enabled analysis, many studies administered the questionnaires during an interview. The individualised version of the CRQ Dyspnoea domain had better completion rates than the standardised version.

There is good evidence of the measurement and practical properties for the CRQ and SGRQ with both including UK evaluations. There is limited evidence for the BPQ, FPI and SOLQ; further evaluations are required.

Some instruments have been identified in this review which report only one evaluation. There is insufficient evidence therefore to make firm recommendations about these at present.

Instrument	**Measurement properties**	**Availability**: Royalty; Scoring methods and interpretation guide	**Acceptability/Feasibility**: Patient acceptability Staff acceptability
COPD-specific			
BPQ	Developed in the UK (2 UK evaluations pre 2000) Limited evidence of reliability and validity	No details	Suitable for self-completion but 33 items 10 to 20 minutes completion No details of patient acceptability or feasibility
CRQ	Developed in Canada. 16 evaluations, 5 UK Several studies evaluating most measurement criteria Several studies evaluating most measurement criteria Good evidence of reliability, validity and responsiveness supporting application	License Agreement required	30 minutes completion time Interview administered Acceptable to patients although the individualised Dyspnoea domain difficult for some
FPI	.Developed in UK Three UK evaluations pre-2000 Some evidence of reliability and validity although two studies with small number of patients	No details	Self reported 68 items related to functional impairment. Some evidence of missing data
SGRQ	Developed in the UK Eight evaluations with people with COPD. Has been used with patients with Asthma Several studies evaluating most measurement criteria Good evidence of reliability and validity	Contribution to the St. George's Research Fund is requested from commercial organisations using the instrument. Permission should be obtained from the authors Scoring algorithms and calculators are available from the developers	Self- report but recommended interview administered 8- 15 minutes to complete Some evidence of missing data

Generic			
Instrument	**Measurement properties**	**Availability**: Royalty; Scoring methods and interpretation guide	**Acceptability/Feasibility**: Patient acceptability Staff acceptability
SF-36	Two UK evaluations Several studies evaluating most measurement criteria. Good evidence of reliability, validity and responsiveness supporting application	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self- report 10 to 15 minutes to complete Some difficulties experiences with completion
SF-20	One USA evaluation	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self-report 5 minutes completion Acceptable to patients
SF-12	One USA evaluation	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self-report 5 minutes completion Acceptable to patients
SIP	Limited evaluations and evidence of validity 2 evaluations, one UK		
Dartmouth COOP	Limited evidence One evaluation from Australia evaluating reliability, validity and responsiveness		
EuroQol	Three evaluations in the UK Some evidence of measurement properties	Permission and licensing should be obtained from the authors Scoring algorithms and manual are available from the developers	Self-report 5 minutes completion VAS higher proportion of missing responses
NHP	Limited evaluations and evidence of measurement properties. It has only been used as a reference measure.		

DISCUSSION AND RECOMMENDATIONS

Many evaluations have been identified in this review of both generic and COPD-specific instruments with patients with different disease severities. The evaluations were conducted mainly in an out-patient setting and although all instruments were completed by the patients, some were administered during interviews. There are limited UK evaluations. Most have been applied in the USA or Canada.

The methodological quality of the studies was variable. Adequate reporting of data enabled abstraction for this review but the small sample sizes of some studies may inflate results. In addition, although most studies predicted associations for example between similar domains, none specified the strength of association a priori.

The SF-36 is the most widely evaluated generic instrument. Amongst the COPD-specific instruments and the Chronic Respiratory Questionnaire and St. George's Respiratory Questionnaire have extensive evidence of measurement properties but further UK evaluations would strengthen evidence of their applicability for UK populations. There is limited evidence for the BPQ, FPI and SOQL. Further evaluations are required.

The generic instruments chiefly the SF-36, included in the review adopt a multi-dimensional perspective to the measurement of patient-reported health.

All reviewed COPD-specific instruments address multi-dimensional aspects of health-related quality of life. All include the assessment of symptoms; physical functioning and most also include psycho-social well-being. Other frequently assessed dimensions include the impact of COPD on role activities, personal constructs and treatment satisfaction.

Formal measurement properties are more commonly a feature of evaluative studies of instruments. The practical properties of instruments are less widely explored or defined in evaluations. No direct evidence was found for feasibility for staff in terms of time to administer, training required and processing of results. Some evidence is reported relating to the percentage of missing responses relating to patient acceptability. The instruments included in the review have been administered by different methods and many have adopted a postal survey or self completion in an out-patient setting.

The lack of studies directly comparing the performance of different generic instruments is disappointing. The SF-36 and NHP performed equally well in one concurrent evaluation (Crockett 1996).

Concurrent evaluations of COPD-specific instruments were dominated by comparative evaluations of modifications of the CRQ and SGRQ. Both instruments have evidence of high levels of internal consistency and reproducibility. Similar strength of correlation is reported for both instruments and generic measures supporting construct validity and both discriminated patients with different health status. The CRQ was slightly more responsive compared to the SGRQ with higher longitudinal correlations with respiratory function. The CRQ standardised dyspnoea domain version was more responsive than the individualised version. Although a

greater number of evaluations were identified for the CRQ, with some in the UK, the SGRQ has been developed in the UK and performance is comparable. In addition, the SGRQ has been evaluated with patients with asthma.

However, several studies report the concurrent evaluation of generic and COPD-specific instruments. Good evidence supports the reliability and validity of both generic (SF-36) and COPD-specific (CRQ, SGRQ) measures, supporting their combined use in people with COPD. However, and as expected, consistently higher levels of responsiveness were reported for the COPD-specific instruments.

Recommendations

Synthesising the available primary evidence reported in this review and extrapolating evidence from concurrent evaluations supports the use of both generic and COPD-specific patient-reported health instruments for people with COPD. The SF-36 is recommended as a generic instrument for the broad evaluation of health-related quality of life. Further evaluations are required, particularly concurrent evaluations of different generic instruments to inform further recommendations and for the UK population.

COPD-specific instruments, particularly the CRQ and SGRQ, are recommended and different administration methods have been evaluated. Further evaluations are needed to support the use of these instruments specifically in the UK.

REFERENCES

Aaron SD, Vandemheen KL, Clinch JJ, Ahuja J, Brison RJ, Dickinson G *et al*. Measurement of short-term changes in dyspnea and disease-specific quality of life following an acute COPD exacerbation. *Chest: the Cardiopulmonary Journal* 2002; **121:**688-96.

Alemayehu B, Aubert RE, Feifer RA, Paul LD. Comparative analysis of two quality of life instruments for patients with chronic obstructive pulmonary disease. *Value in Health* 2002; **5:**436-41.

Belza B, Steele BG, Cain K, Coppersmith J, Howard J, Lakshminarayan S.Seattle Obstructive Lung Disease Questionnaire: sensitivity to outcomes in pulmonary rehabilitation in severe pulmonary illness. *J Cardiopulm Rehabil. 2*005 Mar-Apr; **25**(2):107-14.

Benzo R, Flume PA, Turner D, Tempest M. Effect of pulmonary rehabilitation on quality of life in patients with COPD: the use of SF-36 summary scores as outcomes measures. *Journal of Cardiopulmonary Rehabilitation* 2000; **20:**231-4.

Brightling CE, Monterio W, Green RH, Parker D, Morgan MDL, Wardlaw AJ *et al*. Induced sputum and other outcome measures in chronic obstructive pulmonary disease: safety and repeatability. *Respiratory Medicine* 2001; **95:**999-1002.

Chen H, Eisner MD, Katz PP, Yelin EH, Blanc PD Measuring disease-specific quality of life in obstructive airway disease: validation of a modified version of the airways questionnaire 20.
Chest. 2006 Jun; **129**(6):1644-52.

Crockett AJ, Cranston JM, Moss JR, Alpers JH. The MOS SF-36 Health Survey questionnaire in severe chronic airflow limitation: comparison with the Nottingham Health Profile. *Quality of Life Research* 1996; **5:**330-8.

De Torres JP, Pinto-Plata V, Ingenito E, Bagley P, Gray A, Berger R *et al*. Power of outcome measurements to detect clinically significant changes in pulmonary rehabilitation of patients with COPD. *Chest: the Cardiopulmonary Journal* 2002; **121:**1092-8.

Desikan R, Mason HL, Rupp MT, Skehan M. Health-related quality of life and healthcare resource utilization by COPD patients: a comparison of three instruments. *Quality of Life Research* 2002; **11:739-**51.

Eaton T, Lewis C, Young P, Kennedy Y, Garrett JE, Kolbe J. Long-term oxygen therapy improves health-related quality of life. *Respiratory Medicine* 2004; **98:**285-93.

Fan V, Curtis J et al. Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases. Chest 2002 **122**:429-436.

Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax* 1987; **42:**773-8.

Guyatt GH, Townsend M, Keller J, Singer J, Nogradi S. Measuring functional status in chronic lung disease: conclusions from a randomized control trial. *Respiratory Medicine* 1991; **85:**17-21.

Guyatt GH, King DR, Feeny DH, Stubbing D, Goldstein RS. Generic and specific measurement of health-related quality of life in a clinical trial of respiratory rehabilitation. *Journal of Clinical Epidemiology* 1999; **52:**187-92.

Hazell M, Frank T, Frank P. Health related quality of life in individuals with asthma related symptoms. Respir Med. 2003; 97(11):1211-8.

Halpin DM, Miravitlles M. Chronic obstructive pulmonary disease: the disease and its burden to society. *Proc Am Thorac Soc.* 2006 Sep; **3**(7):619-23

Harper R, Brazier JE, Waterhouse JC, Walters SJ, Jones NMB, Howard P. Comparison of outcome measures for patients with chronic obstructive pulmonary disease (COPD) in an outpatient setting. *Thorax* 1997; **52:**879-87.

Hyland ME, Bott J, Singh SJ, Kenyon CAP. Domains, constructs and the development of the breathing problems questionnaire. *Quality of Life Research* 1994; **3:**245-56.

Hyland M.E.[1]; Singh S.J.[2]; Sodergren S.C.[1]; Morgan M.P. Development of a Shortened Version of the Breathing Problems Questionnaire Suitable for Use in a Pulmonary Rehabilitation Clinic: A Purpose-Specific, Disease-Specific Questionnaire *Quality of Life Research*, Volume **7**, Number 3, 1998 , pp. 227-233(7)

Jones PW, Quirk FH, Baveystock CM. The St George's Respiratory Questionnaire. *Respiratory Medicine* 1991; **85:**25-31.

Katz PP, Eisner MD, Yelin EH, Trupin L, Earnest G, Balmes J, Blanc PD. Functioning and psychological status among individuals with COPD. *Quality of Life Research* 2005; **14**; 1835-1843.

Larson JL, Kapella MC, Wirtz S, Covey MK, Berry J. Reliability and validity of the functional performance inventory in patients with moderate to severe chronic obstructive pulmonary disease. *Journal of Nursing Measurement* 1998; **6:**55-73.

Leidy NK, Knebel AR. Clinical validation of the Functional Performance Inventory in patients with chronic obstructive pulmonary disease. *Respiratory Care* 1999a; **44:**932-9.

Leidy NK. Psychometric properties of the functional performance inventory in patients with chronic obstructive pulmonary disease. *Nurs Res* 1999b;48(1):20-8.

Mahler DA, .Mackowiak JI. Evaluation of the short-form 36-item questionnaire to measure health-related quality of life in patients with COPD. *Chest: the Cardiopulmonary Journal* 1995; **107:**1585-9.

Martin LL. Validity and reliability of a quality of life instrument: the Chronic Respiratory Disease Questionnaire. *Clinical Nursing Research* 1994; **3:**146-56.

McSweeny AJ, Grant I, Heaton RK, Adams KM, Timms RM. Life quality of patients with chronic obstructive lung disease. Arch Intern Med. 1982; **142:**473-478.

NICE (2004) Chronic obstructive pulmonary disease - Management of chronic obstructive pulmonary disease in adults in primary and secondary care. Clinical Guideline

Okubadejo AA, Jones PW, Wedzicha JA. Quality of life in patients with chronic obstructive pulmonary disease and severe hypoxaemia. Thorax 1996; **51:**44-7.

Paterson C, Langan CE, McKaig GA, Anderson PM, MacLaine GDH, Rose LB *et al*. Assessing patient outcomes in acute exacerbations of chronic bronchitis: the Measure Your Medical Outcome Profile/MYMOP, Medical Outcomes Study 6-item general health survey/MOS-6A, and EuroQol/EQ-5D. *Quality of Life Research* 2000; **9:**521-7.

Puhan MA, Scharplatz M, Troosters T, Steurer J. Respiratory rehabilitation after acute exacerbation of COPD may reduce risk for readmission and mortality - A systematic review. *Respiratory-Research* 2005; **6:**:12p.

Quirk FH, Jones PW. Patients' perception of distress due to symptoms and effects of asthma on daily living and an investigation of possible influential factors. *Clin Sci* 1990;79:17-21.

Quirk FH, Baveystock CM, Wilson RC, Jones PW. Influence of demographic and disease related factors on the degree of distress associated with symptoms and restrictions on daily living due to asthma in six countries. *Eur Respir J 1*991;4:167-71.

Ruffin RE, Wilson DH, Chittleborough CR, Southcott AM, Smith BJ, Christopher DJ. Multiple respiratory symptoms predict quality of life in chronic lung disease: a population-based study of Australian adults. *Quality of Life Research* 2000; **9:**1031-9.

Schunemann HJ, Guyatt GH, Griffith LE, Stubbing D, Goldstein RS. A randomized controlled trial to evaluate the effect of informing patients about their pre-treatment responses to two respiratory questionnaires. *Chest: the Cardiopulmonary Journal* 2002; **122:**1701-8.

Schunemann HJ, Griffith LE, Jaeschke R, Goldstein RS, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *Journal of Clinical Epidemiology* 2003; **56:**1170-6.

Schunemann HJ, Griffith LE, Stubbing D, Goldstein RS, Guyatt GH. A clinical trial to evaluate the measurement properties of two direct preference instruments

administered with and without hypothetical marker states. *Medical Decision-Making* 2003; **23:**140-9.

Schunemann HJ, Griffith LE, Jaeschke R, Goldstein RS, Stubbing D, Austin P *et al*. A comparison of the original chronic respiratory questionnaire with a standardized version. *Chest: the Cardiopulmonary Journal* 2003; **124:**1421-9.

Schunemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic Respiratory disease Questionnaire (CRQ). *COPD-Journal-of-Chronic-Obstructive-Pulmonary-Disease* 2005; **2:**81-9.

Singh SJ, Sodergren SC, Hyland ME, Williams JEA, Morgan MDL. A comparison of three disease-specific and two generic health-status measures to evaluate the outcome of pulmonary rehabilitation in COPD. *Respiratory Medicine* 2001; **95:**71-7.

Sprenkle MD, Niewoehner DE, Nelson DB, Nichol KL. The veteran's short form 36 questionnaire is predictive of mortality and health-care utilization in a population of veterans with a self-reported diagnosis of asthma or COPD. *Chest*- 2004; **126:**81-9.

Tu SP, McDonell MB, Spertus JA, Steele BG, Fihn SD. A new self-administered questionnaire to monitor health-related quality of life in patients with COPD. *Chest: the Cardiopulmonary Journal* 1997; **112:**614-22.

Viramontes JL,.O'Brien BJ. Relationship between symptoms and health-related quality of life in chronic lung disease. *Journal of General Internal Medicine* 1994; **9:**46-8.

Weaver TE, Narsavage GL, Guilfoyle MJ. The development and psychometric evaluation of the pulmonary functional status scale: an instrument to assess functional status in pulmonary disease. *Journal of Cardiopulmonary Rehabilitation* 1998; **18:**105-11.

Williams JEA, Singh SJ, Sewell L, Guyatt GH, Morgan MDL. Development of a self-reported Chronic Respiratory Questionnaire/CRQ-SR. *Thorax* 2001;**56**:954-9.

Williams J E A, S J Singh, L Sewell, and M D L Morgan**.** Health status measurement: sensitivity of the self-reported Chronic Respiratory Questionnaire (CRQ-SR) in pulmonary rehabilitation**.** *Thorax*, June 1, 2003; 58(6): 515 - 518.

Wilson CB, Jones PW, O'Leary CJ, Cole PJ, Wilson R. Validation of the St. George's Respiratory Questionnaire in bronchiectasis. *American Journal of Respiratory and Critical Care Medicine* 1997;**156**:536-41.

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care* 1999;**37**:469-78.

Wyrwich KW, Fihn SD, Tierney WM, Kroenke K, Babu AN, Wolinsky FD. Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease: an expert consensus panel report. *Journal of General Internal Medicine* 2003;**18**:196-202.

Yohannes AM, Roomi J, Waters K, Connolly MJ. A comparison of the Barthel Index and Nottingham Extended Activities of Daily Living scale in the assessment of disability in chronic airflow limitation in old age. *Age and Ageing* 1998;**27**:369-74.

Yohannes AM, Roomi J, Waters K, Connolly MJ. Quality of life in elderly patients with COPD: measurement and predictive factors. *Respiratory Medicine* 1998;**92**:1231-6

Yohannes AM, Roomi J, Winn S, Connolly MJ. The Manchester Respiratory Activities of Daily Living Questionnaire: development, reliability, validity, and responsiveness to pulmonary rehabilitation. *Journal of the American Geriatrics Society* 2000;**48**:1496-500.

Chapter 6: Patient-reported Health Instruments used for People with Diabetes

Introduction

Diabetes is a disorder of glucose metabolism caused by a lack of the pancreatic hormone insulin, which results in the accumulation of sugar in the bloodstream (hyperglycaemia) and the appearance of sugar in the urine. Symptoms include thirst, fatigue, weight loss, and excessive urination. The failure to metabolise glucose leads to the breakdown of fats in the body as an alternative source of energy; this process disturbs the acid-base balance in the body and results in the accumulation of ketones in the blood (ketosis) which, if untreated, can lead to convulsions, coma, and death.

There are two main categories of the disease: Type 1, or insulin-dependent diabetes mellitus (IDDM) and Type 2, non-insulin-dependent diabetes mellitus (NIDDM). In Type 1 diabetes, which begins in childhood or adolescence, genetic factors and autoimmune processes damage the insulin-producing (beta) cells in the pancreas, so that patients depend on insulin injections for their survival. Type 2, also called 'mature onset diabetes', generally appears after the age of 40 and also has a hereditary component; Type 2 diabetics usually retain some beta cell function but show insulin resistance, often exacerbated by obesity. In the initial stages of the disease, Type 2 diabetes may be treatable with a combination of diet and exercise alone; in more severe or advanced cases, oral hypoglycaemics and, eventually, insulin injections may be required.

Sufferers face many difficulties, notably self-management of what may be a very complex treatment regimen. Type 1 diabetics have to exercise careful control of their diet balanced with activity, in order to avoid a fall in blood sugar (hypoglycaemia) which can cause dizziness, confusion, and convulsions – symptoms ranging from the unpleasant to the terrifying, and potentially fatal. Patients receiving intensive insulin therapy may have to monitor their blood glucose and inject themselves several times a day. Type 2 diabetes may be asymptomatic at first, so that adherence to dietary restrictions and other lifestyle changes can seem unnecessarily burdensome. In the long term, both types of diabetes are associated with an array of complications, including repeated infections, cardiovascular and peripheral vascular disease, sexual dysfunction, kidney and nerve damage, and loss of vision. Diabetes is a leading cause of blindness, lower extremity amputation, and premature death.

Measuring health-related quality of life (HRQoL) as perceived by diabetic patients is essential if informed and rational choices are to be made amongst the wide range of treatments available, and in order to tailor these to the needs of individual patients. HRQoL measurement in diabetes presents particular challenges to researchers and health-care providers, given the complexities of the condition. There is considerable evidence to show that, in addition to the multiple physical impacts of the disease and its treatment, psychosocial factors, such as depression, social network, and family relationships, significantly affect the course of the disease, and vice versa (Davis et al., 1988; Rosenthal et al., 1998; Lustman et al., 2000). A plethora of measures has been developed for use with diabetes – a recent review (Skovlund, 2005) identified over 150 validated instruments – and choosing between them is far from straightforward.

The following review provides current information available of the patient-reported health questionnaires used to measure health-related quality of life with patients with diabetes.

Search terms and results: identification of articles
At the time of the review, the PHI database contained 12,562 records (up to June 2005). An initial search of record abstracts and titles using the term 'diabet*' generated 495 records, as shown in Table 6.1. All records were reviewed. When assessed against the review inclusion criteria, 187 articles were retrieved and reviewed in full. Of these, 90 articles were included in the review.

Table 6.1 Number of articles identified by the literature review

Source	*Results of search*	*No. of articles considered eligible*	*Number of articles included in review*
PHI database: original search (up to June 2005) Total number = 12,562	326	105	48
Additional PHI database search (July-December 2005) Total number = 4021	169	9	-
Supplementary searching		73	42
TOTAL	495	187	**90**

Supplementary searches included scanning the reference lists of key articles, checking instrument websites, where found, and drawing on other bibliographic resources. All titles of issues of the following journals published between January and September 2006 were scanned:
- Diabetes Care
- Health and Quality of Life Outcomes
- Medical Care
- Quality of Life Research

Identification of patient-reported health instruments
Five generic and six diabetes-specific instruments were included in the review. Instruments targeting paediatric or adolescent populations were excluded, as were those focusing on particular complications of diabetes, and instruments where there was no evidence that an English-language version had been tested. We aimed to include only those which have been applied with both main types of diabetes. The reviews of specific instruments drew substantially on a previous review by the PHIG (Garratt et al., 2002). Developmental and evaluative studies relating to the instruments reviewed are listed in Tables 6.2 to 6.14. Table 6.15 gives an overview of newly developed instruments and those where only one study was identified.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Six generic instruments were identified which met the review criteria. Full details of the content and scoring methods are given in Chapter 3.

The measurement properties of the following instruments are reported:

a) SF-36
b) SF-12
c) Sickness Impact Scale
d) Health Utilities Index
e) Quality of Well-Being Scale
f) EuroQol- EQ-5D

a) SF-36

21 studies provided evidence of measurement properties for the SF-36.

Reliability

In a USA, outpatient clinic, internal consistency reliability scores for all subscales exceeded 0.70 (Jacobson et al., 1994). Similarly positive results were obtained in a UK study (Woodcock et al., 2001) and a second USA study (Wu et al., 1998).

Validity

In a large survey of the general population in eight countries, individuals self-reporting as having diabetes scored significantly worse on all dimensions of the SF-36 compared with respondents without chronic illness (Alonso et al., 2004). Similarly, in the Whitehall II study of over 10,000 civil servants, two scales of the SF-36, namely, General Health and Physical Functioning, showed significantly poorer scores for individuals with diabetes compared to the rest of the sample (Roberts et al., 1997).

In a study of patients with non-insulin-dependent diabetes, SF-36 scale scores were significantly related to number of complications but not to level of glycaemic control (Anderson et al., 1997). Significant correlations were also observed of SF-36 with severity and number of complications in a mixed study of type 1 and type 2 diabetes (Jacobson et al., 1994). In a national survey in the USA of individuals with diabetes, poorer scores for SF-36 were associated with a large number of other variables: less education, lower income, older age, being female, type of health insurance (no medical insurance or Medicare/Medicaid recipients reported lower quality of life than those with either a health maintenance organization or private insurance), number of diabetes complications, number of co morbid illnesses, and lower levels of physical activity (Glasgow et al., 1997). In a study of veterans patients with diabetes attending ambulatory clinics in Boston, SF-36 scales, especially Physical Function were significantly correlated with an index of diabetes severity (for example, eye and foot disease, atherosclerosis) Linzer et al., 2005). In a trial to improve management of patients with stable heart disease, those individuals who also had diabetes had significantly poorer SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) scores than those individuals free of diabetes (Deaton et al., 2006).

A large randomly selected sample of the population in Adelaide was interviewed and also clinically assessed to identify individuals with diabetes (Chittleborough et al., 2006). Individuals with diabetes, compared with the reset of the sample were found to have poorer scores on all scales of SF-36 except Mental Health. These differences also obtained for previously undiagnosed cases of diabetes. Moreover individuals with impaired fasting glucose also had poorer Physical Functioning and Bodily Pain scales. All differences were after controlling for age, sex and cardiovascular disease. The authors also conduct effect size analyses to show that SF-36 identifies mainly small but important differences for all forms of impaired compared with normal glucose.

In a UK study, diabetes-related illness and a greater burden of diabetes treatment were related to poorer SF-36 scores (Woodcock et al., 2001). The SF-36 was included in a questionnaire-based survey of patients who had been patients of a large hospital in Cardiff (Currie et al., 2006). Presence of peripheral neuropathy-related symptoms was related to poorer SF-36 subscale scores. A small-scale study of men with diabetes in Seattle found only modest, but significant correlations between an objective measure of step-count and the physical activity scale of the SF-36 (Smith D.G. et al., 2004). A study comparing older with younger individuals with diabetes did not find predicted differences by age for SF-36 scores (Trief et al., 2003). A trial to evaluate effects of intensive treatment for insulin-dependent diabetes found that using SF-36 scores, there was no adverse effect on quality of life of more intense management (Diabetes Control and Complications [DCCT] Group, 1996). In a study comparing individuals with diabetes, epilepsy and multiple sclerosis, individuals with diabetes had more favourable scores for Mental Health and Role limitations due to Emotional problems (Hermann, 1996).

Responsiveness

In a longitudinal study of veterans with predominantly type 2 diabetes, the majority of SF-36 scales showed significant deterioration over the three years of observation (Ahroni and Boyko, 2000). Moreover, those who were found to have developed diabetic complications experienced significantly more deterioration than other patients. In a four-year follow-up study of individuals with various chronic illnesses, those with diabetes had a significant reduction in the PCS score of SF-36 compared with a reference group of individuals with hypertension only at baseline (Bayliss et al., 2004). In a small two-year RCT of nurse care management, whereas some small but significant improvements were noted for HbA_{1c}, triglycerides, and diastolic blood pressure, only the Vitality scale of SF-36 showed significant improvements (Hill-Briggs et al., 2005). The authors interpret these results as evidence of lack of responsiveness of the majority of scales. In another small RCT of alternative insulin regimens for individuals with poorly controlled diabetes, one of the regimens was associated with significant changes over time in several SF-36 scale scores either at three- or six-month follow-up (Hendra and Taylor, 2004).

Precision

A study of individuals with type 1 diabetes found that compared with HUI3, the SF-36 produced more skewed data and also a distribution that was closer to healthy normative data (Supina et al., 2006). The authors argued that the similarity of scores to healthy individuals ran counter to clinical expectations.

Acceptability
A 'desirable' level of response to the SF-36 of 70% was obtained in a UK study of diabetes sent out by post from general practices (Woodcock et al., 2001). In a very small-scale study of SF-36 in a nurse-led community clinic, although a positive overall report, it was reported that individuals with visual problems had difficulties completing the instrument (Hartley, 2002).

Feasibility
No specific results were found.

b) SF-12

Reliability
No specific evidence was found.

Validity
In cross-sectional evidence, SF-12 some significant associations were found between patients' self-reported self-care and perceptions of quality of care, and SF-12 PCS and MCS scores (Aikens, Bingham, Piette, 2005). In a Canadian study, the SF-12 distinguished between individuals on different treatment regimes and groups with differing amounts of time off work, especially when scored by a Rasch-based method (Johnson and Maddigan, 2004).

Responsiveness
No specific evidence was found.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
A study of SF-12 in a mixed sample of chronically ill subjects from across American medical practices included patients with diabetes (Liu et al., 2005). It showed that 31% of responses had at least one missing value but used modelling to conclude that missing values can be reliably inferred.

Table 6.2 Studies of diabetes using SF-36 and the SF-12

Study	**Country (N)** **Age (years)** **Method of administration** **Setting**	**Measurement and Practical properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Ahroni and Boyko (2000)	USA (331) Type 2 diabetes 63 Interview and questionnaire Outpatients			✓			
Alonso et al. (2004)	Eight countries (24,936) 44 Mixed mail and interview General population, including individuals self-reporting diabetes		Construct ✓				
Anderson et al. (1997)	USA (255) 63 Mail Hospitals in Michigan area	Internal consistency ✓	Construct ✓				
Bayliss et al. (2004)	USA (1574) 58 Mail Various HMO settings			✓			
Chittleborough et al. (2006)	Adelaide, Australia 4006 (266 with diabetes) Age (diabetes) 62 Administered by interview		Construct ✓				
Currie et al. (2006)	Cardiff, UK (1298) Age: type 1 55, type 2 70. Mailed questionnaires, self-completed		Construct ✓				

Study	Country (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Deaton et al. (2006)	USA, UK (1013) Age: 62 Patients with stable heart disease, with and without diabetes Interview at baseline of a RCT		Construct ✓				
Diabetes Control and Complication Trial Group (1996)	Various centres in USA (1441) IDDM Self-completed in clinic		Construct ✓				
Glasgow et al. (1997)	National survey USA (2056) Age 59 Type 1 and type 2 Postal survey		Construct ✓				
Hartley L. (2002)	USA (31) Age: 60 Interview Community-based nurse-led clinic	Internal consistency ✓	Construct ✓				
Hendra and Taylor (2004)	UK (57) Age: 69 Administered by nurse in clinic Clinic-based RCT			✓			
Hermann et al. (1996)	USA (555) Age: 59 Mailed Range of health service settings		Construct ✓				

Study	Country (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hill-Briggs et al. (2005)	Baltimore USA (149) African-American type 2 diabetes RCT of nurse care management Administration not described			✓			
Jacobson et al. (1994)	USA (240) 44 (type 1) 60 (type 2) Type 1 and 2 diabetes Self-completed in clinic One outpatient clinic	Internal consistency ✓	Construct ✓				
Linzer et al. (2005)	Boston, USA (65) Age 64 Type 2 diabetes Mailed Primary care clinic		Construct ✓				
Roberts et al. (1997)	UK (10,308 of whom 65 have diabetes) Age 52 Self-completed questionnaire		Construct ✓				
Smith et al. (2004)	Seattle, USA (57) 68 Men with diabetes Clinic attendees		Construct ✓				
Supina et al. (2006)	Calgary Canada (216) 37 Type 1 diabetes Mailed to home Clinic		Construct ✓		✓		

Study	Country (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Trief P et al. (2003)	Syracuse, USA (191) Completed at clinic		Construct ✓				
Wu et al. (1998)	Wisconsin (143) 52 Type 1 diabetes Mailed from HMO	Internal consistency ✓					
Woodcock et al. (2001)	UK (184) Type 2 diabetes Mailed to home from general practice	Internal consistency ✓	Construct ✓			✓	
SF-12							
Aikens, Bingham, Piette (2005)	USA (752) Type 2 diabetes 63 Telephone interview Outpatients		Construct ✓				
Johnson and Maddigan (2004)	Alberta, Canada (372) Type 2 diabetes 62 Self-completed in clinic Outpatients		Construct ✓				
Liu et al. (2005)	USA (30,308) 53 Mixed chronically ill including diabetes Mailed to home						✓

c) Sickness Impact Profile

Reliability
A small-scale study of a sub-set of SIP subscales in individuals with diabetes found varying levels of test-retest reliability with lowest correlations for the Recreation and Pastimes scale (0.28) Bardsley et al., 1993).

Validity
A small-scale study of a sub-set of SIP subscales found satisfactory agreement with evidence from medical records of foot problems, angina and painful neuropathy, and body mass index (BMI) (Bardsley et al., 1993).

Responsiveness
No specific evidence was found.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 6.3 Study of diabetes using SIP

Study	**Country (N)** **Age** **Method of administration** **Setting**	**Measurement and Practical properties**					
SIP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bardsley et al. (1993)	UK (284) Interview Outpatient setting	Test re-test ✓	Construct ✓				

d) Health Utilities Index

Reliability
No specific evidence was found.

Validity
Maddigan and colleagues compared the performance of HUI2 and HUI3 as alternative scoring systems of a single 15-item questionnaire self-administered (Maddigan et al., 2003; Maddigan et al., 2004). While both scoring systems produced significant associations with clinical evidence such as type of treatment regimen and level of glycaemic control, differences were greater between clinical categories for HUI3 scoring. Maddigan and colleagues (2005) also examined patterns of HUI3 scores for individuals who self-reported diabetes in a Canadian national health survey

(Maddigan et al., 2005). Individuals self-diagnosing as having diabetes had somewhat lower HUI3 scores than healthy controls and further other co-morbidities increased the differences from health respondents. Similarly Bowker and colleagues (2006) found in another Canadian population health survey significantly poorer HUI3 scores for individuals with diabetes compared to healthy respondents, with cancer co-morbidity resulting in further reductions in HUI3 score. Wexler and colleagues (2006) found in a sample of type 2 diabetes patients that, with multiple regression analyses, microvascular complications, heart failure and depression were particularly strongly related to decreased HUI3 scores.

Responsiveness
No specific evidence was found.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 6.4 Studies of diabetes using HUI

Study/	**Country (N)** **Age (years)** **Method of administration** **Setting**	**Measurement and Practical properties**					
HUI		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bowker et al. (2006)	Canada (113,587) Self-completed population survey		Construct ✓				
Maddigan et al. (2003, 2004)	Alberta Canada (372) Age: 62 Type 2 Self-completed Trial of different services		Construct ✓				
Maddigan et al. (2005)	Canada (1193) Self-reported diabetes from survey		Construct ✓				
Wexler et al. (2006)	Boston USA (909) Type 2 diabetes Various types of clinic Supervised completion		Construct ✓				

e) Quality of Well-Being Scale

Reliability
In a small study within a clinical trial, Anderson and colleagues found high correlations between QWB scores, with assessments one day apart (Anderson et al., 1989).

Validity
Schwartz and colleagues administered the QWB in the context of a clinical trial evaluating glimepiride (Schwartz et al., 1999). They identified two distinct factors from items: observable limitations and subjective symptoms. In a study of type 1 and type 2 diabetes, QWB utility scores were significantly associated with frequency of hyperglycaemic symptoms and the occurrence of complications (Tabaei et al., 2004). In a study of type 1 and type 2 diabetes, major complications such as blindness, dialysis, symptomatic neuropathy, foot ulcers, amputation, debilitating stroke, and congestive heart failure were associated with lower utility scores in QWB (Coffey et al., 2002).

Responsiveness
No specific evidence was found.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 6.5 Studies of diabetes using Quality of Well-Being Scale

Study	**Country (N)** **Age (years)** **Method of administration** **Setting**	**Measurement and Practical properties**					
QWBS		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Anderson et al. (1989)	California (76) Non-insulin-dependent diabetes Clinical trial	Test-retest ✓					
Coffey et al. (2002)	Michigan (2048) Type 1 and type 2 diabetes Self-completed University-based clinic		Construct ✓				
Schwartz et al. (1999)	California (588) Age:59 Non-insulin-dependent diabetes Mixed: self-completed and interview Clinical trial		Internal ✓				
Tabaei et al. (2004)	Michigan (1522) Age (type 1) 33 Age (type 2) 56 Self-completed Attendees of clinic		Construct ✓				

f) EQ-5D

Reliability
No specific evidence was found.

Validity
A postal survey of patients in the UK attending one of four centres with diabetes registers included the EQ-5D (Holmes et al., 2000). As well as showing that individuals with type 2 diabetes had poorer EQ-5D scores than the general population, the study showed that complications of diabetes were consistently associated with poorer EQ-5D scores. Patients at a large hospital in Cardiff were sent a questionnaire six weeks after discharge or at an outpatient clinic; the study included 2575 patients with diabetes (Lee et al., 2005). The EQ-5D, included as part of the questionnaire, showed significant differences between type 1 and type 2 diabetes and significantly poorer utility scores with increased BMI.

Responsiveness
The EQ-5D was used as an outcome measure in the UK Prospective Diabetes Study RCT to evaluate benefits of tighter control of blood glucose level and blood pressure (UKPDS Group, 1999; Clarke et al., 2002). While tighter control had significant benefits in reducing complications, there were no differences between standard and more intensive treatment on the EQ-5D. However complications were significantly associated with poorer EQ-5D scores. The authors comment on the skewed nature of EQ-5D scores in analyses.

Precision
Bharmal and Thomas (2006) examined patterns of responses to EQ-5D and SF-6D in a general population survey, the 2000 Medical Expenditure Panel Survey, including respondents with diabetes. Up to 49% of those individuals with no problems identified on EQ-5D reported some negative items on SF-6D, leading the investigators to infer that EQ-5D had important ceiling effects.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 6.6 Studies of diabetes using EQ-5D

Study	**Country (N)** **Age (years)** **Method of administration** **Setting**	**Measurement and Practical properties**					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bharmal and Thomas (2006)	National health survey in USA (5104) with sub-sample with diabetes		Construct ✓		✓		
Holmes et al. (2000)	Four diabetes centres in UK (1578) Type 2 diabetes Age: 67 Self-completed from mailed questionnaire		Construct ✓				
Lee et al. (2005)	Cardiff (2575) Age: Type 1: 52, type 2: 68 Self-completed, mailed from hospital Discharged patients		Construct ✓				
UKPDS Group (1999)	UK (3104) Type 2 diabetes Age: 52 (self-completed in clinic)		Construct ✓				

Miscellaneous measures

Two generic measures, the Duke Health Profile (DUKE) and the General Health Perceptions Questionnaire (GHP) were both examined in a sample of 170 insulin-dependent patients with diabetes from a number of American clinics (Parkerson et al., 1993). Scales of the two instruments were treated as dependent variables in regression analyses. Aspects of diabetes (duration, complications, and severity of treatment) were not generally predictors of scales of either instrument, and stronger associations were found between socio-demographic and psychosocial factors and scales of the two instruments.

Hornquist and colleagues developed a system for rating the quality of life of individuals with diabetes, based on an initial study of 73 patients recruited in 1988 (Hornquist et al., 1993; Hornquist et al., 1995). However descriptions of the instrument are not clear and it is not apparent that the instrument has evolved into a stable form that can be considered for routine use. It is described as a generic instrument.

An Australian study was carried out of a modified version of the Patient Generated Index, termed 'the Client Generated Index' (CGI) (Griffiths et al., 2000). The CGI required trained interviewers to administer. High levels of test-retest reliability were found over a five-week period. Correlations with subscales of the SF-36 provided some evidence of construct validity.

The Behavioral Risk Factor Surveillance System (BRFSS) is a periodic national survey carried out by telephone interview in the United States. A study showed that nine questionnaire items from the BRFSS identified areas of quality of life in terms of which respondents with diabetes scored significantly worse than controls (Smith D.W. 2004).

RESULTS: DIABETES-SPECIFIC PATIENT-REPORTED HEALTH INSTRUMENTS:

Six specific instruments were identified which met the review criteria. Full details of the content and scoring methods are given in Tables 6.7 and 6.8.

The measurement properties of the following instruments are reported:

a) Appraisal of Diabetes Scale/ADS
b) Audit of Diabetes-Dependent Quality of Life/ADDQoL
c) Diabetes 39/D-39
d) Diabetes Health Profile/DHP
e) Diabetes Quality of Life Measure/DQOL
f) Diabetes Quality of Life Clinical Trial Questionnaire/DQLCTQ

a) Appraisal of Diabetes Scale/ADS

The ADS is a brief self-report questionnaire which assesses an individual's thoughts about coping with diabetes (Carey et al. 1991). It was developed in light of 'the transactional relationship between stress and diabetes' – the fact that whilst external stressors can disturb glucose metabolism, hence the course of the disease, adherence to a strict diabetic regime can of itself be stressful. The authors suggest that the ADS may be useful as a screening instrument for adjustment to diabetes, specifically to identify those patients experiencing, or at risk for, dysphoric reactions and problems of adherence to their diabetic regime. The content of the scale is based on theory and research regarding appraisal processes; some items were adapted from a generic Attribution Questionnaire (Hammen and Mayol, 1982).

The instrument consists of seven items covering distress caused by diabetes, control over diabetes (two items), uncertainty due to diabetes, anticipated future deterioration, coping, and effect of diabetes on life goals. The items use a five-point adjectival scale scored from 1 (e.g. control – none at all) to 5 (control – total amount). ADS items are summed to produce a score from 0-35, 0 representing the least and 35 the greatest impact of diabetes.

b) Audit of Diabetes Dependent Quality of Life/ADDQoL

The ADDQoL is an individualized instrument designed to measure an individual's perceptions of the impact of diabetes on their quality of life (Bradley et al., 1999; Speight & Bradley, 2000; Bradley & Speight, 2002). The intention was to create a detailed version for research and in-depth clinical work, and a short form for audit purposes. No further information has been found regarding the latter.

The instrument comprises 18 (originally 13) items where the respondent is invited to indicate, firstly, the effect of diabetes on a particular aspect of life (for example, enjoyment of food, ease of travelling) and, secondly, how important this aspect of life is to overall quality of life. Three (originally ten) of the items – namely, family life, working life, and sex life - have a 'not applicable' response option, allowing patients to exclude items which are not relevant to them. Patients respond by circling a number on a seven-point scale which asks how a particular aspect of their life would be if they did not have diabetes (from -3: 'very much better' to +3: 'very much worse'). They then rate the importance of this aspect of their life on a four-point scale (from 3: very important, to 0: not at all important). Impact ratings are multiplied by importance

ratings to produce a –9 to +9 score, then summed and divided by the number of applicable domains to produce a final score from –9 to +9.

In the original version of the ADDQoL, two additional summary items asked respondents to rate their general QoL, and what their QoL would be if they did not have diabetes, each on a seven-point verbal rating scale. The revised version has a single summary item measuring 'present quality of life' on a seven-point scale from -3 (extremely bad) to +3 (excellent). The wording has also been simplified and amended in order to reduce the number of 'non-applicable' items.

c) Diabetes 39/D-39

The authors of the D-39 intended it to have 'range and reliability', in other words, to be highly relevant to a wide range of diabetes patients over time, easy to use and understand, and to possess good psychometric properties (Boyer and Earp, 1997). A slightly modified version has been developed for use in clinical trials.

The D-39 comprises 39 items in five domains, namely energy and mobility (15 items), diabetes control (12 items), anxiety and worry (four items), social and peer burden (five items), and sexual functioning (three items). Scores are marked on seven-point visual analogue scales ranging from 'not affected at all' to 'extremely affected', then transformed linearly to 0 to 100 scales.

d) Diabetes Health Profile/DHP

The DHP-1 is a multidimensional self-completion instrument originally designed to identify psychosocial dysfunction among adult insulin-dependent and insulin-requiring patients in an ambulatory care setting (Meadows et al., 1996). The instrument has also been adapted for use in non-insulin dependent patients (Meadows et al., 2000). Content was derived from a literature review, a review of available instruments, interviews with IDDM and insulin-requiring patients, and discussions with diabetes health-care professionals (Meadows et al., 1996).

The DHP-1 comprises 32 items covering three dimensions: psychological distress (14 items), barriers to activity (13 items), and disinhibited eating (5 items); it is suggested this last may be appropriate as a screening tool for eating problems. Each item has a four-point adjectival scale; items are summed within the three dimensions and transformed to produce a score from 0-100 where 0 represents no dysfunction.

e) Diabetes Quality of Life Measure/DQOL

The DQoL was originally developed for use in a clinical trial comparing the efficacy of two different treatment regimens on the appearance and progression of chronic complications of patients with IDDM (DCCT Research Group 1988). However its structure allows for application to other patients with IDDM and NIDDM. The developers state that the DQoL could be used in clinical settings as a screening measure to identify patients with concerns about diabetes.

The instrument has 46 core items forming four scales: satisfaction with treatment (15 items), impact of treatment (20 items), worries about future effects of diabetes (four items), and worries about social and vocational issues (seven items). The instrument also includes a generic health item that does not contribute to the scales. Adolescent and youth versions of the DQoL have been developed (Ingersoll and Marrero, 1991).

The dimensions and DQoL total scores (average score across the four dimensions) are scored 0-100 where 0 represents the lowest possible quality of life and 100 the highest.

f) Diabetes Quality of Life Clinical Trials Questionnaire-Revised/DQLCTQ-R

The DQLCTQ was developed for use in multinational clinical trials of patients with IDDM and NIDDM (Shen et al., 1999). It was developed and published alongside a revised version of the instrument referred to as the DQLCTQ-R. The DQLCTQ comprises 142 items across 20 domains, three self-efficacy questions and four demographic questions. For the most part, items use five-point adjectival scales. The DQLCTQ-R, comprises 57 items across eight domains, with between three and ten response options. Mean scores for each domain are transformed into a 100-point scale where higher scores represent better quality of life.

Miscellaneous measures

Several diabetes-specific instruments with a very particular focus have been developed, for example, the Insulin Delivery System Rating Questionnaire (Peyrot and Rubin, 2005) described by the authors as a measure of HRQoL and treatment preference, and the Confidence in Diabetes Self-care scale (Van der Ven et al., 2003) which aims to assess a person's perceived ability to manage their condition. Culturally sensitive measures have also been developed for use with particular ethnic groups, given the higher prevalence of diabetes and greater incidence of long-term complications in, for example, African Americans as compared with 'European' Americans (Elasy et al., 2000). Although such specifically-focused instruments may have merit in targeted research studies, they are not included in this review.

More numerous still are diabetes-specific measures of psychosocial functioning. This is hardly surprising given that diabetes has been described as one of the most psychologically demanding of chronic illnesses (Cox and Gonder-Frederick, 1992). The best-known of these measures is the Well-Being Questionnaire/WBQ developed by Clare Bradley and colleagues (Bradley and Lewis, 1990) which measures depression, anxiety and, notably, positive well-being. Although not intrinsically specific to diabetes, this measure has been so widely used with diabetic patients; it is generally regarded as such. It is often used in conjunction with the Diabetes Treatment Satisfaction Questionnaire (Bradley and Lewis, 1990), developed concurrently with the WBQ.

Other such measures include the ATT39 (Dunn et al., 1986) which assesses emotional adjustment in diabetic patients, the Problem Areas in Diabetes scale/PAID (Polonsky et al., 1995) measuring diabetes-related distress, and the Fear of Hypoglycaemia Survey/HFS (Cox et al., 1987), developed in response to the phenomenon of insulin-dependent patients whose fear of hypoglycaemic episodes leads them to maintain undesirably elevated levels of blood glucose. Also noteworthy is the Diabetes Care Profile (Fitzgerald et al., 1996), a lengthy (234-item) survey instrument to assess the social and psychological factors related to diabetes and its treatment, which has been tested for reliability with a minority ethnic population (Fitzgerald et al., 1998). Again, this group of measures, whilst undoubtedly of importance, is beyond the scope of the present review.

DIABETES-SPECIFIC INSTRUMENTS: Domains, items and scoring methods

Table 6.7 Diabetes-specific patient-reported health instruments

Instrument (no. items)	***Domains (no. items)***	***Response options***	***Score***	***Administration/ Completion (time)***
Appraisal of Diabetes Scale/ADS	*Single index (7)* Distress Control (2 items) Uncertainty Future condition Coping Impact on life goals	5-point adjectival scales: 1 (not at all) to 5 (extremely/totally)	Scale scores summed to give an overall total 0-35	5 mins
Audit of Diabetes-Dependent Quality of Life/ADDQoL	*18 items:* Freedom to eat as I wish Enjoyment of food Family life* Working life* Sex life* Physical activity Worries about the future Holidays/leisure activities Freedom to drink as I wish Self-confidence Friendships, social life Motivation to achieve things Ease of travelling Physical appearance Finances Living conditions Unwanted dependence on others Reaction of society *1 summary item:* Present QoL	Impact: -3 (very much better without diabetes) to +3 (very much worse) Importance: 0 (not at all important) to 3 (very important) 3 items with N/A option (*)	Impact x importance = weighted score (range -9 to +9). Scores for each item summed, then divided by no. applicable items to give average weighted impact (AWI) score (i.e. N/A items do not contribute to score).	<10 mins
Diabetes 39/D-39	*39 items:* Anxiety and worry (4) Social and peer burden (5) Sexual functioning (3) Energy and mobility (15) Diabetes control (12)	7-point visual analogue scales; 1 = not affected at all, 7 = extremely affected	Scores transformed into 0-100 scores; 0 – lowest, 100 – highest possible score	*Not reported*

Instrument (no. items)	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Diabetes Health Profile/DHP 1/18	Psychological distress (14/6) Barriers to activity (13/7) Disinhibited eating (5/5)	Four-point adjectival scales	Item scores 0-3 in each dimension summed & transformed to produce score between 0 (no dysfunction) and 100	*Not reported – probably 5-10*
Diabetes Quality of Life Measure/ DQOL	Worries - future effects of diabetes (4) Worries - social/vocational issues (7) Impact of treatment (20) Satisfaction with treatment (15)	5-point Likert scale	*No details*	15- 20 minutes
Diabetes Quality of Life Clinical Trials Questionnaire-Revised/ DQLCTQ-R	*57 items in 8 domains:* Physical function Energy/fatigue Health distress Mental health Satisfaction (DQOL) Treatment satisfaction Treatment flexibility Frequency of symptoms *1 global health question* *1 transition question*	Variety of ordinal scales, with 3 to 10 response options.	Mean scores for each domain converted to a 100-point scale	'10 mins' – probably 15-20

Table 6.8 Summary of diabetes-specific instruments: health status domains

Instrument	*Instrument domains (after Fitzpatrick et al., 1998)*								
	Physical function	**Symptoms**	**Global judgment**	**Psychological well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal constructs**	**Treatment satisfaction**
Appraisal of Diabetes Scale/ADS				x	x		x	x	
Audit of Diabetes-Dependent Quality of Life/ADDQoL	x			x	x		x	x	
Diabetes 39/ D-39	x			x	x			x	
Diabetes Health Profile/DHP 1/18				x	x		x		
Diabetes Quality of Life Measure/ DQOL		x		x	x		x		x
Diabetes Quality of Life Clinical Trials Questionnaire-Revised/DQLCTQ-R	x	x	x	x	x		x	x	x

DIABETES- SPECIFIC PATIENT- REPORTED HEALTH INSTRUMENTS

a) Appraisal of Diabetes Scale/ADS

Reliability

Internal consistency

Item-total correlations were adequate: in the range of 0.28-0.59. Cronbach's alpha was 0.73, demonstrating sufficient reliability for use in groups. Principal component analysis yielded a single dimension; all items had loadings above 0.40. These analyses indicate the scale assess an internally consistent dimension of diabetes appraisal (Carey et al., 1991). Test-retest reliability was assessed by giving the ADS to a sub-sample of patients (n = 98) on three occasions: just before blood withdrawal, one hour after completing clinic visit, and one week later. Pearson product moment correlations for the one-hour and one-week retest were 0.89 and 0.85, respectively, demonstrating stability.

Validity

Convergent validity

Positive correlations were expected with measures of negative affect (anxiety, anger, and depression), perceived stress, diabetes-related hassles, perceived severity of diabetes/susceptibility to complications, and non-adherence to the diabetic regimen. A second sub-sample of patients (n = 102) was asked to complete these measures. Strong relationships were found between ADS scores and measures of negative affect, perceived stress, and diabetes-related hassles (Pearson product moment correlations 0.39-0.59). A modest relationship was found between the ADS and the measure of adherence (0.17), suggesting that patients reporting negative appraisal were less likely to adhere to their diabetic regimen.

Criterion validity

A low correlation (0.18) was found between ADS scores and the standard measure of glycaemic control, HbA_{1c} (glycosylated haemoglobin), indicating that those reporting negative appraisal were more likely to have experienced poor glycaemic control during the weeks prior to the test. Further studies have used the ADS as part of a battery of measures including the SF-36 and the DQOL, to examine the impact of family systems (Trief et al., 1998) and the work environment (Trief et al., 1999) on glycaemic control and psychosocial adaptation. Both studies found that the ADS strongly predicted glycaemic control and, in the 1998 study, scores on all DQOL subscales, and SF-36 role-physical, role-emotional, and bodily pain domains.

In the 1999 study, having more complications, older age, and shorter duration of diabetes were significant predictors of more negative appraisal on the ADS, whilst greater perceived supervisor support significantly predicted more positive appraisal. This study appeared to show inconsistent findings in that older age predicted more negative appraisal (ADS) but greater diabetes-related satisfaction (DQOL). However, it was concluded that the two measures tap into different aspects of adaptation: the DQOL satisfaction subscale assessing primarily current satisfaction with the diabetes care regimen, whilst the ADS focuses on coping efficacy in the present but uncertainty about the future.

Paradoxically, in a more recent study to compare the HRQoL of elderly and younger persons with diabetes (Trief et al., 2003), using the same array of instruments plus the PAID, the elderly group reported significantly better appraisal of diabetes (ADS) despite having more role limitations due to physical problems (SF-36). The authors hypothesize that this may reflect a cohort phenomenon, and acknowledge other potential confounders (e.g. specific complications, non-diabetic co-morbidities). Trief et al. also contend that diabetes-specific measures, including the ADS, have not included elderly patients in validation samples so that they may not be truly valid with the over-65 age-group; they suggest that future research should explore the validity of existing measures and pursue the development of diabetes-specific HRQoL measures specifically relevant to elderly individuals.

Responsiveness
No specific evidence was found.

Acceptability
The majority of patients found the instrument quick and easy to complete, requiring five minutes or less. In the initial study, of the 98 patients asked to complete the ADS on three occasions, 79% returned complete retest data.

Feasibility
No specific evidence was found.

Table 6.9 Developmental and evaluation studies relating to the Appraisal of Diabetes Scale/ADS

Study Reference Country	**Population & setting (n) Age; male/female Type 1/Type 2; duration Method of administration**	**Measurement properties**					
Appraisal of Diabetes Scale/ADS		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Carey et al. (1991) USA	Diabetic outpatients (200) Mean age 58.4; all M 66% T1; duration 15 yrs Self-report; 1-week retest mailed to home	Internal consistency ✓ Test-retest ✓	Convergent ✓			✓	✓
Trief et al. (1998) USA	Diabetes clinic patients (150) Mean age 51, range 20-79 M 84 (56%), F 66 (44%) T1 81 (54%) T2 68 (46%); duration 15.6 yrs. White 97% Self-report after clinic visit/at home, returned by mail. Battery incl. SF-36, DQOL, WBQ, ADS & family system measures		Criterion ✓				
Trief et al. (1999) USA	Diabetics in employment (129) Mean age 40.5, range 19-70 M 68 (53%), F 61 (47%) T1 93 (72%) T2 35 (27%); duration 14 yrs White 96% Self-report after clinic visit. Battery: DQOL, ADS & work systems measures		Criterion ✓				
Trief et al. (2003) USA	Diabetes clinic patients, all insulin users (191) a) 30-64 yrs (100); M 51, F 49 T1 52, T2 48; duration 13.5 yrs White 96% b) >/= 65 yrs (91); M 46, F 45 T1 18, T2 73; duration 18.3 yrs White 93% Self -report battery: SF-36, DQOL, PAID, ADS		Criterion ✓				

b) Audit of Diabetes Dependent Quality of Life/ADDQoL

Reliability

The original design of the ADDQoL was influenced by the principles underlying development of the interview measure, the Schedule for the Evaluation of Individual Quality of Life/SEIQoL (McGee et al., 1991), as well as discussions with health professionals, and in-depth interviews with diabetic patients. The content was then reviewed by the British Diabetic Association/Royal College of Physicians Working Group, and patients with diabetes. Further development of the instrument, with the addition of items aiming to extend its relevance to people with complications of diabetes, has resulted from work on the Renal-Dependent Quality of Life/RDQoL (Bradley, 1997).

Evidence for the unidimensionality of the 18-item instrument (Speight & Bradley, 2000) was found through a forced one-factor analysis; all 18 items had factor loadings above 0.5. Item-total correlations for the original 13-item version ranged 0.37 to 0.67 (item-total correlations for the 18-item version not found). Cronbach's standardized item alpha for the 18-item instrument was 0.92, indicating high reliability (Speight and Bradley, 2000). No evidence was found for test-test reliability.

In the original development study (Bradley et al., 1999), six of the ADDQoL items elicited responses which indicated positive effects of diabetes, illustrating the need for bipolar scales to measure the impact of diabetes. All four importance ratings were used for the 13 domains. The authors cite this as evidence in support of the importance ratings, which take individual perceptions of impact on QoL into account.

In the original study (Bradley et al., 1999), mean weighted ADDQoL scores were correlated with the two summary items and, as hypothesized, correlated better with the diabetes-specific item ($r = 0.47$) than with the generic item ($r = 0.31$); both were highly significant. The correlations fell well below 1.00, indicating that ADDQoL scores provided information additional to that elicited by the summary items.

Validity

Clinical and QoL variables

ADDQoL scores were significantly correlated with perceptions of hypoglycaemia ($r = 0.32$) and the number of reported complications ($r = 0.21$). As hypothesized, ADDQoL scores showed a greater negative impact of diabetes on quality of life for insulin-treated patients. This difference was significant for seven out of 13 dimensions (Bradley et al., 1999).

Generic heath status measures

A study by Woodcock et al., (2001) compared the performance of an 11-item version of the ADDQoL and the SF-36 in a group of patients with Type 2 diabetes, and concluded that the two were complementary. The authors found that ADDQoL scores were skewed towards good general QoL, although indicating a negative impact of diabetes. Correlations between the two instruments were greater amongst patients with diabetes alone, compared with patients reporting non-diabetic co-morbidity.

Responsiveness
A study reporting the DAFNE (dose adjustment for normal eating) trial of a five-day education programme, which aimed to teach patients how to match their insulin dose to food choices, found significant improvements in the negative impact of diabetes on dietary freedom, as measured by the 'Freedom to eat as I wish' item, and in the impact on general quality of life, measured by the summary item. For the former, the improvement was apparent at six months follow-up; for the latter it reached significance by one year (DAFNE Study Group, 2002).

Precision
The authors argue that the use of importance ratings to weight item scores prevents the impact of particular items from being either under- or overestimated in the individual case, enhancing precision (Bradley et al., 1999).

Acceptability
The principle behind the ADDQoL is to enable patients to show how diabetes affects them as an individual, allowing them to give added weight to those aspects which are particularly important to them. However, it has been argued (Polonsky, 2000) that the stem question of the scale ('If I did not have diabetes, [x aspect of my life] would be [from 'a great deal better' to 'a great deal worse']) represents a complex cognitive task, somewhat removed from direct questions about diabetes-specific QoL. On the other hand, it offers respondents the chance to indicate areas where they feel diabetes may have had a positive impact.

Response rates for the original samples ranged between 62% and 93%. Missing data for the three items presumed to relate to everyone, namely physical activity, motivation, and enjoyment of food, ranged from 3% to 8% (Bradley et al., 1999). In the Woodcock study (Woodcock et al., 2001), response rates exceeded 70%. It is estimated the instrument takes under ten minutes to complete.

A recent study in Singapore (Wee et al., 2006) found the ADDQoL to be culturally appropriate for English-speaking Chinese, Indian, and Malay patients, as well as confirming the reliability, validity, and acceptability of the instrument.

Feasibility
No specific evidence was found.

Table 6.10 Developmental and evaluation studies relating to the Audit of Diabetes-Dependent Quality of Life/ADDQoL

Study Reference Country	**Population & setting (n) Age; male/female Type 1/Type 2; duration Method of administration**	**Measurement properties**					
Audit of Diabetes-Dependent Quality of Life/ADDQoL		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bradley, Todd et al. (1999) UK	1) Cambridge – outpatients (52) Mean 52.4 yrs; 54% M, 46% F T1 & T2; duration 12.7 yrs Self-completed questionnaire 2) Bromley – patients attending education open day (102) 61.7 yrs; 54% M, 46% F IDDM & NIDDM; duration 7.3 yrs Self-completed questionnaire	Internal consistency ✓	Content ✓ Construct ✓		✓	✓	
Woodcock et al. (2001) UK	GP patients (184) [131/71% responded] Range: 30-70 yrs (most 55-64 yrs) T2 Mailed questionnaire		Construct ✓			✓	
Bradley & Speight (2002) UK	Outpatients attending annual review at hospital clinic (795) T1 & T2 Self-completed questionnaire	Internal consistency ✓	Content ✓ Construct ✓		✓	✓	
DAFNE study group (2002) UK	RCT participants (169) Mean 40 yrs; 44% M, 56% F Moderate-to-poorly controlled T1; duration 16.6 yrs Questionnaire battery (ADDQoL, DTSQ, WBQ) at baseline, 6 mths & 1 yr			✓		✓	
Wee, Tan et al. (2006) Singapore	English-speaking Chinese, Indian, Malay patients–tertiary acute care referral hospital (173) mean age: 52 yrs; range: 18-80 yrs 55% M, 45% F duration of diabetes: mean 10 yrs, range 0-62	Internal consistency ✓	Content ✓ Construct ✓		✓	✓	

c) Diabetes 39/D-39

Reliability

Despite the authors' intention of enabling patients to express their individual experience of diabetes and its impact on their lives, it is not clear whether patients were involved in item derivation for the D-39 (Boyer and Earp,1997).

Instrument development was in two phases. In the first, information derived from a literature review, existing quality of life instruments, and unstructured interviews with diabetes patients and health professionals (physicians, diabetes educators, pharmacists) was used to develop 93 items considered to address important aspects of patients' lives. Each item asked the respondent to assess the extent to which their quality of life was affected during the previous month by the action or activity within the item.

Following the application of factor analysis and item analysis, the instrument was reduced to 42 items in six domains. Item standard deviations were found to be approximately equal within each scale. With the exception of two items, larger correlations were found between items and scale scores than with the remaining scales. Item-total correlations were in the range 0.50-0.84.

In the second phase of the study, confirmatory factor analysis was used to confirm the presence of the six domains previously identified. Items were assessed for equivalent variances and item-total correlation. Item-total correlations were in the range 0.45-0.84. The instrument was reduced to 39 items and five domains.

The six domains from the first phase (Cary sample) produced Cronbach's alpha coefficients in the range 0.81-0.92. The final five-domain instrument produced alpha coefficients in the range 0.82-0.93 and 0.81-0.93 for the patients recruited from the community (Iowa) and from the hospital outpatient department (North Carolina), respectively. Estimates of internal consistency were all above the criterion of 0.70 for sub-groups of older patients and patients with no high school education.

Validity

In the first phase of instrument development, D-39 scores were correlated with global ratings of quality of life. There were no a priori hypotheses. Not all the results were significant but they were all in the anticipated direction. Four of the six dimension scores were significantly related to self-ratings of diabetes severity. Patients with seven or more concomitant conditions had the poorest scale scores (data not shown). Patients with no concomitant conditions had the best scores on five of the six dimensions (data not shown). Patients reporting depression as a concomitant condition had poorer scores on each of the six scales (data not shown).

Compared to younger patients, those aged over 75 had significantly poorer scores on the scales of energy and mobility. Younger patients had poorer scores, although not always significant, on the scales of diabetes control, anxiety and worry, social and peer burden, and diabetes medication (data not shown). Women had significantly poorer scores for the scales of energy and mobility, diabetes control, and anxiety and worry (data not shown). Patients who were not married had significantly poorer

scores for the scales of energy and mobility, and anxiety and worry, and significantly better scores on the sexual functioning scale.

In the second phase, the D-39 scores were compared with those for the eight scales of the SF-36. The instrument was assessed in both the community and the outpatient groups. As hypothesized, the largest correlations were found between the D-39 dimension of energy and mobility, and the SF-36 scale of physical functioning ($r = 0.71$), between the D-39 dimension of anxiety and worry, and the SF-36 scale of mental health ($r = 0.64$), and between the D-39 dimension of social burden and the SF-36 scale of social functioning ($r = 0.48$). Most correlations were statistically significant. All five dimensions of the D-39 had significant correlations with the self-reported global quality of life ($r = 0.21$-0.44) and self-reported diabetes severity ($r = 0.15$-0.56).

Relative to patients with no co-morbidity, patients with co morbid conditions had significantly poorer scores on the D-39 energy and mobility dimensions. Compared to younger patients, those aged over 75 had poorer scores on the scales of energy and mobility. Although not always statistically significant, younger patients had poorer scores on the scales of diabetes control, anxiety and worry, social and peer burden, and diabetes medication. With the exception of sexual functioning, in which men had significantly poorer scores, there were no significant score differences between men and women. Finally, compared to patients with no employment-limiting disabilities, those with employment-limiting disabilities had poorer scores across all five dimensions.

Compared to NIDDM patients, IDDM patients had significantly poorer scores for the D-39 dimensions of diabetes control, and anxiety and worry. These results had the greatest levels of statistical significance in the sample of patients recruited from the community. Patients using a combination of insulin and oral therapies had poorer scores across the five dimensions.

In a more limited study, involving low income respondents with diabetes in North Carolina, Camacho and colleagues (2002) found some additional evidence relevant to validation of D-39, with poorer subscale scores being associated with self-reported leg and foot complaints and a longer duration of diabetes.

Precision
The authors suggest that further research is needed to simplify the scoring system of the D-39 which may be unnecessarily precise.

Responsiveness
The instrument has not been assessed for responsiveness.

Acceptability
Of the 1000 questionnaires mailed to the Cary sample, 542 were returned (54.2%). There was a 73.3% response rate from the community pharmacy sample and a 45.8% response rate from the outpatient sample (see Table 6.11). Of the questionnaires returned, 70.8% and 41.4% were deemed usable from the community and outpatient samples, respectively. This suggests the questionnaire in its present form has poor acceptability.

Table 6.11 Developmental and evaluation studies relating to the Diabetes-39 instrument

Study Reference Country	**Population, setting (n) Age; male/female Type 1/Type 2; duration Method of administration**	**Measurement properties**					
Diabetes 39/D-39		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Boyer & Earp (1997) USA	Suburban Cary NC, pharmacist-run regional diabetes centre: 1000 selected from mailing list - 516 'usable' returned Age 52; M 240 (45.5%), F 288 (54.5%) T1 32.5%, T2 67.5%; duration 14 yrs White 88% Rural Iowa, GP patients (165) Age 62; M 74 (45%), F 90 (55%) T1 20%, T2 81%; duration 11.5 yrs White 100% Ethnically diverse Chapel Hill, NC (262) Age 55; M 93 (36%), F 169 (65%) T1 10%, T2 90%; duration 10 yrs White 42%, Black 54% Mailed questionnaire	Internal consistency ✓	Construct ✓ Criterion ✓		✓	✓	
Camacho et al. (2002) USA	Ethnically diverse (249) Age 54.4, range 18-87 M 77, F 172 T1 16% , T2 84%; mean duration 7.8, range 0-60 Insulin 32% yes, 68% no Mode of administration unclear		Construct ✓				

d) Diabetes Health Profile/DHP

Reliability

The content of the DHP was derived following a literature review, a review of available instruments, interviews with 25 IDDM and insulin-requiring patients and discussions with diabetes health-care professionals (Meadows et al., 1996). The interviews were analysed on the basis of thematic content which generated 95 items. Four assessors independently grouped the items into five areas. All four allocated 81% of the items to the same five areas; the remainder were allocated following discussion. No additional content was suggested but some items were re-worded.

Following a survey of patients, 24 items with poor levels of endorsement and low or high levels of intercorrelation were removed from the instrument. The structure of the instrument was assessed in three samples of patients using principal axis factoring (PAF). The first PAF analysis showed that there were two additional factors to those hypothesized. The 16 items loading onto these factors were removed, together with 12 items with low factor loadings. The level of correspondence between composition of the three resultant factors, and item grouping carried out by the assessors was found to be moderate but satisfactory.

After application of a forced three-factor PAF analysis on the remaining 43 items, a further 11 items were removed that had either low factor loadings or high loadings on more than one factor. The remaining 32 items contributed to three dimensions labelled psychological distress (PD; 14 items), barriers to activity (BA; 13 items) and disinhibited eating (DE; 5 items). Item-total correlations were in the range 0.47-0.75 and all items had higher item-total correlations within their own dimensions than with the other dimensions. The PAF results were confirmed across sexes and age-groups, and when the sample was randomly split in two to form two separate sub-samples. One final sample of patients confirmed the factor structure of the 32-item DHP in this evaluation.

Cronbach's alphas for two of the samples in which the instrument was developed were: PD (0.85-0.86), BA (0.82-0.85), and DE (0.77-0.80) (Meadows et al., 1996). Test-retest reliability has not been reported by the developers, although Whitty et al., (1997) tested the PD and BA dimensions, along with the other test items used, by administering the instrument twice, three weeks apart, to a sample of patients with NIDDM. 95% confidence intervals for the intraclass correlation coefficients were 0.90-0.96.

The DHP-1 has been adapted for use with Type 2 diabetics following studies with UK and Danish samples (Meadows et al., 2000). The instrument has the same three subscales comprising 18 items, irrelevant content (e.g. items relating to insulin therapy) having been removed. Cronbach's alpha for the modified scale (DHP-18) ranged 0.70-0.88, and all item-scale correlations exceeded 0.40.

Validity

Face and content

The authors state that the methods of item derivation and dimension development are evidence of satisfactory face and content validity for the DHP (Meadows et al., 1996). However, they acknowledge that a number of important areas, such as lack of social

support, fear/worry about late complications, and satisfaction with treatment and care providers, are absent (Meadows et al., 2000). It is suggested the DHP be used in combination with other disease-specific measures such as the DTSQ and PAID, as well as generic measures when appropriate, in order to obtain a full picture of the patient's level of functioning (Meadows et al., 2000).

Convergent
DHP scores were compared with those for the Hospital Anxiety and Depression Scale/HADS and the SF-36. Correlations were in the range 0.17-0.68 and all were statistically significant. As hypothesized, the highest correlations were found between the PD and BA dimensions and the HADS and the SF-36.

Discriminant
The authors also hypothesized that women would score higher than men on the PD and DE dimensions. These predictions were in part supported in one of the initial study samples, with women under 40 scoring significantly higher than men on the PD dimension, and women aged 65 years and under scoring significantly higher than men on the DE dimension. In another smaller sample, the PD dimension did not significantly differ between women and men, but women had a significantly higher mean score on the DE dimension. In the study to develop the DHP-18 (Meadows et al., 2000), it was hypothesized that insulin-treated patients would have higher PD and BA scores due to the increased demands of their treatment. This proved to be the case for the BA subscale where there were significant differences; however, for the PD subscale, it was true only in the UK sample.

Responsiveness
The DHP-1 has not been formally assessed for responsiveness. However, the PD and BA dimensions within the earlier version of the DHP have been assessed for responsiveness (Whitty et al., 1997). Following a literature review and discussions with clinicians, it was hypothesized that changing NIDDM patients to insulin treatment should result in improvements in psychological distress and energy. The PD and BA dimensions produced standardised response means (SRM) of 0.23 and 0.02 at six weeks follow-up, compared to an SRM of 0.85 for the Newcastle Diabetes Symptoms Questionnaire/NDSQ (McColl et al., 1995). Smaller SRMs were found at three months follow-up.

Precision
In the initial studies (Meadows et al., 1996), all three dimensions showed a positive skew (less dysfunctioning) and less than six percent of patients scored at the floor or ceiling on any dimension. Floor effects (a high percentage of patients scoring at the lowest level of dysfunction) were found for both language versions of the DHP-18 (Meadows et al., 2000), particularly in respect of the PD scale in the diet-treated group. The percentage of patients obtaining the maximum score (ceiling effect) on the three subscales was acceptable, however.

Acceptability
Two of the samples recruited for the development of the DHP-1 produced response rates of 79.0-86.0%. Anonymity meant that the response rate could not be calculated for one of the samples (Meadows et al., 1996). In the larger sample of 2239 patients, all 43 items were answered by 84.85% of the sample, with a significant association

between lower completion rate and increasing age. There was a response rate of 81.8% for the UK arm of the development study for the DHP-18

Feasibility
No specific evidence was found.

Table 6.12 Developmental and evaluation studies relating to the Diabetes Health Profile

Study Reference Country	**Population & setting (n) Age; male/female Type 1/Type 2; duration Method of administration**	**Measurement properties**					
Diabetes Health Profile/DHP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Meadows et al. (1996) UK	Insulin-dependent or insulin-requiring outpatients All ID/IR at one clinic (278) Mean age 41, range 20-65 Duration 13.7 yrs Mailed out questionnaire Outpatients 54 hospitals England & Wales (2239) Age 39.8, range 16-84 M 51%, F 49% Duration 13.1 yrs Questionnaire completed during waiting time, or returned by post 7 hospitals NE England (295) Age 51.5 range 19-90 M 52%, F 48% Mailed out questionnaire	Internal consistency ✓	Convergent ✓ Construct ✓		✓	✓	
Whitty et al. (1997) UK	Prospective follow-up of patients (48) commencing insulin at six diabetic clinics NE England Age: 54% <60, 46% >60 M 42%, F 58% T2; mean duration 7 yrs, range 1-17 yrs Self-complete items from SF-36, HADS, NDSQ, DHP, at baseline, 6 wks, 3 mths	Test-retest ✓		✓			
Meadows et al. (2000) UK	Consecutive patients (650), 175 insulin-treated, Age 57, M/F 50/50; 69 diet-treated Age 65; M 62%, F 38% 182 tablet-treated Age 64; M 59%, F 41% Mailed out questionnaire	Internal consistency ✓	Construct ✓		✓	✓	

e) Diabetes Quality-of-Life Measure/DQOL

Reliability

The development and initial validation of the DQoL was carried out by the Diabetes Control and Complications (DCCT) Research Group. The content of the DQoL was derived from the following three sources: a literature review identifying the concerns of diabetic patients and problems that impact on their lives, consultation with clinicians knowledgeable about diabetes, and patients with IDDM. The meaning, relevance and readability of the instrument were assessed during its development by giving draft versions to IDDM patients; drafts were also reviewed by health professionals. The initial item pool comprised items considered to be of greatest relevance to patients with IDDM undergoing treatments of differing intensity.

In the original reporting of the instrument, the DCCT Research Group (1988) reported Cronbach's alpha coefficients ranging from 0.69 to 0.92 for the scales of diabetes-related worry and total scores, respectively. Only the former dimension fell below 0.70. Parkerson et al., (1993) reported alpha values in the range 0.52-0.88 for the diabetes worry and total DQOL scores, respectively. Jacobson et al. (1994) reported alpha values in the range 0.47-0.87 for patients with IDDM and NIDDM. With the exception of the diabetes worry scale ($r = 0.47$-0.49), the reliability estimates were regarded as similar to those reported in previous studies.

Test-retest reliability was assessed by asking patients to complete a second questionnaire at a mean of nine days after it was first administered. Pearson correlations were in the range 0.78-0.92 for the social/vocational worry and total scores, respectively (DCCT Research Group, 1988).

Validity

The DQOL items were derived from IDDM patients and clinicians, together with the literature on psychosocial aspects of diabetes. Selected patients, as well as clinicians, then reviewed the items for content relevance. On the basis of patient input, the instrument was expanded to include worries about the future (DCCT Research Group, 1988). In this original evaluation, the DQOL was compared with three instruments: the Symptom Checklist-90-R/SCL, the Bradburn Affect Balance Scale/ABS, and the Psychosocial Adjustment to Illness Scale/PAIS. Several hypotheses were constructed. First, the DQOL worry scales would have larger correlations with the SCL total score than the PAIS and ABS. Second, the DQOL worry scales would have similar levels of correlation with the ABS and PAIS. Third, the DQOL satisfaction scale would have the largest correlation with the ABS. Fourth, the DQOL impact scale would have the largest correlation with the PAIS scales, with the exception of the PAIS distress scale. Finally, the DQOL total scores would have significant correlations with all instrument scores and the DQOL scales would have positive correlations with all instrument scores. Correlations were expected to fall within the range 0.3-0.7, indicating that constructs were similar but not identical.

The two DQOL worry scales were significantly correlated with the SCL total score ($r = 0.40$-0.50) and these were stronger than all correlations with the PAIS and ABS except for the PAIS scale of psychological distress ($r = 0.46$). The DQOL worry scales had similar low levels of correlation with the ABS and the PAIS except for the aforementioned psychological distress scale of the PAIS. The DQOL satisfaction

scale had a significant correlation (r = -0.55) with the ABS but had a slightly larger correlation with the PAIS scale of health-care orientation. The DQOL impact scale did produce the largest correlations with the PAIS scales although the PAIS psychological distress scale correlated more highly with the DQOL impact scale than expected. Finally, the DQOL total scores did have significant correlations with all the instrument scores and all correlations were positive.

The DCCT study found two small but significant associations with sex: women reported DQOL scores reflecting a greater impact of diabetes and greater diabetes-related worries. Two studies have compared the DQOL with generic instruments. The first compared the DQOL with the Duke-UNC Health Profile/DUHP, the General Health Perceptions Profile/GHP, and the Health and Daily Living Form/HDL (Parkerson et al., 1993). There were no formal hypotheses but the authors expected DQOL scores to explain greater variance in disease indicators than scores for the generic instruments. Of the disease indicators (duration of diabetes, complications and intensity of treatment), only the complications variable was a statistically significant predictor.

The DQOL total scores had 28% of variation explained by four co-morbidity and psychosocial variables. The DQOL social/vocational worry dimension had the most variance explained (41%) by these variables. The impact dimension had the least variance explained (12%). Similar analyses of a modified DQOL that separated the instrument into generic and disease-specific components found that more variance was explained by the generic component. Neither of the modified scales had a statistically significant relationship with the diabetes-related variables.

In a stepwise regression analysis, sex and age did not enter the equation when DQOL total scores and satisfaction, impact and diabetes worry scales were the dependent variables. However, age did enter the equation when social/vocational worry was the dependent variable. Age was predictive of less social worry. Marriage entered the equation when the two DQOL worry dimensions were dependent variables: being married was predictive of less worry and better mental health.

The second study compared the DQOL with the SF-36 scales of physical functioning, social functioning, role limitations due to physical problems, pain, and general health perception (Jacobson et al., 1994). The total DQOL had small to moderate levels of correlation with the SF-36 scales (r = 0.33-0.60). The DQOL scales of satisfaction and impact had the largest correlations with the SF-36 scales, ranging 0.28-0.50 and 0.30-0.59, respectively.

This study also assessed the relationship between the DQOL and complications using regression analysis, after adjusting for sociodemographic factors. The DQOL impact and satisfaction scales, and total scores had a significant relationship with the number of complications among patients with IDDM. The DQOL total scores, and impact, satisfaction and diabetes worry scales had a significant relationship with the severity of diabetes among patients with IDDM. The DQOL satisfaction scale had a significant relationship with the number of complications in patients with NIDDM. The DQOL impact and satisfaction scales, and total scores had a significant relationship with the severity of diabetes among patients with NIDDM. DQOL total scores were significantly correlated with age. Separated or divorced patients were

found to experience worse quality of life than their counterparts, but data were not presented.

Lloyd and colleagues (1992) followed up in adulthood a sample who had been first identified as IDDM in childhood. Patients with specific complications (macrovascular disease or nephropathy) had significantly poorer scores for DQOL and also there was deteriorating DQOL scores with each additional complication.

A study comparing younger with older patients with diabetes, found that older age and type 1 diabetes were independently associated with more favourable scores on DQOL (Trief et al., 2003).

In an application of the instrument to a sample of young adults with Type 1 diabetes in England, the instrument was found to have a different factor structure with three scales emerging, namely, social relationships, diabetes concerns, and impact (Eiser et al., 1992). Apart from a correlation with poor attendance at the study clinic, there were few significant correlations between scales of DQOL and clinical or disease-related variables.

Responsiveness

The responsiveness of the DQOL has not been formally assessed but the instrument's authors cite two studies as evidence for the responsiveness of the instruments. In the first, patients with end-stage renal disease were given either a kidney transplant or a combined pancreas/kidney transplant. There was a significant improvement in the DQOL total scores and all subscales in patients who received the combined transplant, while there was no improvement for those receiving the kidney transplant alone (Nathan et al., 1991). The second study compared the quality of life of patients who received an implantable pump with those receiving normal insulin treatments (Selam et al., 1992). The DQOL scale of satisfaction showed an improvement but there were no other changes. More recently, a small scale study appeared (Weinger and Jacobson, 2001) in which patients with diabetes attending a clinic intended to provide intensive treatment to improve control were assessed longitudinally on a number of measures. Patients showing improved glycaemic control also exhibited small but significant changes in scales of DQOL.

Precision

No specific evidence was found.

Acceptability

Information relating to the acceptability of the DQOL is available only for the two studies reporting comparisons with generic instruments. In the first, 131 out of 179 IDDM patients completed the DQOL and there were no missing items (Parkerson et al., 1993). The analysis was limited to those patients completing the DQOL. There were no significant differences between responders and non-responders to the DQOL for any of the demographic, psychosocial, or co morbidity variables collected. There were also no significant differences for disease duration and complications. However, a significant difference was found for intensity of treatment and 79.5% of non-responders were insulin-pump patients. The second study reported that 88% of patients agreed to participate (Jacobson et al., 1994). There were differences in responses to DQOL subscales, reflecting the fact that the social/vocational worry

subscale is less suitable for older NIDDM patients. The responses were different for the subscales of satisfaction (n = 228), diabetes worries (n = 219), impact (n = 217), and social/vocational worries (n = 61).

Feasibility

No specific evidence was found.

Table 6.13 Developmental and evaluation studies relating to the Diabetes Quality of Life Measure/DQOL

Study **Reference** **Country**	**Population (n)** **Age** **Method of administration** **Setting**	**Measurement properties**					
Diabetes Quality of Life Measure/DQOL		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
DCCT Research Group (1988) USA	Outpatients (192) Age 28 (adults) & 16 (adolescents) IDDM in adults and adolescents Administered during clinic visit	Internal consistency ✓ Reproducibility ✓	Construct ✓				
DCCT Research Group (1996) USA	Outpatients (1441) Age 27 IDDM Administered during clinic visit			✓			
Eiser et al. (1992) UK	Outpatients (69) Age: 21 Type 1 diabetes Self-completed in clinic		Construct ✓				
Jacobson et al. (1994) USA	Boston (240) Age: (Type 1) 44 Age (Type 2) 60 Self-completed during clinic visit	Internal consistency ✓	Construct ✓				
Lloyd et al. (1992) USA	Pittsburgh hospital register follow-up (175) Age: >28 Childhood IDDM Postal questionnaire		Construct ✓				
Nathan et al. (1991) USA	Recipients of transplant surgery at Boston hospital (33) Age: 34 IDDM Self-completed questionnaire			✓			

Study Reference Country	Population (n) Age Method of administration Setting	Measurement properties					
Diabetes Quality of Life Measure/DQOL		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Parkerson et al. (1993) USA	8 clinics (170) Age: 34 IDDM Self-completed during clinic visit	Internal consistency ✓	Construct ✓				
Selam et al. (1992) USA	Multi-centre trial (56) IDDM			✓			
Trief et al. (2003) USA	Specialist diabetes centre, Syracuse (191) Age: (younger group) 47 (older group) 71		Construct ✓				
Weinger and Jacobson (2001) USA	Specialist clinic in Boston (55) Age: 34 Completed during clinic visit			✓			

f) Diabetes Quality of Life Clinical Trial Questionnaire/DQLCTQ

Reliability

To develop the DQLCTQ, patient focus groups (30 patients) and expert clinician panels (11 clinicians) in the USA and France identified domains of importance. The literature was then reviewed to find generic and disease-specific instruments containing these domains. The major components of the draft instrument were based on these findings, and data were extracted from validated generic and disease-specific instruments. Generic instruments used were the SF-20 and SF-36; specific measures were the Diabetes Quality of Life Measure/DQOL (DCCT Group, 1988), the Questionnaire on Stress in Diabetes (Waadt et al., 1992) and the Hypoglycaemia Fear Survey/HFS (Cox et al., 1987). Instruments dealing with social stigma, treatment satisfaction, and symptoms were not available; these items were therefore developed. This process produced 293 items that were assessed for face and content validity by a group of researchers expert in the measurement of health-related quality of life.

The draft instrument was evaluated in patients attending five internal medicine practices and diabetes care centres in the USA. Following this study, and using the results of the focus groups and clinician panels, the instrument was reduced to make it acceptable for multinational clinical trials. Redundant items and domains were removed and domains with poor psychometric properties were modified or removed. Two domains were created for insulin-specific comparisons, treatment satisfaction and treatment flexibility.

Cronbach's alpha coefficients for the DQLCTQ domains were greater than 0.70, with the exception of the DQOL dimensions of social worry (0.62) and the DQOL diabetes worry (0.53) – i.e. these domains did not reach the levels of reliability required for group comparisons. The newly developed domains produced alpha coefficients ranging from 0.77 to 0.89 for the frequency of symptoms and treatment flexibility domains, respectively.

Test-retest reliability was assessed 7-10 days after baseline among the initial pilot sample of patients. Intraclass correlation coefficients were in the range 0.49-0.90 for the social stigma and health distress dimensions, respectively. The diabetes worry and the social stigma dimensions produced coefficients below 0.70. Revisions to the instrument meant that the test-retest reliability was not reported for the newly created domains of treatment satisfaction, treatment flexibility, frequency of symptoms, or bothersomeness of symptoms. The revised version of the instrument, the DQLCTQ-R, has good levels of reliability with alpha and test-retest coefficients all above 0.70 (Shen et al., 1999).

Validity

The draft instrument was assessed for face and content validity by a group of researchers with expertise in the measurement of health-related quality of life. The DQLCTQ was further assessed for validity through comparisons with clinical and sociodemographic variables. On the whole, the hypotheses were supported by the data. Patients with good metabolic control had significantly higher mean DQLCTQ scores than those with poor metabolic control. Patients who considered themselves to be in good control of their diabetes had significantly higher mean DQLCTQ scores than those who felt they were in poor control. With the exception of the domains of

social worry (DQOL), worry (HFS), treatment satisfaction, and treatment flexibility, patients with IDDM had higher mean DQLCTQ scores than those with NIDDM. With the exception of the dimensions of satisfaction, impact, and social worry and social stigma, women patients had poorer mean DQLCTQ scores than men.

Responsiveness

The responsiveness of the DQLCTQ was assessed by mean changes in DQLCTQ scores for patients whose metabolic control had improved or worsened over six months. For the improved group, the satisfaction (DQOL) and treatment satisfaction scales produced significantly better scores compared to baseline. For the worsened group, the mental health scale produced a significantly worse score compared to baseline. The DQLCTQ-R scales of treatment satisfaction, health/distress, mental health, and DQOL satisfaction produced significantly better scores compared to baseline for the improved group.

Acceptability

Less than 10% of items were missing for 83% of questionnaires administered (Kotsanos et al., 1997).

Feasibility

The authors report that the revised DQLCTQ can easily be administered and completed in ten minutes. From an examination of the questionnaire, this would appear somewhat optimistic.

Table 6.14 Developmental and evaluation studies relating to the Diabetes Quality of Life Clinical Trial Questionnaire/DQLCTQ

Study Reference Country	**Population & setting (n) Age; male/female Type 1/Type 2; duration Method of administration**	**Measurement properties**					
Diabetes Quality of Life Clinical Trial Questionnaire/DQLCTQ		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Kotsanos et al. (1997) Shen et al. (1999) Canada USA France Germany	Pilot study USA (123) Multinational randomised open-label crossover trial (T1 468, T2 474) T1 age 33.8; duration 12.6 M 56%, F 44% White 97% T2 age 58.2; duration 12.5 Questionnaire administered during clinic visit	Internal consistency ✓ Test-retest ✓	Face & content ✓ Construct ✓ Discriminant ✓ Criterion ✓	✓		✓	✓

Other diabetes-specific instruments identified from the review

Table 6.15 Overview of newly developed diabetes-specific instruments or single study reporting of measurement properties and/or evaluation.

Instrument Reference Country	Population, setting (n) Age; male/female Type 1/Type 2; duration Method of administration	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Summary; comments No other records identified unless stated
Diabetes Impact Measurement Scales/DIMS (Hammond and Aoki 1992) USA	Diabetes clinic patients (130) Mean: 45 yrs, range: 18-78 yrs; 42% M, 58% F Type 1 & 2, mean 11 yrs Self-administered questionnaire	Internal consistency ✓ Test-retest ✓	Construct ✓		✓	✓	✓	Designed to measure therapeutic impact in CTs Based on literature review, Rand instruments, SIP, AIMS, discussion with clinicians. 44 Likert-scale items in four subscales: symptoms (5 specific, 11 non-specific), 10 general well-being, 10 diabetes-related morale, 5 social role. Scores summed to produce 0-10 overall score. 15-20 mins completion time. No patient input. Correlated with VAS scales, clinical & demographic variables. Not all correlations >0.70. Some missing values. Test-retest not standard interval. No evidence for responsiveness.
Bournemouth Impact of Diabetes Scale (Everett, Kerr, 2005) UK	Diabetes clinic (237) Age: 37 Type 1 diabetes	Internal consistency ✓	Construct ✓					25 questions with responses on range 1-10. Is described as taking 5 minutes to complete. Good internal consistency and obtained significant differences between analogue and soluble forms of insulin. However no evidence of responsiveness. No other evidence for this recently reported instrument.
Three related measures - un-named (Brod et al., 2006) USA	Web-based survey of individuals with diabetes (418)		✓					Three scales developed from focus groups and web-based survey: satisfaction (21 items), symptoms (30 items), and productivity (14 items). Validation included distinguishing type of medication, age. No other studies and no evidence of responsiveness.

DISCUSSION AND RECOMMENDATIONS

Generic measures

Five generic measures were considered in detail because of the availability of potentially supportive evidence: SF-36 (and its variant SF-12), SIP, HUI, QWB, EQ-5D. By far the most substantial evidence was available in relation to SF-36. It has been widely used to capture the health status consequences of diabetes, in a wide range of settings and populations. It provides important evidence of the personal impact of the disorder. It has been extensively validated in the context of cross-sectional applications in diabetes. There is a small body of evidence also to support its use longitudinally to capture changes over time in health status and health-related quality of life. Most of this small body of evidence does suggest that the SF-36 is sensitive to change over time in important experiences for individuals with diabetes, although one small trial did fail to detect changes over time where other clinical evidence led investigators to expect change (Hill-Briggs et al., 2005). Responsiveness is always a critical requirement of patient-reported outcome measures. It becomes a particularly important issue in the context of diabetes where there is debate about whether interventions to achieve tight control of diabetes may have adverse effects on quality of life and such adverse effects need to be distinguished from consequences of the illness.

There is encouraging evidence for the use of instruments such as HUI and EQ-5D, instruments that may be important where assessment of utilities is needed. Even less evidence was found to address responsiveness of these types of instrument in diabetes than was found for non-utility generic instruments.

Diabetes-specific measures

In terms of volume of discussion, it is clear that patient-reported health instruments have an important role in improving understanding of diabetes and interventions for diabetes. However, given the clear importance of patients' experience and health-related quality of life in the condition, it is remarkable how few well-conducted studies were found independently to examine the measurement properties and practical usefulness of patient-reported health instruments. Even less common were studies directly comparing the performance of alternative instruments within samples of individuals with diabetes.

Three instruments have some evidence of measurement properties that might make them appropriate for further evaluation in the context of the NHS: ADDQOL, DHP and DQOL. They appear reasonably short for routine, regular use with adequate response rates and have some supportive evidence of measurement properties. Importantly, there is some evidence of responsiveness for each of the three instruments, although in no case was formal rigorous evidence of responsiveness found. Additional limitations include the limited coverage of the domains of the DHP, and evidence that the originally intended scales of DQOL may not be stable and may not pertain when completed by individuals with diabetes in the UK.

Recommendations

For assessment of broader aspects of health status in diabetes, the SF-36 clearly provides reliable insights; substantial evidence exists to support its use in diabetes and more from a wide range of other applications. Where, specifically, utility values are

required, there is also evidence to support use of EQ-5D and HUI. Normally it is recommended that a disease-specific measure is used in conjunction with a generic measure to assess particular problems of any given long-term condition. There is insufficient evidence strongly to single out any particular disease-specific instrument in diabetes. Of the large number of such instruments, ADDQOL, DHP and DQOL may warrant more attention to establish the case for a disease-specific instrument.

REFERENCES

Ahroni JH, Boyko EJ. Responsiveness of the SF-36 among veterans with diabetes mellitus. *Journal of Diabetes and its Complications* 2000;**14**:31-9.

Aikens JE, Bingham R, Piette JD. Patient-provider communication and self-care behavior among type 2 diabetes patients. *Diabetes Educator* 2005;**31**:681-90.

Alonso J, Ferrer M, Gandek B, Ware JEJr, Aaronson NK, Mosconi P *et al*. Health-related quality of life associated with chronic conditions in eight countries: results from the International Quality of Life Assessment (IQOLA) Project. *Quality of Life Research* 2004;**13**:283-98.

Anderson JP, Kaplan RM, Berry CC, Bush JW, Rumbaut RG. Interday reliability of function assessment for a health status measure: the Quality of Well-Being scale. *Medical Care* 1989;**27**:1076-84.

Anderson RM, Fitzgerald JT, Wisdom K, Davis WK, Hiss RG. A comparison of global versus disease-specific quality-of-life measures in patients with NIDDM. *Diabetes Care* 1997;**20**:299-305.

Bardsley MJ, Astell S, McCallum A, Home PD. The performance of three measures of health status in an outpatient diabetes population. *Diabetic Medicine* 1993;**10**:619-26.

Bayliss EA, Bayliss MS, Ware JEJr, Steiner JF. Predicting declines in physical function in persons with multiple chronic medical conditions: what we can learn from the medical problem list. *Health and Quality of Life Outcomes.* 2004;**2**:47.

Bharmal M, Thomas J. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value in Health* 2006;**9**:262.

Bowker SL, Pohar SL, Johnson JA. A cross-sectional study of health-related quality of life deficits in individuals with co-morbid diabetes and cancer. *Health and Quality of Life Outcomes* 2006;**4**.

Boyer JG, Earp JA. The development of an instrument for assessing the quality of life of people with diabetes: Diabetes-39. *Medical Care* 1997;**35**:440-53.

Bradley C. Design of a renal-dependent individualised quality of life questionnaire. *Advances in Peritoneal Dialysis* 1997;**13**:116-20.

Bradley C, Lewis KS. Measures of psychological well-being and treatment satisfaction developed from the responses of people with tablet-treated diabetes. *Diabetic Medicine* 1990;**7** :445-51.

Bradley C, Todd C, Gorton T, Symonds E, Martin A, Plowright R. The development of an individualized questionnaire measure of perceived impact of diabetes on quality of life: the ADDQoL. *Quality of Life Research* 1999;**8**:79-91.

Bradley C,.Speight J. Patient perceptions of diabetes and diabetes therapy: assessing quality of life. *Diabetes and Metabolism Research and Reviews* 2002;**18**:S64-S69.

Brod M, Skovlund SE, Wittrup-Jensen KU. Measuring the impact of diabetes through patient report of treatment satisfaction, productivity and symptom experience. *Quality of Life Research* 2006; **15**:481-91.

Camacho F, Anderson RT, Bell RA, Goff DCJ, Duren-Winfield V, Doss DD *et al*. Investigating correlates of health related quality of life in a low-income sample of patients with diabetes. *Quality of Life Research* 2002;**11**:783-96.

Carey MP, Jorgensen RS, Weinstock RS, Sprafkin RP, Lantinga LJ, Carnrike CLJ *et al*. Reliability and validity of the Appraisal of Diabetes Scale. *Journal of Behavioral Medicine* 1991;**14**:43-51.

Chittleborough C, Baldock K, Taylor A, Phillips P. Health status assessed by the SF-36 along the diabetes continuum in an Australian population. *Quality of Life Research* 2006;**15**:687-94.

Clarke P, Gray A, Holman R. Estimating utility values for health states of Type 2 diabetic patients using the EQ-5D (UKPDS 62). *Medical Decision-Making* 2002;**22**:340-9.

Coffey JT, Brandle M, Zhou Hea. Valuing health-related quality of life in diabetes. *Diabetes Care* 2002;**25**:2238-43.

Cox D, Irvine A, Gonder-Frederick L, Nowacek G, Butterfield J. Fear of hypoglycemia: quantification, validation, and utilization. *Diabetes Care* 1987;**10**:617-21.

Cox DJ, Gonder-Fredrick L. Major developments in behavioral diabetes research. *Journal of Consulting and Clinical Psychology* 1992;**60**:628-38.

Currie CJ, McEwan P, Peters JR, Patel TC, Dixon S. The routine collation of health outcomes data from hospital treated subjects in the Health Outcomes Data Repository (HODaR): descriptive analysis from the first 20,000 subjects. *Value in Health* 2006;**8**:581-90.

DAFNE Study Group. Training in flexible, intensive insulin management to enable dietary freedom in people with type 1 diabetes: dose adjustment for normal eating (DAFNE) randomised controlled trial. *BMJ: British Medical Journal* 2002;**325**:746.

Davis WK, Hess GE, Hiss RG. Psychosocial correlates of survival in diabetes. *Diabetes Care* 1988;**11** :538-45.

Deaton C, Kimble LP, Veledar E, Hartigan P, Boden WE, O'Rourke R *et al*. The synergistic effect of heart disease and diabetes on self-management, symptoms, and health status. *Heart and Lung* 2006;**35**:315-23.

Diabetes Control and Complications Trial Research Group. Reliability and validity of a diabetes quality-of-life measure for the Diabetes Control and Complications Trial (DCCT). *Diabetes Care* 1988;**11**:725-32.

Diabetes Control and Complications Trial Group. Influence of intensive diabetes treatment on quality-of-life outcomes in the Diabetes Control and Complications Trial. *Diabetes Care* 1996;**19**:195-203.

Dunn SM, Smartt HH, Beeney LJ, Turtle JR. Measurement of emotional adjustment in diabetic patients: validity and reliability of ATT39. *Diabetes Care* 1986;**9**:480-9.

Eiser C, Flynn M, Green E, Havermans T, Kirby R, Sandeman D *et al*. Quality of life in young adults with type 1 diabetes in relation to demographic and disease variables. *Diabetic Medicine* 1992;**9**:375-8.

Elasy TA, Samuel-Hodge CD, DeVellis RF, Skelly AH, Ammerman AS, Keyserling TC. Development of a health status measure for older African-American women with type 2 diabetes. *Diabetes Care* 2000;**23**:325-9.

Everett J, Kerr D. Measuring the impact of type 1 diabetes from the patient's perspective. *Journal of Diabetes Nursing* 2005;**9**:186-90.

Fitzgerald JT, Davis WK, Connell CM, Hess GE, Funnell MM, Hiss RG. Development and validation of the Diabetes Care Profile. *Evaluation and the Health Professions* 1996;**19**:208-30.

Fitzgerald JT, Anderson RM, Gruppen LD, Davis WK, Aman LC, Jacober SJ *et al*. The reliability of the Diabetes Care Profile for African Americans. *Evaluation and the Health Professions* 1998;**21**:52-65.

Garratt AM, Schmidt LJ, Fitzpatrick R. Patient-assessed health outcome measures for diabetes: a structured review. *Diabetic Medicine* 2002;**19**:1-11.

Glasgow RE, Ruggerio L, Eakin EG, Dryfoos J, Chobanian L. Quality of life and associated characteristics in a large national sample of adults with diabetes. *Diabetes Care* 1997;**20**:562-7.

Griffiths R, Jayasuriya R, Maitland H. Development of a client-generated health outcome measure for community nursing. *Australian and New Zealand Journal of Public Health* 2000;**24**:529-35.

Hammen C, Mayol A. Depression and cognitive characteristics of stressful life-event types. *Journal of Abnormal Psychology* 1982;**91**:165-74.

Hammond GS, Aoki TT. Measurement of health status in diabetic patients: Diabetes Impact Measurement Scales. *Diabetes Care* 1992;**15**:469-77.

Hartley LA. Functional health status of persons with diabetes in a nurse-managed clinic. *Diabetes Educator* 2002;**28**:106-14.

Hendra TJ, Taylor CD. A randomised trial of insulin on well-being and carer strain in elderly type 2 diabetic subjects. *Diabetes Complications* 2004;**18**:148-54.

Hermann BP, Vickrey BG, Hays RD, Cramer JA, Devinsky O, Meador K *et al*. A comparison of health-related quality of life in patients with epilepsy, diabetes and multiple sclerosis. *Epilepsy Research* 1996;**25**:113-8.

Hill-Briggs F, Gary TL, Baptiste-Roberts K, Brancati FL. Thirty-Six–Item Short-Form outcomes following a randomized controlled trial in Type 2 diabetes. *Diabetes Care* 2005;**28**:444.

Holmes J, McGill S, Kind P, Bottomley J, Gillam S, Murphy M. Health-related quality of life in Type 2 diabetes (T²ARDIS-2). *Value in Health* 2000;**3**:47.

Hornquist JO, Wikby A, Hansson B, Andersson PO. Quality of Life: status and change (QLsc) reliability, validity and sensitivity of a generic assessment approach tailored for diabetes. *Quality of Life Research* 1993;**2**:263-79.

Hornquist JO, Wikby A, Stenstrom U, Andersson PO. Change in quality of life along with type 1 diabetes. *Diabetes Research in Clinical Practice* 1995;**28**:63-72.

Ingersoll GM, Marrero DG. A modified quality-of-life measure for youths: psychometric properties. *The Diabetes Educator* 1991;**17**:114-8.

Jacobson AM, De Groot M, Samson JA. The evaluation of two measures of quality of life in patients with type I and type II diabetes. *Diabetes Care* 1994;**17**:267-74.

Johnson JA, Maddigan SL. Performance of the RAND-12 and SF-12 summary scores in type 2 diabetes. *Quality of Life Research* 2004;**13**:449-56.

Kotsanos JG, Vignati L, Huster W, Andrejasich C, Boggs MB, Jacobson AM *et al*. Health-related quality-of-life results from multinational clinical trials of insulin lispro. Assessing benefits of a new diabetes therapy. *Diabetes Care* 1997;**20**:948-58.

Lee AJ, Morgan CL, Morrissey M, Wittrup-Jensen KU, Kennedy-Martin T, Currie CJ. Evaluation of the association between the EQ-5D (health-related utility) and body mass index (obesity) in hospital-treated people with Type 1 diabetes, Type 2 diabetes and with no diagnosed diabetes. *Diabetic Medicine* 2005;**22**:1482-6.

Linzer M, Pierce C, Lincoln E, Miller DR, Payne SM, Clark JA *et al*. Preliminary validation of a patient-based self-assessment measure of severity of illness in type 2 diabetes: results from the pilot phase of the Veterans Health Study. *Journal of Ambulatory Care Management* 2005;**28**:167-76.

Liu H, Hays RD, Adams JL, Chen WP, Tisnado D, Mangione CM *et al*. Imputation of SF-12 health scores for respondents with partially missing data. *Health Services Research* 2005;**40**:905-21.

Lloyd C, Matthews KA, Wing RR, Orchard TJ. Psychosocial factors and complications of IDDM. The Pittsburgh Epidemiology of Diabetes Complications Study. VIII. *Diabetes Care* 1992;**15**:166-72.

Lustman PJ, Anderson RJ, Freedland KE, De Groot M, Carney RM, Clouse RE. Depression and poor glycemic control: a meta-analytic review of the literature. *Diabetes Care* 2000;**23**:934-42.

Maddigan SL, Feeny DH, Johnson JA. A comparison of the Health Utilities Indices Mark 2 and Mark 3 in type 2 diabetes. *Medical Decision-Making* 2003;**23**:489-501.

Maddigan SL, Feeny DH, Johnson JA. Construct validity of the RAND-12 and Health Utilities Index Mark 2 and 3 in type 2 diabetes. *Quality of Life Research* 2004;**13**:435-48.

Maddigan SL, Feeny DH, Johnson JA. Health-related quality of life deficits associated with diabetes and co-morbidities in a Canadian National Population Health Survey. *Quality of Life Research* 2005;**14**:1311-20.

McColl E, Steen N, Meadows KA, Hutchinson A, Eccles MP, Hewison J *et al*. Developing outcome measures for ambulatory care: an application to asthma and diabetes. *Social Science and Medicine* 1995;**41**:1339-48.

McGee HM, O'Boyle CA, Hickey AM, O'Malley KM, Joyce CRB. Assessing the quality of life of the individual: the SEIQoL with a healthy and a gastroenterology unit population. *Psychological Medicine* 1991;**21**:749-59.

Meadows KA, Steen N, McColl E, Eccles MP, Shiels C, Hewison J *et al*. The Diabetes Health Profile (DHP), a new instrument for assessing the psychosocial profile of insulin requiring patients: development and psychometric evaluation. *Quality of Life Research* 1996;**5**:242-54.

Meadows KA, Abrams C, Sandbaek A. Adaptation of the Diabetes Health Profile (DHP-1) for use with patients with Type 2 diabetes mellitus: psychometric evaluation and cross-cultural comparison. *Diabetic Medicine* 2000;**17**:572-80.

Nathan DM, Fogel H, Norman D, Russell PS, Tolkoff-Rubin N, Delmonico FL *et al*. Long-term metabolic and quality of life results with pancreatic/renal transplantation in insulin-dependent diabetes mellitus. *Transplantation* 1991;**52**:85-91.

Parkerson GRJ, Connis RT, Broadhead WE, Patrick DL, Taylor TR, Tse CK. Disease-specific versus generic measurement of health-related quality of life in insulin-dependent diabetic patients. *Medical Care* 1993;**31**:629-39.

Peyrot M, Rubin RR. Validity and reliability of an instrument for assessing health-related quality of life and treatment preferences: the insulin delivery system rating questionnaire. *Diabetes Care* 2005;**28**:53-8.

Polonsky WH, Anderson BJ, Lohrer PA, Welch G, Jacobson AM, Aponte JE *et al*. Assessment of diabetes-related distress. *Diabetes Care* 1995;**18**:754-60.

Polonsky WH. Understanding and assessing diabetes-specific quality of life. *Diabetes Spectrum* 2000;**13**:36.

Roberts R, Hemingway H, Marmot M. Psychometric and clinical validity of the SF-36 general health survey in the Whitehall II study. *British Journal of Health Psychology* 1997;**2**:285-300.

Rosenthal MJ, Fajardo M, Gilmore S, Morley JE, Naliboff BD. Hospitalization and mortality of diabetes in older adults: a three-year prospective study. *Diabetes Care* 1998;**21**:231-5.

Schwartz CE, Genderson MW, Kaplan RM, Anderson JP, Holbrook T. Covariation of physical and mental symptoms across illnesses: results of a factor analytic study. *Annals of Behavioral Medicine* 1999;**21**:122-7.

Selam JL, Micossi P, Dunn FL, Nathan DM. Clinical trial of programmable implantable insulin pump for type I diabetes. *Diabetes Care* 1992;**15**:877-85.

Shen W, Kotsanos JG, Huster WJ, Mathias SD, Andrejasich CM, Patrick DL. Development and validation of the Diabetes Quality of Life Clinical Trial Questionnaire. *Medical Care* 1999;**37**:AS45-AS66.

Skovlund SE. Patient-reported assessments in diabetes care: clinical and research applications. *Current Diabetes Reports* 2005;**5**:115-23.

Smith DG, Domholdt E, Coleman KL, del Aguila MA, Boone DA. Ambulatory activity in men with diabetes: relationship between self-reported and real-world performance-based measures. *Journal of Rehabilitation Research and Development* 2004;**41**:571-9.

Smith DW. The population perspective on quality of life among Americans with diabetes. *Quality of Life Research* 2004;**13**:1391-400.

Speight J, Bradley C. ADDQol indicates negative impact of diabetes on quality of life despite high levels of satisfaction with treatment. *Diabetologia* 2000;**43**:A225 (Abstract 864).

Supina AL, Feeny DF, Carroll LJ, Johnson JA. Misinterpretation with norm-based scoring of health status in adults with type 1 diabetes. *Health and Quality of Life Outcomes* 2006;**4** .

Tabaei BP, Shill-Novak J, Brandle M, Burke R, Kaplan RM, Herman WH. Glycemia and the quality of well-being in patients with diabetes. *Quality of Life Research* 2004;**13**:1153-61.

Trief PM, Grant W, Elbert K, Weinstock RS. Family environment, glycemic control, and the psychosocial adaptation of adults with diabetes. *Diabetes Care* 1998;**21**:241-5.

Trief PM, Aquilino C, Paradies K, Weinstock RS. Impact of the work environment on glycemic control and adaptation to diabetes. *Diabetes Care* 1999;**22**:569-74.

Trief PM, Wade MJ, Pine D, Weinstock RS. A comparison of health-related quality of life of elderly and younger insulin-treated adults with diabetes. *Age and Ageing* 2003;**32**:613-8.

UK Prospective Diabetes Study Group. Quality of life in type 2 diabetic patients is affected by complications but not by intensive policies to improve blood glucose or blood pressure control (UKPDS 37). *Diabetes Care* 1999;**22**:1125-36.

Van der Ven NCW, Weinger K, Yi J, Pouwer F, Ader H, Van der Ploeg HM *et al*. The Confidence in Diabetes Self-Care scale. Psychometric properties of a new measure of diabetes-specific self-efficacy in Dutch and US patients with type 1 diabetes. *Diabetes Care* 2003;**26**:713-8.

Waadt S, Duran G, Waadt M, Herschbach P. Quality of life in people with type 2 diabetes mellitus. In Lefèbvre PJ, Stabdl E, eds. *New aspects in diabetes*, Berlin, Germany: Walter de Gruyter, 1992.

Wee H-L, Tan C-E, Goh S-Y, Li S-C. Usefulness of the Audit of Diabetes-Dependent Quality-of-Life (ADDQoL) questionnaire in patients with diabetes in a multi-ethnic Asian country. *Pharmacoeconomics* 2006;**24**:673-82.

Weinger K, Jacobson AM. Psychosocial and quality of life correlates of glycemic control during intensive treatment of type 1 diabetes. *Patient Education and Counseling* 2001;**42**:123-31.

Wexler DJ, Grant RW, Wittenberg E, Bosch JL, Cagliero E, Delahanty L *et al*. Correlates of health-related quality of life in type 2 diabetes. *Diabetologia* 2006;**49**:1489-97.

Whitty R, Steen N, Eccles MP, McColl E, Hewison J, Meadows KA *et al*. A new self-completion outcome measure for diabetes: is it responsive to change? *Quality of Life Research* 1997;**6**:407-13.

Woodcock AJ, Julious SA, Kinmonth AL, Campbell MJ. Problems with the performance of the SF-36 among people with type 2 diabetes in general practice. *Quality of Life Research* 2001;**10**:661-70.

Wu SY, Sainfort F, Tomar RH, Tollios JL, Fryback DG, Klein R *et al*. Development and application of a model to estimate the impact of type 1 diabetes on health-related quality of life. *Diabetes Care* 1998;**21**:725-31.

Chapter 7: Patient-reported Health Instruments used for people with epilepsy

Epilepsy comprises a group of disorders in which there are recurrent episodes of altered cerebral function, the clinical accompaniment of which is a seizure. These vary in severity from brief lapses of awareness to prolonged unconsciousness, jerking of limbs and incontinence. Seizures are divided into generalised (arising from a wide area of the brain) or partial (arising from a limited area of damage to the brain). Treatment may be medical or surgical, and aims to control seizures. Surgery does not always prevent seizures from occurring.

About one person in 200 suffers from epilepsy. Many people lead normal lives with no symptoms between seizures. For others, epilepsy can have an adverse impact on everyday life, psychological well-being and feelings of stigma, and can have a slight adverse effect on mental ability. Even within the same group of seizures, differences in seizure frequency and severity can lead to differences in the impact on a person's life. It is an area where multidimensional social, psychological, physical and cognitive patient based outcome assessment is highly relevant.

Search terms and results: identification of articles

At the time of the review, the PHI database contained 12,562 records (up to June 2005). An initial search of record abstracts and titles using the terms 'epilep* *or* seizure*' generated 183 records, as shown in Table 7.1. All records were reviewed. When assessed against the review inclusion criteria, 106 articles were retrieved and reviewed in full. Of these, 71 articles were included in the review.

Table 7.1 Number of articles identified by the literature review

Source	***Results of search***	***No. of articles considered eligible***	***Number of articles included in review***
PHI database: original search (up to June 2005) **Total number = 12,562**	182	82	58
Additional PHI database search (July-December 2005) **Total number = 4021**	1	1	-
Hand searching		23	13
TOTAL	183	106	**71**

Supplementary searches included scanning the reference lists of key articles, checking instrument websites, where found, and drawing on other bibliographic resources. All titles of issues of the following journals published between January and September 2006 were scanned:

- Epilepsia
- Epilepsy Research
- Health and Quality of Life Outcomes
- Medical Care
- Quality of Life Research

Identification of patient-reported health instruments
Eight generic and eight epilepsy-specific instruments were included in the review. Instruments targeting paediatric or adolescent populations were excluded, as were those where there was no evidence that an English-language version had been tested. Developmental and evaluative studies relating to the instruments reviewed are listed in Tables 7.2 to 7.15. Table 7.16 provides an overview of records of newly developed epilepsy-specific instruments and single-study reporting of measurement properties and/or evaluation.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Seven generic instruments were identified which were evaluated with patients with epilepsy. For full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36 and SF-12
b) EQ-5D
c) HUI
d) Q-TWIST
e) NHP
f) SIP

a) SF-36 and SF-12

Nine published studies (two of which relate to the same study) described the use of the SF-36, as a measure of health status, or quality of life, with patients with diagnosed epilepsy. One study compared the SF-36 and the SF-12. The studies were based on outpatients, convenience samples or mixed groups. In two cases it was not clearly specified whether the patients were in- or outpatients.

Reliability
Jacoby et al. (1999), in their European study, reported that item-scale correlations for each subscale of the SF-36 all exceeded 0.40. Reliability coefficients exceeded the standard recommended for group comparisons. The lowest coefficient reported was for social function ($\alpha = 0.73$) and the highest was for bodily pain ($\alpha = 0.92$). Scaling success was reported to be high at 96% of comparisons made (using a definition of scaling success of any item/same scale correlation exceeding item/other scale correlation by 0.10 or more).

Wagner et al. (1996) reported similar scaling success with their sample of US out-patients, and their reported Cronbach's alpha coefficients similarly ranged from $\alpha =$ 0.73-0.93. Wagner et al. (1995) reported more variable scaling success with UK patients, with correlations varying from 31.3% to 100% (number of hypothesized correlations higher/total number of correlations). They reported modest to high internal consistency coefficients with UK patients (Cronbach's alpha) ($\alpha = 0.43$-0.92) and also high test-retest correlations ($r = 0.55$-0.88).

Validity
Five studies reported evidence of validity. Jacoby et al. (1999) provided the most explicit data for construct validity, reporting on associations between the SF-36 and seizure frequency and type, additional health problems and side-effects, in hypothesised directions. For example, the mean SF-36 scores for the Bodily Pain subscale were 82.7 (SE [standard error] 0.55) for those with no seizures in the past year, to 77.3 (0.77) for those with one per month, and 68.8 (0.70) for those with 1+ per month. Wagner et al. (1995) reported that the Role Physical scale discriminated best among patients' disease severity.

Epilepsy-specific measures of quality of life
SF-36 and SF-12
Two of the nine studies compared the results of the SF-36 and SF-12 with epilepsy-specific measures of quality of life and/or health utilities. Birbeck et al. (2000) compared the SF-36 and SF-12 with the Quality of Life in Epilepsy (QOLIE) shorter and long (31- and 89-item) versions. The QOLIE is an epilepsy-specific measure that includes the SF-36 as a generic core. They reported that the epilepsy-specific measure had larger responsiveness indices than the SF-36 or SF-12, although Wagner et al. (1995) reported stronger results for the generic measures. (See also Wiebe et al., 2002, below.)

Health utilities
Wiebe et al. (2002) compared the SF-36 with the QOLIE-31 and -89, and the Health Utilities Index version 3 (HUI-III). They reported all instruments to be robust, and able to distinguish accurately between different levels of patient-assessed changes in their condition.

Responsiveness
Birbeck et al. (2000) and Wiebe et al. (2002) provided some evidence of responsiveness to change (see earlier). They reported that the epilepsy-specific measure had larger responsiveness indices than the SF-36 or SF-12, although they were comparable in relation to mental and global health. However, Wiebe et al. (2002) reported all instruments to be robust below (see above).

Precision
Jacoby et al. (1999) reported negligible floor effects for all but the two role disability subscales, but substantial ceiling effects for five of the SF-36 subscales. Leidy et al., 1999a (see later in Table 7.7) reported that the SF-36 generic core embedded in the QOLIE-89 had the largest ceiling effects in the instrument: Role limitations-Emotional, Role limitations-Physical, Physical Function and Pain. Wagner et al. (1995) reported noteworthy ceiling effects for Role Functioning and Bodily Pain.

Acceptability
Jacoby et al. (1999) reported high item-completion.

Feasibility
Birbeck et al. (2000) compared scoring methods for the SF-36, using Rand's item response theory-based scoring versus equal weighting and scoring. They reported that the choice of method influenced scale results (overall, the Rand scoring method provided stronger results).

Table 7.2: Evaluative studies relating to the SF-36 and SF-12 *(both Rand MOS and Ware et al. versions)* when completed by patients with epilepsy

Study/ Country	**Population (N) Age (years) Method of administration Setting**	**Measurement and Practical properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Birbeck et al. (2000) USA	Participants in RCT medication for epilepsy (142) Age: range 18.8–66.8, mean 38.2 Mode of administration not specified In- or outpatients not specified		Construct ✓	✓			
*Buck et al. (1999) 8 European countries	Epilepsy (4929) Age: range 16-90, mean 37 Postal Outpatients and support groups		Construct ✓				
Hermann et al. (1996) USA	Epilepsy (271), multiple sclerosis/MS (85), diabetes (555) Age: mean 36.3, 44.6, 58.9, respectively Mode of administration not specified Diabetes patients from medical Outcomes Study, MS patients from neurology referrals, epilepsy centre patients; in- or outpatients not specified		Construct ✓				
*Jacoby et al. (1999) 8 European countries	Epilepsy (4929) Age: range 16-90, mean 37 Postal Outpatients and support groups	Item total ✓	Internal validity ✓ Construct ✓		✓	✓	
Leidy et al. (1999a) USA	Epilepsy (139), Age: 18+ , mean 38.5 Self-completed Convenience sample based on clinic records and outpatients		Construct ✓		✓		

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Wagner et al. (1995) UK	Epilepsy (136) Age; range 15-78, mean 34.9 Self-administered Outpatients	Internal consistency ✓ Item-total ✓ Test-retest ✓	Construct ✓		✓	✓	
Wagner et al. (1996) USA	Epilepsy (148) Age: 18+, mean 38.5 Self-administered Outpatients	Internal consistency ✓ Item-total ✓	Internal ✓				
Wagner et al. (1997) USA	Participants in RCT medication for epilepsy (163) Age: mean 43 (intervention group), 45 (control group) Self-administered Outpatients					✓	
Wiebe et al. (2002) Canada	Patients with difficult-to-control focal epilepsy investigated for surgery (136) Age: mean 36 Self-administered In- or outpatients not specified		Construct ✓	✓	✓		

***** These papers report on different findings from the same study

c) EuroQoL- EQ-5D

Three studies, all in the UK, used the EQ-5D. Two were studies of in- and outpatients (Remák et al., 2004; Selai et al., 2000), and one was based on a market research database of people with and without epilepsy (Trueman and Duthie 1998).

Reliability
No specific evidence was found.

Validity
Selai et al. (2000) reported that the measure was not valid in detecting changes pre- and post-surgical treatment for epilepsy. They also questioned the scales content validity (see Acceptability). Trueman and Duthie (1998) simply reported significant bivariate associations between the EQ-5D and the HADS (Hospital Anxiety and Depression Scale).

Responsiveness
Selai et al. (2000) reported that, in contrast to the Epilepsy Surgery Inventory-55 (ESI-55), there were no significant changes in the EQ-5D pre- and post-surgery. They concluded that the measure was unable to detect changes pre- and post-surgical treatment for epilepsy, and not valid or responsive. The EQ-5D visual analogue scale (VAS) was, however, responsive to clinically defined outcome. Remák et al. (2004) reported that the EQ-5D had mixed responsiveness to change in patient condition at six months after their commencement of one of five different epilepsy medications (the EQ-5D increased for only two of the medication groups). However, they stated that the lower power of their study might have been the cause.

Precision
No specific evidence was found.

Acceptability
Selai et al. (2000) reported that 42% of their sample queried the EQ-5D VAS, mainly because ‘health does not include epilepsy’ and if it did, the score would be up to 70 points lower.

Feasibility
No specific evidence was found.

Table 7.3: Developmental and evaluation studies relating to the EQ-5D

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
EQ-5D		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Remák et al. (2004) UK	Patients with intractable epilepsy on five different medical therapies (125) Age: mean 35.7 to 38 Outpatients Interviews			✓			
Selai et al. (2000) UK	Epilepsy patients (145, 45 followed up) Age: not given Interview Inpatients		Construct ✓	✓		✓	
Trueman & Duthie (1998) UK	Market research database of people of with (289) and without 9389) epilepsy Age: mean 46 Mode of administration: Self-administration		Construct ✓ Concurrent ✓			✓	

d) Health Utilities Index

Wiebe et al. (2001; 2002) examined minimum clinically important change; small, medium, and large changes; and changes needed to exclude chance error in the Health Utilities version III, along with the SF-36, and the Quality of Life in Epilepsy Inventory 31- and 89-item versions (QOLIE-31, QOLIE-89). They reported (2002) that the HUI-III, and the other instruments, all differentiated between no change and minimum important change. Only the two QOLIE instruments distinguished accurately between minimum important change and medium or large change. In 2001, Wiebe et al. reported that threshold values for the HUI-III were larger than expected, due to large between-patient variance, which they attributed to the nature of the instrument. Langfitt and Wiebe (2002) reviewed methodological issues in determining health values in epilepsy.

Table 7.4: Developmental and evaluation studies relating to the Health Utilities Index, version 3

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
HUI-III		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Wiebe et al. (2001) Canada	Stable epilepsy patients, candidates for surgery (40) Age: mean 36 In- or outpatients not specified Mode of adminstration unclear			✓	✓		
Wiebe et al. (2002) Canada	Patients with difficult-to-control focal epilepsy investigated for surgery (136) Age: mean 36 Self-administered In- or outpatients not specified		Construct ✓	✓	✓		

e) Q-TWIST

Schwartz et al. (1995) used the approach of 'quality-adjusted time without symptoms and toxicities' (Q-TWIST) as a hypothetical example. The paper is methodological, and not empirical, and explains their adapted Q-TWIST approach, which includes additional dimensions relevant to epilepsy.

f) The Nottingham Health Profile

The Nottingham Health Profile was used in two trials of medical therapy for epilepsy (Chadwick 1994; Smith et al., 1993) and a study of a patient population (Baker et al., 1993).Chadwick presented no data for the NHP, simply reporting that it gave 'poor information' and 'lacked sensitivity'. Smith et al. (1993) reported there were no differences between control and placebo groups with the NHP subscales post-treatment, despite a significant reduction in seizure frequency among the treatment group. Baker et al. (1993) reported high internal consistency for the NHP, but only the Physical mobility subscale was able significantly to distinguish between patients taking medication or a placebo.

Table 7.5: Developmental and evaluation studies relating to the Nottingham Health Profile

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
NHP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Chadwick (1994) UK	Epilepsy patients in drug trial (81) Age: range 15-67, mean 33.7 LSSS only In- or outpatients not specified Mode of administration not specified		✓				
Smith et al. (1993) UK	Patients with medically refractory partial seizures (100) Age: range 15-67, mean 32.7 In- or outpatients not specified Self-administration		✓				
Baker et al. (1993) UK	Patients with refractory epilepsy (81) Age: range 15-67, mean 33.7 Self-administration	Internal consistency ✓	✓	✓			

g) Sickness Impact Profile

Langfitt (1995) compared the Sickness Impact Profile with the Epilepsy Surgery Inventory-55 and the Washington Psychosocial Seizure Inventory. All measures were judged to be valid for use with epilepsy patients, and the SIP was preferred in studies of the broad impact of epilepsy on quality of life. All summary scales and most scales exceeded the Cronbach's alpha 0.70 criterion suggested for group comparisons (reliability). Scales with low internal consistency were examined and items generally covaried according to item content. Construct validity was supported by correlations between comparable subscales. Criterion validity was supported by correlations between the scales and disease severity (with the exception of the WPSI family background and interpersonal adjustment subscales). Feasibility analyses showed that the SIP took an average of 34.5 minutes to complete. (WPSI took an average of 15.8 mins to complete; ESI-55 16 mins.

RESULTS: EPILEPSY-SPECIFIC PATIENT-REPORTED HEALTH INSTRUMENTS:

Eight epilepsy-specific instruments were identified which were evaluated with patients with epilepsy. For full details of the development, domains and scoring methods are detailed in Tables 7.6 and 7.7.

The following instruments measurement properties are reported:

a) Epilepsy Surgery Inventory-55
b) Katz Adjustment Scale
c) Liverpool Quality of Life (LQOL) Battery and Seizure Severity Scale
d) Quality of Life in Epilepsy-89
e) Quality of Life in Epilepsy-31
f) Quality of Life in Epilepsy-10
g) Side-Effect and Life Satisfaction (SEALS) Inventory
h) Washington Psychosocial Seizure Inventory

a). Epilepsy Surgery Inventory-55 (ESI-55)

The ESI-55 is a 55-item measure of health-related quality of life, designed to assess outcome of epilepsy surgery. It was constructed after a literature review, and includes the Rand SF-36 as a generic core, plus 19 epilepsy-specific items (Vickrey et al., 1992a). The ESI-55 contains 11 multi-item subscales of health perceptions, energy/fatigue, overall QoL, social functioning, emotional well-being, cognitive functioning, and role limitations due to emotional problems, role limitations due to memory problems, role limitations due to physical health problems, physical functioning and pain. The initial scale was tested on a small sample of epilepsy patients and then reviewed by health-care professionals, before administration to a sample of epilepsy patients to evaluate its reliability and validity (Vickrey et al., 1992a).

Subgroups of these scales can be weighted and summed to form scores for mental health, physical health and role functioning. The scale scores are weighted and summed to produce the overall score. The health perceptions subscale has been reported to have the greatest sensitivity in discriminating between patients varying by seizure type and frequency (Vickrey et al., 1995). The ESI-55 takes an average of 15 minutes to complete. It was reviewed by Devinsky and Vickrey (1994), Selai and Trimble (1995), Jacoby (1996), Leidy et al., (1998), and Buelow and Ferrans (2001).

b) Katz Adjustment Scale

This instrument was originally developed to measure social behaviour and adjustment of people with a diagnosis of schizophrenia (Katz and Lyerly, 1963), but has been extended and adapted for use with people with epilepsy (Vickrey et al., 1992b; see also summary by Trimble, 1994). However the scale is completed by relative/friend proxies and not the patients themselves. Vickrey et al. (1992b) increased the items and supported the scale's validity for use with epilepsy patients. Their tested version contains 126 items (127 items should have been included but one was omitted in error), in 14 rather than 12 subscales. Proxies are asked to rate the patient according to

'how your relative or friend has looked to you during the past few weeks on these things'. For each item there are four response choices ('almost never' to 'almost always'). The revised instrument by Vickrey et al. (1992b) includes 14 subscales: Over-sensitivity/fearfulness, Social, Irritability, Dependency, Acting out, Paranoia, and Abnormal thought process, Withdrawal-R, Emotional liability, Nervousness-R, Sociopathy, Bizarreness-R, Hyperactivity-R, Disorientation. The responses are summed and transformed to a 0-100 point scale. Higher values indicate better functioning. The scale has been reviewed by Hermann (1995).

c) Liverpool Quality of Life (LQOL) Battery and Seizure Severity Scale (LSSS)

The aim of the LQOL was to focus on issues relevant and important to people with epilepsy. The LQOL was initially tested using a wide range of existing and new scales. The final version consists of two epilepsy-specific subscales: adverse drug effects (21 adverse drug effects, rated on a four-point Likert scale; total sum scores are used in analyses) and impact of epilepsy (nine areas of life that can be affected by epilepsy or treatment, rated on four-point Likert scales; mean scores are used in analyses). It also includes three general subscales. The first scale is Affect Balance (an existing, well tested scale encompassing five items describing negative and five items describing positive states, rated dichotomously as present or absent. The second scale is sense of mastery (an existing, well tested, 7-item scale, rated on 4-point Likert scales, summed with higher scores representing higher mastery. Thirdly there is the life fulfilment scale (ten items on areas of life, rated for importance on 4-point rating scales) and then again for satisfaction, on 4-point Likert scales. The scale together with the LSSS takes up to 45 minutes to complete. It was reviewed by Hermann (1995), Leidy et al. (1998), and Buelow and Ferrans (2001).

The LSSS contains 20 clinical features or symptoms of seizures over the previous four weeks, rated on 4-point Likert scales (total scores range from 20-80). Two subscales measure perceived control over seizures and ictal and post-ictal (11 items) symptoms. Scores are computed by summing item scores, with scores ranging from 9-36 for perceptions, and 11-44 for the ictal scale. Higher scores indicate worse severity.

d) Quality of Life in Epilepsy-89 (QOLIE-89)

The QOLIE-89 is an epilepsy-specific measure that includes the seven subscales of the Rand SF-36 as a generic core. It is an extension of a 55-item, QoL questionnaire (the Epilepsy Surgery Inventory, ESI-55) which was designed for use with epilepsy surgery patients. Items judged by the investigators to be missing were included in the QOLIE. It was developed with 304 epilepsy patients and their relative/friend proxies from 25 epilepsy centers in the USA (130 men and 174 women), with a mean age of 36 years (range 17-63). It was repeated two to three weeks later (Devinsky et al., 1995). The questionnaire takes an average of 28.4 minutes (SD [standard deviation] 15.6, range 6-135) to complete.

It contains 17 multi-item subscales comprising 86 items plus three single item measures of change in health, sexual relations, overall health (the SF-36 core is supplemented by 53 items specific to epilepsy) grouped into four factors. It aims to assess physical, mental, and social areas of life. Standardised methods are used to convert each item to a 0-100 score, with higher scores indicating better QoL. Subscale

scores involve averaging across the items in the subscale, with the number of items as the division. The overall score is a weighted sum of the individual subscale scores. Factor-based, standardised regression coefficients (weights) are used to calculate domain scores.

The instrument initially included 99 items at administration, 86 of which, across 17 subscales, were retained after multitrait scaling. Factor analysis of the 17 subscales yielded four underlying dimensions of health: an epilepsy-targeted dimension, cognitive, mental health, and physical health. Construct validity was supported by significant patient-proxy correlations, and correlations between the instrument and seizure frequency over the previous year, neuropsychological tests, and emotional and cognitive function. It was reviewed by Devinsky and Vickrey (1994), Jacoby (1996), Leidy et al. (1998), Leppik (1998), and Buelow and Ferrans (2001).

e) Quality of Life in Epilepsy-31 (QOLIE-31)

The QOLIE-31 was developed, using the original dataset, from the 99 items used to develop the QOLIE-89 by Cramer et al. (1998). They selected the subscales that were reported to be most important by people with epilepsy (as determined by an expert panel), with the result that generic topics (e.g. pain) were excluded.

Following psychometric and factor analyses of the full scale, variables with loadings of equal to or greater than 0.4 were included in the subscales for the QOLIE-31. This resulted in a 31-item questionnaire, with seven subscales, forming two factors: Emotional/Psychological Effects (seizure worry, overall QoL, emotional well-being, energy/fatigue) and Medical/Social Effects (medication effects, work-driving-social limits, cognitive function). Cross-cultural translations were developed. Analyses supported the reliability and validity of the QOLIE-31. It was reviewed by Leidy et al. (1998), and more briefly by Jacoby (1996) and Leppik (1998).

f) Quality of Life in Epilepsy-10 (QOLIE-10)

This was also developed, using the original dataset, from the 99 items used to develop the QOLIE-89 (Cramer et al., 1996). Items from the QOLIE-89 were selected for inclusion in the QOLIE-10 by an expert panel, which also identified seven domains considered to be important for epilepsy patients. The panel selected items with high item-scale correlations, consistent or appropriate wording and sentence structure. The 10-item questionnaires covers general and epilepsy-specific areas, grouped into three factors: Epilepsy effects (memory, physical effects, mental effects of medication), Mental health (energy, depression, overall QoL), Role functioning (seizure worry, work, driving, social limits).There is some support for its reliability and validity.

g) Side-Effect and Life Satisfaction (SEALS) Inventory

The SEALS is a 50-item self-completion questionnaire designed to measure satisfaction with anti-epileptic drug (AED) therapy. The original also contains an ADL (activities of daily living) subscale with items on frequency of daily activities, from household to social roles. A less diffuse, 38-item version is available (Gillham et al., 2000). The items relate to the patients' feelings and behaviour over the previous week, and were grouped into five subscales, supported by factor analysis: General

cognitive difficulties, Satisfaction/Dysphoria. Fatigue/Tiredness, Temper, Worry, each with 4-point Likert frequency response scales. Answers are summed for each domain and for an overall score. The SEALS was designed by Brown and Tomlinson (1982) with 125 epilepsy patients and 79 people without epilepsy. Fatigue discriminated well between patients and non-patients.

h) Washington Psychosocial Seizure Inventory (WPSI)

The WPSI is the oldest instrument for the evaluation of psychosocial concerns in adults with epilepsy. It was not intended to cover broader health or QoL. The questions are anchored in actual performance in life, and assess adaptation and functioning. It has good reliability and validity when compared to clinical ratings (Dodrill et al., 1980).

The instrument was developed with a group of behavioural scientists, who compiled a list of categories of psychosocial problems they had encountered in this area. These were: Family background, Emotional adjustment, Interpersonal problems, Vocational adjustment, Financial status, Adjustment to seizures, Medicine and medical management, and overall psycho-social functioning. Item development and scaling was constructed next. After two piloting studies, 132 items with Yes/No responses resulted. Professionals then rated 127 adults. Each subscale had to satisfy empirical requirements for inclusion. Inter-rater reliability, retest and internal consistency reliability were established, although patients were not consulted. The resulting scale has 132 items in three validity subscales and the eight clinical subscales (Family background, Emotional adjustment, Interpersonal problems, Vocational adjustment, Financial status, Adjustment to seizures, Medicine and medical management, and overall psychosocial functioning). Later an item measuring QoL was added (Dodrill and Batzel, 1996). There are four profiles: 1) No problems; 2) Possible or slight difficulties; 3) Definite problems; 4) Severe or major problems. Higher scores indicate poorer adjustment. It is a lengthy instrument, taking 15-20 minutes to complete, using a trained interviewer.

The instrument has been reviewed by Hermann (1995), Jacoby (1996) and Selai and Trimble (1995). An overview of its development and widespread use was published by Dodrill and Batzel (1994), who reported 48 published papers on the WPSI.

EPILEPSY-SPECIFIC INSTRUMENTS: RESULTS

Table 7.6: Epilepsy-specific patient-reported health instruments

Instrument	***Domains (no. items)***		***Response options***	***Score***	***Administration/ Completion time***
Epilepsy Surgery Inventory-55 (ESI-55)	*5 subscales/55 items* 1. SF-36* 2. Cognitive function (5) 3. Role limitations (8) 4. Health perceptions (4) 5. Overall QoL (2) ** Includes the 7 subscales of Rand SF-36 as generic core*		Various, including 5- and 6-point scales, dichotomous responses and VAS	Three summary composite scores computed: mental functioning, physical functioning, role functioning	15 minutes to complete
Katz Adjustment Scales (adapted for epilepsy)	Original KAS-R: 12 subscales/76 of 127 potential items 1. General psychopathology (24) 2. Suspiciousness (4) 3. Anxiety (6) 4. Negativism (9) 5. Confusion (3) 6. Belligerence (4) 7. Withdrawal (5) 8. Bizarreness (5) 9. Hyperactivity (3) 10. Helplessness (4) 11. Verbal expansiveness (5) 12. Nervousness (4) Misc. *(not used)* (50)	*Revised by Vickrey et al. (1992b) for epilepsy*: 14 subscales/127 items 1. Oversensitivity/ fearfulness (18) 2. Social (10) 3. Irritability (9) 4. Dependency (15) 5. Acting out (12) 6. Paranoia (5) 7. Abnormal thought process (5) 8. Withdrawal-R (11) 9. Emotional lability (6) 10. Nervousness-R (5) 11. Sociopathy (4) 12. Bizarreness-R (4) 13. Hyperactivity-R (4) 14. Disorientation (5) Misc *(not used)* (13)	4-point scales	Summed. Higher scores indicate better functioning (transformed to 0-100 point scales).	No details

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Liverpool Quality of Life (LQOL) Battery *and* Liverpool Seizure Severity Scale (LSSS)	LQOL: 1. Adverse drug effects Scale(21) 2. Impact of Epilepsy Scale (8) General: 1. Affect Balance Scale (10) 2. Mastery Scale (7) 3. Life Fulfilment Scale (20) *[Early versions included other existing psychological and health status scales]* LSSS: Seizure Severity Scale (2 subscales): 1. Perceived control over seizures (9) 2. Ictal and post-ictal symptoms (11)	4-point Likert; rating scales, dichotomous Present/Absent	Adverse drug effects summed, with higher scores indicating more problems. Impact: mean scores used. Affect Balance: range 1-9 with higher scores indicating more positive balance; Mastery: range 7-28, with higher scores indicating greater mastery; Fulfilment: computed difference between actual and ideal life satisfaction scores.	30-45 minutes to complete
Quality of Life in Epilepsy-10 (QOLIE-10)	*7 subscales/10 items:* 1. Seizure worry(1) 2. Overall QoL (1) 3. Emotional well-being (1) 4. Energy/fatigue (1) 5. Cognitive functioning (1) 6. Medication effects (2) 7. Social function (3)	5 point Likert; 0-10 QoL rating scale	Summation and domain score (weighted) Higher scores represent better function on all scales Index: 0 = worst QoL, 100 = best QoL Higher scores represent better function on all scales	Self-report 'few minutes to complete'
Quality of Life in Epilepsy-31 (QOLIE-31)	*7 subscales/31 items:* 1. Seizure worry (5) 2. Overall QoL (2) 3. Emotional well-being (5) 4. Energy/Fatigue (4) 5. Cognitive functioning (6) 6. Medication effects (3) 7. Social functioning (5) Health status (1) *[not included in total score]*	5 point Likert; 0-10 QoL rating scale	Summation and domain score (weighted) Higher scores represent better function on all scales Index: 0 = worst QoL, 100 = best QoL Higher scores represent better function on all scales	Self-report 15 minutes to complete

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Quality of Life in Epilepsy-89 (QOLE-89)	*17 subscales/89 items [SF-36 supplemented by 53 items specific to epilepsy] grouped into four factors:* 1.Seizure-specific effects (seizure worry, health discouragement, medicine effects, work or driving or social function) 2. Cognition (language, memory, attention) 3. Physical health (role limitations/physical pain, health perceptions, or physical function) 4. Mental health (overall quality of life, emotional well-being, role limitations/ emotional, social isolation, social support, and energy or fatigue) *[Original 99 items reduced to 87, two items added on overall health perception and sexual functioning to produce 89 items.]*	4- & 6-point Likert scales; dichotomous Yes/No; 0-10 QoL rating scale, 1-5 VAS scale, 0-100 VAS scale	Summation and domain score (weighted) Overall score = weighted sum of subscale scores Subscale scores = mean scores across items within the subscale Index: 0 = worst QoL, 100 = best QoL Higher scores represent better function on all scales	Self-report 28.4 (±15.6) minutes to complete
Washington Psychosocial Seizure Inventory (WPSI)	*7 subscales/132 items:* 1. Family background (11) 2. Emotional adjustment (34) 3. Interpersonal adjustment (22) 4. Vocational adjustment (13) 5. Financial status (7) 5. Adjustment to seizures (15) 6. Medicine/medical management (8) 7. Overall psychosocial functioning (57) *Later an item measuring QoL was added*	Dichotomous Yes/No	Summation and domain score Higher scores indicate poorer adjustment. There are four profiles: 1) No problems; 2) Possible or slight difficulties; 3) Definite problems; 4) Severe or major problems.	Interviewer 15-20 minutes to complete
Side-Effect and Life Satisfaction (SEALS)	5 subscales/50 *and shorter form versions; Gillam et al. (1996) standardisation:* 1. General cognitive difficulties (17) 2. Satisfaction/Dysphoria (8) 3. Fatigue/Tiredness (5) 4. Temper(4) 5. Worry (4) *[early version included frequency of ADL)*	4-point Likert	Summation and domain scores	

Table 7.7: Summary of epilepsy-specific instruments: health status domains *(after Fitzpatrick et al., 1998)*

Instrument	***Instrument domains***								
	Physical function	**Symptoms**	**Global judgement of health**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct***	**Treatment satisfaction**
Liverpool Quality of Life (LQOL) Battery		x	x	x	x		x	x	
Epilepsy Surgery Inventory-55 (ESI-55)	x	x	x	x	x	x	x	x	
Katz Adjustment Scale				x		x			
Quality of Life in Epilepsy-10 (QOLIE-10)		x		x	x	x	x	x	
Quality of Life in Epilepsy-31 (QOLIE-31)		x	x	x	x	x	x	x	
Quality of Life in Epilepsy-89 (QOLE-89)	x	x	x	x	x	x	x	x	
Washington Psychosocial Seizure Inventory (WPSI)				x	x		x	x	x
Side-Effect and Life Satisfaction (SEALS)	x	x		x		x	x	x	

*** Includes global (HR) QoL ratings**

EPILEPSY-SPECIFIC PATIENT-REPORTED HEALTH INSTRUMENTS

a) Epilepsy Surgery Inventory-55 (ESI-55)

The seven studies included contained cohorts of patients who has undergone surgery, those who had undergone or been assessed for surgery, a cohort eligible for surgery, and mixed groups of patients (Langfitt, 1995; McLachlan et al., 1997; O'Donogue et al., 1998; Selai et al., 2000; Vickrey et al., 1992a, 1995; Wiebe et al., 1997). Both males and females were included. Mean ages ranged from 28 to 34 (actual age-ranges not given).

Reliability
Vickrey et al. (1992a) reported internal consistency reliability coefficients for the ESI-55 (Cronbach's alpha: 0.68–0.88). Multi-trait scaling supported item discrimination across subscales. Good internal consistency correlation coefficients and Cronbach's alphas of 0.62 to 0.94 were also reported by Langfitt (1995). In addition, high inter-rater agreement (kappa 0.91) has been obtained (Langfitt, 1995).

Validity
Factor analysis has confirmed mental and physical health factors, and a third defined by cognitive function and role limitations (Vickrey et al., 1992a).

Epilepsy-specific patient-reported health instruments
Selai et al. (2000) reported that the measure correlated well with the QOLAS. Construct validity was further supported by correlations between the ESI-55 and the WPSI emotional adjustment domain (Langfitt, 1995).

Measures of epilepsy function
O'Donoghue et al. (1998) reported that only some ESI-55 subscales were associated with seizure frequency, and the ESI-55 was less sensitive to outcome after surgery than the SHE (Subjective Handicap of Epilepsy scale), which measures subjective evaluations of handicap in epilepsy. The health perceptions subscale has been reported to have the greatest sensitivity in discriminating between patients varying by seizure type and frequency (Vickrey et al., 1995). Vickrey et al. (1992a) reported that patients who were seizure-free following surgery were significantly more likely to have higher ESI-55 scores than patients who continued to have seizures. The ESI-55 was able to discriminate between patients having only auras and seizure-free patients, but not between aura-only and seizure-free patients (Vickrey et al., 1995). (See also 'Responsiveness'.)

Generic heath status
Construct validity was supported by correlations between the ESI-55 and comparable functioning domains on the SIP; and by correlations between the ESI-55 and measures of mood (Vickrey et al., 1992a).

Responsiveness
McLachlan et al. (1997) reported that seizure-free patients and those with at least a 90% reduction in seizure frequency, reported improved QoL on five of 10 ESI-55 subscales and overall score at 24 months. The ESI-55 was also sensitive to < 90% seizure reduction. But only one ESI-55 subscale at six months and two at 12 months

showed significant differences between groups. Selai et al. (2000) reported that the ESI-55 scales for Mental health and Physical health showed improvements at one-year patient follow-up, although Role functioning did not achieve significance. Wiebe et al. (1997) examined the responsiveness at one year of the ESI-55, and supported the responsiveness of the ESI-55.

Expert consensus
Review of piloted instrument by panel of nine health professionals (Vickrey et al., 1992a). No further details given.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
The ESI-55 takes an average of 15-16 minutes to complete, compared with 15.8 for the WSPI and 34.5 for the SIP (Langfitt, 1995).

Table 7.8: Developmental and evaluation studies relating to the Epilepsy Surgery Inventory-55 (ESI-55)

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Epilepsy Surgery Inventory-55 (ESI-55)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Langfitt (1995) USA	Patients with intractable epilepsy grouped into complex partial seizures with secondary generalization, complex partial seizures only, having undergone anterior temporal lobectomy 6 months+ earlier (71) Age: mean 34.6, 34.3, 31.2, respectively Postal Inpatients	Internal consistency ✓ Inter-rater ✓	Face ✓ Content ✓ Construct ✓ Concurrent ✓				✓
McLachlan et al. (1997) Canada	Epilepsy patients, eligible for temporal lobectomy, who had surgery or medical therapy (81) Age: mean 31.9, 34.2, respectively Self-administered Inpatients		Construct ✓	✓			
O'Donogue et al. (1998) UK	Epilepsy patients (287) Age: median 34 Outpatients		Construct ✓				
Selai et al. (2000) UK	Epilepsy patients (145, 45 followed up) Age: not given Interview Inpatients		Construct ✓ Concurrent ✓	✓			
Vickrey et al. (1995) USA	Cohort of patients who had undergone surgery for intractable epilepsy (133) Age: mean 28 Postal		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Epilepsy Surgery Inventory-55 (ESI-55)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Vickrey et al. (1992a) USA	Cohort of patients who had undergone surgery or assessed without having surgery (200) Age: mean 34 Postal	Internal consistency ✓	Construct ✓		✓		
Wiebe et al. (1997) Canada	Surgically and medically treated epilepsy patients (57) Age: mean 32.6 and 36.7, respectively		Construct ✓	✓			

b) Katz Adjustment Scales

The Katz Adjustment Scale was developed to assess social behaviour and adjustment among patients with schizophrenia (Katz and Lyerly, 1963) and most of the validation studies are with mental health patients. The measure was revised and tested for use with epilepsy patients by Vickrey et al. (1992b). The revisions to the scale improved its scaling success, comparing item-scale correlations, and also increased the number of Cronbach's alpha reliability coefficients equalling or exceeding 0.70 from five out of 12 to 12 out of 14 scales. Their analyses overall supported construct validity. Nervousness, Dependency, Oversensitivity/fearfulness, and Withdrawal subscales were the most sensitive to seizure status, while Sociopathy and Hyperactivity were the least sensitive.

Table 7.9: Developmental and evaluation studies relating to the Katz Adjustment Scales

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Katz Adjustment Scales		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Vickrey et al. (1992b) USA	Epilepsy patients (328 and 193 cross-validation sample) Age: 328 patients - range 12-63, mean 30; 193 patients - range 16-66, mean 34 Self-completion by relative/close friend proxy Postal Outpatients, proxies	Internal consistency ✓	Construct ✓				

c) Liverpool Quality of Life (LQOL) Battery and Seizure Severity Scale (LSSS)

The nine studies examining these instruments studies included both men and women, and mainly outpatients (where specified) with ages ranging from 15 to 78 years (Baker et al., 1993, 1994; Buck et al., 1999; Chadwick 1994; Jacoby et al., 1993; Rapp et al., 1998; Smith et al., 1991; Wiebe et al., 2001; Wagner et al., 1995). Overviews of the scales have been published by Baker (1998); Baker et al. (1994) Cramer and French (2001).

Reliability

Tests of internal consistency of an early LQOL model showed Cronbach's alphas ranging from weak to strong (0.35 to 0.84) and from 0.69 to 0.85 for the two subscales of the LSSS (Baker et al., 1993). Cronbach's alphas for Personal and Material Fulfilment were later reported to be 0.68-0.77 (Baker et al., 1994). Jacoby et al. (1993) tested the Impact subscale within the LQOL and reported the Cronbach's alpha to be 0.65, but increasing to 0.82 if the work item was removed. With the exception of the Perceptions subscale, the minimum criterion for internal consistency for scales under early evaluation (> 0.50) were met for all scales (Wagner et al., 1995). The internal consistency of the LQOL has since been found to be adequate, although test-retest correlations are more variable (Rapp et al., 1998).

Validity

Smith et al. (1991) tested an early version of the LQOL model, and found that while no associations were found between seizure frequency and psychological factors, seizure severity was the most significant predictor of self-esteem, control and anxiety. Baker et al. (1993) also tested an early version of the model, and reported that the happiness and mastery subscales of the LQOLS and the subscales of the LSSS were able to detect treatment effects, supporting the construct validity of both.

Baker et al. (1994) reported moderate to high correlations for the Impact subscale of the LQOL and affect balance, anxiety, self-esteem and mastery (> 0.4) depression and perceived QoL (> 0.6). The only significant correlations for the Material fulfilment subscale were with Impact of epilepsy and Perceived QoL. The revised Impact subscale correlated significantly with the other psychological subscales in the battery by -0.21 to 0.6, with the exception of the partner item which failed to correlate significantly with three of the seven subscales tested. The total Impact score was significantly correlated with all psychological subscales (r = 0.45-0.66), supporting construct validity (Jacoby et al., 1993).

Epilepsy-specific patient-reported health instruments

Rapp et al. (1998) reported that the LSSS and the LQOL instrument correlated well with the ESI-55.

Measures of epilepsy function

There is some inconsistency of results in this area. The Ictal subscale, but not the Perceptions subscale, has been found to discriminate between seizure types (Baker et al., 1993). Rapp et al. (1998) reported that most of the LQOL subscales were significantly associated with seizure severity (LSSS), although none distinguished between patients with different seizure types. But both seizure type and frequency have also been found to be key predictors on all items of the Impact subscale (Buck et al., 1999). Wagner et al. (1995) found that the scales varied widely in their ability to discriminate between groups of patients known to differ clinically. Chadwick (1994) did find that seizure frequency was reduced with medication, compared with a placebo group. Differences with seizure severity and seizure ratings were small but significant. A critical review of the LSSS can be found in Cramer and French (2001).

Generic heath status

The NHP, and a range of generic psychological scales (HADS [Hospital Anxiety and Depression Scale], POMS [Profile of Mood States], Rosenberg Self-Esteem Scale [RSES]) as well as the SEALS ADL measure, were included in an early LQOL battery. The Cronbach's alpha of the NHP was reported to be 0.76 (Baker et al., 1993). No treatment effects were found for the NHP, nor for the HADS, RSES and POMS. The NHP, along with the Social Problems Questionnaire (SPQ), ADL scale of the SEALS inventory and the POMS, were excluded in later versions.

Wagner et al. (1995) used the LQOL and LSSS with the SF-36, and reported that, although the SF-36 had large ceiling effects, it discriminated better than epilepsy-specific scales among different disease severity groups. Buck et al. (1999) reported that the SF-36 subscales were all significantly correlated with seizure type and frequency.

Responsiveness

Threshold values for detecting clinically important changes were small for the LSSS and for the Impact of Epilepsy and Adverse drug events subscales of the LQOL (Wiebe et al., 2001).

Precision

Floor and ceiling effects were small in one study of the Impact subscale of the LQOL and the LSSS (Wiebe et al., 2001), although larger ceiling effects were reported for both LQOL and LSSS by Wagner et al. (1995).

Acceptability

Smith et al. (1991) commented on the high completion rate of the LSSS and an early version of the LQOL, and the acceptability of the battery of questionnaires to patients.

Feasibility

No evidence reported.

Table 7.10: Developmental and evaluation studies relating to the Liverpool Quality of Life (LQOL) Battery and Liverpool Seizure Severity Scale (LSSS)

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Liverpool Quality of Life (LQOL) Battery and Liverpool Seizure Severity Scale (LSSS)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Baker et al. (1993) UK	Patients with refractory epilepsy (81) Age: range 15-67, mean 33.7 Self-administration	Internal consistency ✓	Construct ✓	✓			
Baker et al. (1994) UK	Patients with epilepsy (75) Age: range 15-68, mean 33.3 Self-administration Outpatients	Internal consistency ✓	Construct ✓ Concurrent ✓				
Buck et al. (1999) Eight European countries	Epilepsy (4929) Age: range 16-90, mean 37 Postal Outpatients and support groups		Construct ✓				
Chadwick (1994) UK	Epilepsy patients in drug trial (81) Age: range 15-67, mean 33.7 LSSS only In- or outpatients not specified Mode of administration unclear		Construct ✓ Concurrent ✓				
Jacoby et al. (1993) UK	Epilepsy patients (75) Age: range 15-68, mean 33 Self-administration Outpatients	Internal consistency ✓	Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Liverpool Quality of Life (LQOL) Battery and Liverpool Seizure Severity Scale (LSSS)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Rapp et al. (1998) USA	Epilepsy patients experiencing seizures in previous 4 weeks (92) Age: mean 39.04 Self-administered Outpatients	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓				
Smith et al. (1991) UK	Patients with medically refractory partial seizures (100) Age: range 15-67, mean 32.7 Self-administration In- or outpatients not specified		Construct ✓			✓	
Wagner et al. (1995) UK	Epilepsy patients on AED therapy in multicentre study (136) Age: range 15-78, mean 34.9 Self-administration Outpatients	Internal consistency ✓	Construct ✓		✓		
Wiebe et al. (2001) Canada	Stable epilepsy patients, candidates for surgery (40) Age: mean 36 In- or outpatients not specified Mode of administration unclear			✓	✓		

QUALITY OF LIFE IN EPILEPSY 10-, -31-, & 89-ITEM VERSIONS (QOLIE)

The early development of the QOLIE as a test battery of 98 items, constructed using the Rand SF-36 as a generic core, was described by Perrine (1993). This version was reported to have good reliability and validity. Two studies were identified which evaluated the Quality of Life in Epilepsy 10-item version, eight which evaluated the 31-item version, and 13 which evaluated the 89-item version.

d) Quality of Life in Epilepsy-89 (QOLIE-89)

13 studies of the QOLIE-89 were identified (Birbeck et al., 2000; Breier et al., 1998; Devinsky et al., 1995; Fargo et al., 2004; Hays et al., 1995; Kim et al., 2003; Leidy et al., 1999b; Loring et al., 2004, 2005; Perrine et al., 1995; Vickrey et al., 2000; Wiebe et al., 2001, 2002). Of these, most were observational, based on convenience, epilepsy clinic or centre out- or inpatients (the latter was not always specified), and one study involved randomisation of patients to telephone interview or self-completion questionnaire. All involved self-completion of the questionnaire. The studies included both adult men and women, with an age-range (where given) of 16-90 years.

Reliability

Devinsky et al. (1995), in the development phase of the instrument, reported the Cronbach's alpha for the subscales as high, ranging from 0.78 (medication effects) to 0.92 (attention/concentration), and 0.97 for the overall score. These exceeded the generally accepted criterion for acceptability of 0.70. Leidy et al. (1999b) also reported on internal consistency in their study comparing self- and telephone completion. For both methods they reported Cronbach's alpha for the 17 subscales to be high, ranging between 0.76 and 0.95.

Devinsky et al. (1995) reported test-retest reliability (up to 91 days) to be good overall, with product moment correlations ranging from $r = 0.58$ to $r = 0.86$ for the 17 scales. Apart from the two role limitations, pain and medication effects subscales, the remaining subscales exceeded the generally accepted criterion for acceptability for group comparisons of 0.70. Leidy et al. (1999b) reported high test-retest correlations at two weeks.

Devinsky et al. (1995) reported patient-proxy agreement product moment correlations to be low to modest, although significant, ranging from $r = 0.29$ (role limitations, emotional) to $r = 0.57$ (work/social function). Hays et al. (1995), however, reported that while proxy ratings can be substituted for patients' ratings in group comparisons with adequate reliability and validity, caution is needed for individual comparisons given the discrepancies for more subjective measures (cognitive functioning, health perceptions, seizure distress).

The instrument included 99 items at initial administration, 86 of which, across 17 subscales, were retained after multitrait scaling. Factor analysis of the 17 subscales yielded four underlying dimensions of health: an epilepsy targeted dimension (seizure worry, health discouragement, medicine effects, work/driving/or social function), cognition (language, memory, attention), mental health (overall QoL, emotional well-being, role limitations-emotional, social isolation, social support, energy/fatigue), and

physical health (role limitations-physical, pain, health perceptions, physical function) (Devinsky et al., 1995).

Leidy et al. (1999b) reported that the mode of administration (telephone interview or self-completion) did not influence the reliability or validity of results.

Validity

Initial testing was conducted by Devinsky et al. (1995), with 304 epilepsy patients and their relative/friend proxies from 25 epilepsy centers in the USA (130 men and 174 women), with a mean age of 36 years (range 17-63). The authors reported that construct validity was supported by significant patient-proxy correlations (although these were low to modest and significant correlations between the instrument and seizure frequency over the past year, neuropsychological tests and emotional and cognitive function. Hays et al. (1995) have reported caution in interpretation – see 'Reliability').

Loring et al. (2004) reported significant associations between the QOLIE-89 and measures of depressive symptoms (using the Beck Depression Inventory) as well as seizure worry (using the EFA [Epilepsy Foundation of America] Concerns Index), supporting construct validity. The EFA Concerns Index, a measure of the experience of epilepsy in relation to everyday activities, correlated variously with QOLIE-89; the strongest correlations were with the Work/Driving/Social Function subscale (Loring et al., 2005). Breier et al. (1998) reported that the QOLIE-89 Memory, Language and Attention/Concentration scales correlated significantly with the Minnesota Multiphasic Personality Inventory-2, supporting construct validity.

Socio-demographic variables

Reporting of results by socio-demographic variables was rare. Loring et al. (2004) reported linear regression analyses showing that older patients developing seizures had lower QOLIE-89 scores than those developing epilepsy at younger ages. Higher years of education were also an independent influencer of higher QOLIE-89 scores.

Measures of epilepsy function

Devinsky et al. (1995) reported significant correlations between the instrument and seizure frequency over the past year, as well as neuropsychological tests, emotional and cognitive function. Vickrey et al. (2000) reported weak, but significant, associations between seizure severity score (National Hospital Seizure Severity Scale-3) and the QOLIE-89 subscale and overall scores, although not all items on either scale achieved significant correlations. However, Fargo et al. (2004) reported that, in patients with epilepsy or psychogenic non-epileptic seizures, while the self-reports of neurocognitive functioning (memory, language, attention/concentration) with the QOLIE-89 correlated significantly with mood, not all self-reports of neuro-psychological functioning were accurate when tested against neuropsychological tests. But Perrine et al. (1995) reported the QOLIE-89 correlated adequately with a wide range of tests of neuropsychological measures, and supported the validity of the QOLIE-89.

Generic heath status

Birbeck et al. (2000) compared the SF-36 and SF-12 with the Quality of Life in Epilepsy (QOLIE) shorter and long (31- and 89-item) versions. They reported that the

epilepsy-specific measure had larger responsiveness indices than the SF-36 or SF-12, although they were comparable in relation to mental and global health for change in seizure frequency. Wiebe et al. (2002) compared the QOLIE-31 and QOLIE-89 with the SF-36 and the Health Utilities Index version 3 (HUI-III). They reported all instruments to be robust, and able to distinguish accurately between different levels of patient-assessed changes in their condition

Responsiveness
Kim et al. (2003) examined responsiveness in the QOLIE-89 at baseline and 28-week follow-up and reported mixed results. Wiebe et al. (2001) reported that the threshold values for QOLIE-31 and QOLIE-89 were similar. The additional 58 items in QOLIE-89 did not significantly improve its ability to detect real (clinically important) change.

Interpretation
Wiebe et al. (2001, 2002) examined minimum clinically important change, small, medium and large changes, and changes needed to exclude chance error in the Health Utilities version 3, along with the SF-36, and the Quality of Life in Epilepsy Inventory 31- and 89-item versions (QOLIE-31, QOLIE-89). They reported (2002) that the HUI-III, and the other instruments, all differentiated between no change and minimum important change. Only the two QOLIE instruments distinguished accurately between minimum important change and medium or large change.

Precision
Floor and ceiling effects were reported for the QOLIE-89 by Leidy et al. (1999b). The subscales with the largest ceiling effects (> 25%) were generic SF-36 subscales: role limitations-emotional, role limitations-physical, physical function and pain. Breier et al. (1998) stated that there was evidence of possible floor effects in some subscales but did not produce the full data to illustrate this comment. Wiebe et al. (2001) found no floor and ceiling effects.

Acceptability
Devinsky et al. (1995) analysed patients' responses to an open-ended question on QoL. The analyses showed that patients had additional concerns not captured in the QOLIE questionnaire (e.g. about finances, athletic activities, pregnancy, birth defect, stigma, bother with medication and insomnia). However, the authors justified the questionnaire as it covered the areas raised by 'many other responses'. Devinsky et al. (1995) reported that the questionnaire took an average of 28.4 minutes (SD 15.6, range 6-135) to complete.

Feasibility
Devinsky et al. (1995) reported that the questionnaire took an average of 28.4 minutes (SD 15.6, range 6-135) to complete. No estimation of costs was provided.

Table 7.11: Developmental and evaluation studies relating to the Quality of Life in Epilepsy-89 (QOLIE-89) instrument

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Quality of Life in Epilepsy-89 (QOLIE-89)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Birbeck et al. (2000) USA	Participants in RCT medication for epilepsy (142) Age: range 18.8- 66.8, mean 38.2 Mode of administration not specified In- or outpatients not specified		Construct ✓	Time period ✓ [responsiveness indices for SF-36 and SF-12 also compared]			
Breier et al. (1998) USA	Patients with seizures and pseudo-seizures (68) Age: mean 35.1 (pseudo-seizure), 35.7 (epileptic) Self-administered Inpatients		Construct ✓		✓		
Devinsky et al. (1995) USA	Epilepsy patients (seizure-free for one year) and their accompanying friend/relative proxies (304) Age: range 17-63, mean 36 Self-administered Outpatients	Internal consistency ✓ Test-retest ✓ Inter-rater ✓	Content ✓ Construct ✓			✓	✓
Fargo et al. (2004) USA	Patients with epilepsy (45) or psychogenic non-epileptic seizures (37) Age: median 34.66 and 37, respectively Inpatients Self-administered		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Quality of Life in Epilepsy-89 (QOLE-89)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hays et al. (1995) USA	Epilepsy patients and their and their accompanying friend/relative proxies (292) Age: range 18-64, mean 36 Outpatients Self-administered	Internal consistency ✓ Inter-rater ✓ Test-retest ✓	Construct ✓				
Kim et al. (2003) USA	Patients in anti-epileptic drug trial (147) Age: mean 38.2 Self-administered			✓			
Leidy et al. (1999b) USA	Patients with epilepsy identified with patient records and clinic visits (139) Age: mean 38 Self- and telephone-administered	Internal consistency ✓ Test re-test ✓	Construct ✓		✓	✓	✓
Loring et al., (2004) USA	Patients with epilepsy undergoing evaluation for surgery (115) Age: mean 34.2 Self-administered within cognitive assessments		Construct ✓				
Loring et al. (2005) USA	Epilepsy patients assessed for surgery (189) Age: mean 34		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Quality of Life in Epilepsy-89 (QOLIE-89)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Perrine et al. (1995) USA	Epilepsy patients across epilepsy centres and clinics (304) Age: mean 36.1 Self-administered		Construct ✓				
Vickrey et al. (2000) USA	Epilepsy patients enrolled in seven centre prospective study (340) Age: range 18-66, mean 37.3 Self-administered within interview study		Construct ✓				
Wiebe et al. (2001) Canada	Stable epilepsy patients, candidates for surgery (40) Age: mean 36 Self administered In- or outpatients not specified			✓	✓		
Wiebe et al. (2002) Canada	Patients with difficult-to-control focal epilepsy investigated for surgery (136) Age: mean 36 Self-administered In- or outpatients not specified			✓			

e) Quality of Life in Epilepsy -31 (QOLIE-31)

Seven studies were identified that examined the QOLIE-31 (Cramer et al., 1998; 2000, 2004; Gunter, 2004; Wiebe et al., 2001, 2002). Of these, one was based on the dataset to develop the parent measure (QOLIE-89). The studies included both men and women, with an age-range of 17-92 years (range not always provided).

Reliability

Cramer et al. (1998) used the original QOLIE-89 developmental dataset to assess the reliability and validity of the 31-item version. Internal consistency was high, and Cronbach's alpha ranged from 0.77 (social functioning) to 0.85 (cognitive functioning). Test-retest results were good, and correlations ranged from r = 0.64-0.85. In every instance, individual scale items correlated more significantly with the scale on which that item loaded than with other scales, and item-scale correlations were uniformly high: (seizure worry (r = 0.68-0.79), overall QoL (r = 0.90-0.92), emotional well-being (r = 0.71-0.82), energy/fatigue (r = 0.81-0.85), cognitive functioning (r = 0.66-0.81), medication effects (r = 0.75-0.89), and work/driving/social functioning (r = 0.68-0.80).

Factor analysis of the 30 items yielded seven factors, paralleling the QOLIE-31 scale structure, with the exception of broader QoL. Factor analysis of the seven subscales yielded two factors: emotional and psychological issues (seizure worry, overall QoL, emotional well-being, energy/fatigue subscales) and mental efficiency (medical/social effects, work/driving/social and cognitive functioning subscales).

Validity

The QOLIE-31 correlated significantly with seizure frequency, supporting its construct validity. It was reported by Cramer et al. (2000) to discriminate between treatment groups in relation to seizure worry, cognitive functioning and total scores, and to detect a difference in the overall QoL subscale. Cramer et al. (2004) reported that the QOLIE-31 emotional well-being subscale correlated significantly with the Profile of Mood States (POMS), among patients in a drug trial, supporting its construct validity. All seven subscales (especially energy and well-being) correlated well with each of the six POMS subscales (tension, depression, anger, vigour, fatigue, confusion) (Cramer et al., 1998).

Socio-demographic variables

Cramer et al. (1998) reported that the overall QOLIE-31, the cognitive and the work subscales correlated with employment status.

Measures of epilepsy function

Cramer et al. (1998) reported that the QOLIE-31 subscales correlated significantly with neurotoxicity scores, but not with systematic toxicity scores.

Generic heath status

Birbeck et al. (2000) compared the SF-36 and SF-12 with the Quality of Life in Epilepsy (QOLIE) shorter and long (31- and 89-item) versions. They reported that the epilepsy-specific measure had larger responsiveness indices than the SF-36 or SF-12, although they were comparable in relation to mental and global health for change in seizure frequency. Wiebe et al. (2002) compared the QOLIE-31 and -89 with the SF-

36 and the Health Utilities Index version 3 (HUI-III). They reported all instruments to be robust, and able to distinguish accurately between different levels of patient-assessed changes in their condition.

Responsiveness

Gunter et al. (2004) reported results from a study of a disease management programme, showing that the patients in the intervention group significantly improved their scores at 6-8 weeks post-baseline, on the QOLIE-31 for Seizure worry and Emotional well-being subscales, while there were no changes in the control group. Cramer et al. (2000) also reported evidence of responsiveness to change (baseline and 18 weeks) among respondents in a medication trial. Wiebe et al. (2001) reported that the threshold values for QOLIE-31 and -89 were similar, and the additional 58 items in QOLIE-89 did not significantly improve its ability to detect real (clinically important) change.

Interpretation

Wiebe et al. (2001, 2002) examined minimum clinically important change, small, medium and large changes, and changes needed to exclude chance error in the Health Utilities version 3, along with the SF-36, and the Quality of Life in Epilepsy Inventory 31- and 89-item versions (QOLIE-31, QOLIE-89). They reported (2002) that the HUI-III, and the other instruments, all differentiated between no change and minimum important change. Only the two QOLIE instruments distinguished accurately between minimum important change and medium or large change.

Expert consensus

The subscales were selected from the full QOLIE, on the basis of those believed to be the most important to people with epilepsy, as determined by an expert panel; no further details were provided (Cramer et al., 1998).

Precision

Wiebe et al. (2001) found no floor and ceiling effects.

Acceptability

No specific evidence was found.

Feasibility

No specific evidence was found.

Table 7.12: Developmental and evaluation studies relating to the Quality of Life in Epilepsy-31 (QOLIE-31) instrument

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Quality of Life in Epilepsy-31 (QOLIE-31)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Birbeck et al. (2000) USA	Participants in RCT medication for epilepsy (142) Age: range 18.8-66.8, mean 38.2 Mode of administration not specified In- or outpatients not specified		Construct ✓	✓ [responsiveness indices for SF-36 and SF-12 also compared]			
Cramer et al. (1998) USA	Patients recruited from epilepsy clinics (304) Age: range 17-60, mean 36 Self-administered Outpatients	Internal consistency ✓ Test re-test ✓	Construct ✓				
Cramer et al (2000) USA	Patients in RCT epilepsy medication therapy (246) Age: range 16-70 Self-administered		Construct ✓	✓			
Cramer et al. (2004) USA	Epilepsy patients with poorly controlled seizures or experiencing unacceptable adverse effects from current medication before and after changes to medication (two comparative treatment arms) (196) Age: mean 43.4 and 44.9 in two arms Self-administered		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Quality of Life in Epilepsy-31 (QOLIE-31)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Gunter et al. (2004) USA	Patients in pre-and post-intervention evaluation of management of epilepsy (225) Age: range 18-92, mean 92 Self-administered		Construct ✓				
Wiebe et al. (2001) Canada	Stable epilepsy patients, candidates for surgery (40) Age: mean 36			✓	✓		
Wiebe et al. (2002) Canada	Patients with difficult-to-control focal epilepsy investigated for surgery (136) Age: mean 36 Self administered In- or outpatients not specified			✓			

f) Quality of Life in Epilepsy-10 (QOLIE-10)

Two studies reported psychometric properties of the QOLIE-10 (Cramer et al., 1996, 2000). One was based on patients recruited from seizure clinics (the same dataset used for the development of the full QOLIE-89 and the QOLIE-31) and the other was based on patients participating in a medication trial. All were adults, with mean ages of 36 and 38.7 years.

Reliability

Cramer et al. (1996) reported that test-retest correlations for all items and subscales were significant, and were moderate to high (Pearson's r = 0.48-0.81). Cronbach's alphas for the three subscales ranged from 0.48 to 0.51.

Validity

Three subscales were confirmed by factor analyses: Epilepsy effects, Mental health and Role function. The three resultant QOLIE-10 subscales correlated well with their QOLIE-89 counterpart subscales (r -0.78-0.92) (Cramer et al., 1996).

Measures of epilepsy function

Cramer et al. (1996) reported that measures of systemic toxicity and neurotoxicity scores correlated best with different QOLIE-10 subscales. Correlations were weak to modest. The authors interpreted these variations as suggesting that patients' perceptions approximately reflected clinical test results. Scales also varied by seizure frequency. Patients with low seizure frequency had better Role function scores (driving, work, social issues) than patients with moderate or high seizure frequency. The authors stated this supported the discriminant validity of the QOLIE-10.

Responsiveness

Cramer et al. (2000) reported evidence of responsiveness to change (baseline and 18 weeks) among respondents in a medication trial.

Precision

No specific evidence was found.

Acceptability

No specific evidence was found.

Feasibility

Cramer et al. (2000) reported that, while the QOLIE-10 was able to detect changes over time among patients in a drug trial, the longer QOLIE-31 is preferred as it provides more detailed information.

Table 7.13: Developmental and evaluation studies relating to the Quality of Life in Epilepsy-10 (QOLIE-10) instrument

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Quality of Life in Epilepsy 10 (QOLIE-10)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Cramer et al. (1996) USA	Patients recruited from epilepsy clinics (304) Age: range 17-60, mean 36 Self-administered Outpatients	Internal consistency ✓ Test-retest ✓	Construct ✓				✓
Cramer et al. (2000) USA	Patients in RCT epilepsy medication therapy (246) Age: range 16-70 Self-administered		Construct ✓	✓			

g) Washington Psychological Seizure Inventory (WPSI)

Four studies were identified on the WPSI, based on outpatients where specified. The age-range of patients included was 18-69, where given. Men and women were included. It was not always clear whether the instrument was adapted for self-completion, rather than being administered during an interview.

Reliability

Dodrill et al. (1980) reported on the development of the WPSI. Internal consistency, evaluated by split-half reliability, was modest to good, with most correlations ranging from 0.68 to 0.95, and one was 0.37. Test-retest correlations ranged from 0.58 to 0.58 to 0.84, with one at 0.28. Ratings by professionals showed good reliability, but ratings of patients' 'significant others' were less good.

Chang and Gehlert (2003) used item-response theory to evaluate how items in each clinical scale performed in relation to representing the underlying constructs being measured. They reported that most items within each scale fitted the measurement model well. All subscales were found to be acceptably unidimensional.

Validity

The WPSI was independently associated with preoperative adjustment and seizure-free outcomes, supporting its construct validity (Hermann et al., 1992).

Responsiveness

Wiebe et al. (1997) examined the responsiveness at one year of the WPSI and the ESI-55. They reported that all instruments registered some change, and supported the responsiveness of the ESI-55. But the WPSI was relatively unresponsive to small or medium changes.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 7.14: Developmental and evaluation studies relating to the Washington Psychosocial Seizure Inventory (WPSI)

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Washington Psychosocial Seizure Inventory (WPSI)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Dodrill et al. (1980) USA	Epilepsy patients (127) Age: range 18-56, mean 29.16 Outpatients Interview	Internal consistency ✓ Test-retest ✓ Inter-rater reliability ✓					
Chang and Gehlert (2003) USA	Epilepsy patients (145) Age: range 18-69, mean 39.6 Outpatients Interview	Item-response theory ✓					
Hermann et al. (1992) USA	Epilepsy patients with complex partial seizures, seizure-free or significantly improved post-surgery (97) Age: mean 30.6 and 30.4, respectively		Construct ✓				
Wiebe et al. (1997) Canada	Surgically and medically treated epilepsy patients (57) Age: mean 32.6 and 36.7, respectively			✓			

Other epilepsy-specific instruments

Less well-known and well-tested seizure severity and other epilepsy-specific scales can be found in reviews by Cramer and French (2001), and Trimble and Dodson (1994). The section below includes better known scales, but with relatively little evidence of testing with epilepsy patients than the scales reviewed above.

h) Side-Effect and Life Satisfaction Inventory (SEALS)

The SEALS, a measure of side-effects and satisfaction with drug therapy, has undergone some limited testing since its early development, with larger numbers of male and female patients, and ages ranging from 15-60 years. While five factors were confirmed, their structure is slightly different to the original (Gillham et al., 1996). Split-half coefficient of test-retest reliability was 0.792; while less than perfect, it was regarded as adequate. The SEALS is able to discriminate in the expected direction between patients taking two or more drugs compared with those taking none. It was also able to detect expected changes in patients' condition.

The initial version included a frequency of ADL subscale, but Baker et al. (1993) reported that it was unable to discriminate between patients taking medication or a placebo. Gillham et al. (2000) reported that a 38-item version of SEALS (subscale scores and the total score) did correlate significantly with generic psychological measures (POMS, HADS, Rand Medical Outcomes Study Cognitive Functioning Scale) (r = 0.53-0.84).

Table 7.15: Developmental and evaluation studies relating to the Side-Effect and Life Satisfaction Inventory

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Side Effect and Life Satisfaction Inventory (SEALS)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Baker et al. (1993) UK	Patients with refractory epilepsy (81) Age: range 15-67, mean 33.7 Self-administration		Construct ✓				
Gillham et al. (1996) UK	Epilepsy patients from 5 centres (45 for test-retest and 923 for validity) Age: 923 -range 15-60, mean 33.43	Test-retest ✓ Split half ✓	Construct ✓				
Gillham et al. (2000)	Epilepsy patients Age: range 18-71, mean 37.82 Self-administration Outpatients		Construct ✓				

Table 7.16: Other epilepsy-specific instruments identified from the review

The following table provides an overview of other records of epilepsy-specific instruments identified of either newly developed instruments or single-study reporting of measurement properties and/or evaluation.

Instrument/reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsive-ness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Epilepsy Foundation of America (EFA) Concerns Index **Loring et al. (2005)** **USA**	Epilepsy patients assessed for surgery (189) Age: mean 34 In- or outpatients not specified nor mode of administration		✓					EFA Concerns Index aims to measure the experience of epilepsy in relation to everyday activities. Modest correlations between EFA Concerns Index and cognitive measures, and varying correlations with QOLIE-89 reported - the strongest correlations were with the Work/Driving/Social Function subscale. Five factors identified: affective impact on enjoyment of life, general autonomy concerns, fear of seizure recurrence, concern of burdening family, perceived lack of understanding by others.

Instrument/reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsive-ness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Epilepsy Social Effects Scale **Chaplin et al. (1990)** **UK**	Epilepsy patients in two centres No further details provided In- or outpatients not specified Interview	Internal consistency ✓	✓					The questionnaire was developed to investigate the social effects of epilepsy. Developmental work reported with patients who were asked to generate statements, which were compared with existing questionnaires. Following piloting, the areas included were: Attitudes towards accepting attacks, Attitude to label Epilepsy, Fear of having seizures, Fear of stigma in employment, Lack of confidence about the future, Concern about performance at work, Concern about sexual relationships, Concern about platonic relationships, Concern about housing, Lack of confidence travelling, Adverse reaction on social life, Adverse reaction on leisure pursuits, Change of outlook on life, Difficulty communicating with family, Problems taking medication, Distrust of medical profession, Misconceptions about epilepsy, Depression or emotional reactions, Feeling increased social isolation,, Lethargy/lack of energy, Sleep disturbance. The response format was Yes/No answers, which, during piloting, some respondents found difficulty with, thus these were changes to levels of agreement/disagreement. Some statements were weighted. Validity was assessed against staff ratings of patients' behaviour which led to weak or modest correlations. Inter-scale correlations were generally high.

Instrument/reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsive-ness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Glasgow Epilepsy Outcome Scale (GEOS-90) **Espie et al. (1998, 2001)** **Scotland, UK**	1998: epilepsy patients (39), carers 2001: family of epilepsy patients (384), clinicians and staff Outpatients and in care Age: 39 in care environments, 31 living with family	Internal consistency ✓	✓					The scale measures types and degrees of concern in the treatment of people with epilepsy and mental retardation. The final scale, tested for factor structure, contains four subscales of Concerns about seizures (30 items), Concerns about treatment (26), Concerns about caring (14), and Concerns about social impact (20). Internal consistency was reported as high, and the scale could discriminate between patient groups. A short 35 item version was also developed (2001).
National Hospital Seizure Severity Scale (SSS) **Vickrey et al. (2000)** **USA**	Epilepsy patients enrolled in 7-centre prospective study (340) Age: range 18-66, mean 37.3 Interview		✓		✓			This was a refinement of a previously developed measure, containing 7 items on seizure severity. Instructions recommend completion by a witness to seizures in addition to the person with epilepsy. SSS was significantly, although weakly, associated with QoL. There were no floor, and few ceiling effects.

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsive-ness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Quality of Life Assessment Schedule (QOLAS) **Kendrick and Trimble (1994)** **UK**	Patients with chronic epilepsy (50) and undergoing surgery (11) Age: not given In- or outpatients not specified Mode of administration not specified	Test-retest ✓	✓					QOLAS uses repertory grid techniques, to elicit epilepsy patients' own constructs and concerns. Patients are first asked to define what aspects of their life are important to their QoL. The method permits an objective assessment of subjective feelings. Initial test-retest results were good; construct validity was partly supported. Lack of standardisation is likely to limit its appeal in clinic settings.
Selai et al. (2000) **UK**	Epilepsy patients (145, 45 followed up) Age: not given Interview Inpatients	Internal consistency ✓	✓	✓				Coefficient alpha 0.7; correlated well with ESI-55.
Quality of Life Index Epilepsy Version III **Ferrans and Powers (1985, 1992)**	No specific evidence was found							This scale was developed for use with haemodialysis patients, and a version was developed for use with epilepsy patients. Other versions exist – e.g. for cardiac and cancer patients. It consists of satisfaction and importance ratings of various areas of life. It is a well known (in cancer) but little used scale.

Instrument/reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsive-ness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Quality of Life in Newly Diagnosed Epilepsy Instrument (NEWQOL) **Abetz et al. (2000)** **UK and USA**	Patients with new-onset epilepsy (48 UK, 60 USA) Age: mean 35.3 UK, 36.5 USA Self-administration Outpatients Self-completion, followed by interviews	Internal consistency ✓ Test-retest ✓	✓					The NEWQOL includes a battery of previously validated multi-item scales and items, containing 93 items, 81 of which form eight multi-item subscales. It aims to measure epilepsy-specific QoL, and includes 13 subscales which measure: Anxiety, Depression, Social activities, Symptoms, Locus of Control/Mastery, Neuropsychological Problems, Social Stigma, Worry and Work Limitations. Single items measure general Health, Number of Seizures, Social Limitations, Social Support, Self-Concept, Ambition Limitations, Health Transition, and General Limitations. Five additional items measure supportive networks. All use Likert scaling and scores are summed. All multi-item scales had good test-retest reliability, acceptable internal consistency, and high item discriminant validity. The NEWQOL was able to discriminate between patient groups (particularly symptoms, psychological problems).
Subjective Handicap of Epilepsy (SHE) **O'Donoghue et al. (1998)** **UK**	Epilepsy patients (287) Age: median 34 Outpatients Postal	Internal consistency ✓ Test-retest ✓	✓			✓		SHE, measures subjective evaluations of handicap in epilepsy. It contains 32 items in 6 subscales: Work and activities, Social and personal, Self-perception, Physical, Life satisfaction, and Change. It takes 10 mins to complete. Cronbach's alphas were high (0.79-0.88), and test-retest results satisfactory. SHE was sensitive to seizure frequency and was more sensitive than the ESI-55 to outcome after surgery.

SUMMARY - GENERIC INSTRUMENTS

Seven generic instruments were included in the review. The SF-36 was the most used and well-tested patient-based generic measure for use with patients with epilepsy. It was recommended by Vickrey et al. (1992a) and is the generic core in the most established and well-tested epilepsy-specific measures. One study compared the SF-36 and the shorter form SF-12. Other generic instruments included the EuroQoL/EQ-5D, the Health Utilities Index, the Nottingham Health Profile, and the Sickness Impact Profile. The instruments were tested on a wide range of patient types and severities. Not all studies specified whether they included in- or outpatients. The content of the generic instruments are detailed in Chapter 3.

Most SF-36 studies tested this instrument for construct validity, and reported satisfactory results (Buck et al., 1999; Hermann et al., 1996; Jacoby et al., 1999; Leidy et al., 1999a; Wagner et al., 1995, 1996). Jacoby et al. (1999) reported the most explicit data for construct validity, and reported associations in the expected directions between the SF-36 and seizure frequency and type, other health problems and side-effects.

Three studies reported good results for internal consistency (Jacoby et al., 1999; Wagner et al., 1995, 1996), two for time-period responsiveness (Birbeck et al. 2000; Wiebe et al. 2002), three for precision (Jacoby et al., 1999; Leidy et al., 1999a; Wagner et al., 1995), and one reported patients' and doctors' preferences (Wagner et al., 1997). Overall, US results for internal consistency reliability were high (Cronbach's alpha = 0.73-0.93), although they were lower among UK patients (0.43-0.92) Wagner et al. (1996). Other tests for reliability have indicated that they are satisfactory, except in Germany, for unknown reasons (Jacoby et al., 1999). Floor and ceiling effects were evident in many of the SF-36 subscales (Jacoby et al., 1999; Leidy et al., 1999a; Wagner et al., 1995). Overall, the evidence supports the use of the SF-36 as a generic instrument with people with epilepsy. But it is less responsive than epilepsy-specific measures, such as the QOLIE (Birbeck et al., 2000).

The evidence is limited for the other instruments reviewed. Three studies used the EQ-5D on community, market research and hospital samples. Selai et al. (2000) found that it was not valid in detecting changes pre-and post-treatment for epilepsy. They reported that 42% of their sample questioned the EQ-5D VAS, and thus questioned the scale's content validity. There was less evidence for the HUI-III. Compared with the QOLIE, the HUI-III was not able to distinguish accurately between minimum important changes. One paper commented only on the Q-TWIST, generally poor results were reported for the NHP, and one study reported good results for reliability and validity for the SIP.

In conclusion, the SF-36 is the best tested generic instrument for use with epilepsy patients, although it has floor and ceiling effects. There is little evidence in relation to other generic or utility instruments.

SUMMARY - EPILEPSY-SPECIFIC INSTRUMENTS

Eight epilepsy-specific instruments were reviewed, and a small number of others were summarised. The most extensively used and tested specific instruments are the Epilepsy Surgery Inventory, the QOLIE-31 and QOLIE-89, the Washington Psychosocial Seizure Inventory, and the Liverpool QoL Battery and Seizure Severity Scale. Overall, there was good evidence of concurrent validity, when compared with generic measures.

The Washington Psychosocial Seizure Inventory was the most popular measure in the past. Four studies were identified. But it does not cover all important areas in relation to cognition, physical functioning, energy and overall QoL. An improvement on the scale is the Epilepsy Surgery Inventory which has been shown to be reliable and valid (Vickrey et al., 1992a). Seven studies were identified for the ESI. It includes the SF-36 as a generic core, and supplements it with epilepsy-specific items. Although it contains 55 items, the completion time is about 15 minutes. Vickrey et al. (1992a) reported acceptable to high internal consistency reliability coefficients (Cronbach's alpha: 0.68-0.88). Cronbach's alphas of 0.62 to 0.94 were also reported by Langfitt (1995). Selai et al. (2000) reported that the ESI-55 scales for Mental health and Physical health showed improvements at one year patient follow-up, although Role functioning did not achieve significance.

The Liverpool Battery aimed to focus on issues important to people with epilepsy, although it is lengthy and time-consuming. Nine studies were identified. The studies relating to these are not all easily identified as the measures were unlabelled during their early development. Internal consistency reliability coefficients have generally been reported to be just adequate to good (Cronbach's alpha: 0.68-0.88) (Rapp et al., 1998). Most of its subscales have been reported to be associated with seizure severity, although not all with seizure type (Baker et al., 1993; Rapp et al., 1998). Wagner et al. (1995) reported that the SF-36 discriminated better between different disease severity types.

An increasingly more popular measure is the QOLIE, particularly the 31- and 89-item versions. The 89-item version includes the Rand SF-36 as a generic core. 13 studies were identified for the QOLIE-89, eight for the QOLIE-31, and two for the QOLIE-10. The QOLIE has also been tested on a wide range of patients (whether in- or outpatients was not always clearly specified). Cronbach's alphas were high for the QOLIE-89 (Cronbach's alpha: 0.76-0.97) (Devinsky et al, 1995; Leidy et al., 199b). Other tests for reliability were generally good. Caution is needed if substituting proxy for patient assessments, as concordance is not always good (Devinsky et al., 1995; Hays et al., 1995). More mixed results for validity have been reported, although construct validity was generally supported. One problem is that not all self-reports of neuropsychological functioning correlate well with neuropsychological measures (Fargo et al., 2004), although others have reported the correlations to be adequate (Perrine et al., 1995). While 12 studies evaluated the QOLIE-89, just eight were included for the shorter form QOLIE-31. Overall, the reliability and validity of the QOLIE-31 was reported to be good.

In sum, the most robust and popular measures are the ESI-55 and the QOLIE-89, although all require further evidence and many need better clarification of sample sources and types.

DISCUSSION AND RECOMMENDATIONS

A large number of studies were reviewed, the majority of which were based in North America, with the exception of the Liverpool Battery which was developed in the UK.

The SF-36 is the most widely applied and tested generic instrument. The most psychometrically sound and popular measures epilepsy-specific measures are the ESI-55 and the QOLIE-89. One possible limitation of the ESI-55 is that it was developed for use in the context of surgery for epilepsy and the majority of evidence derives from that context. It would be helpful to have evidence from more general contexts. There was little evidence in support of the use of the utility measures.

The generic instruments were multidimensional indicators of broader health or health-related quality of life. Few examined acceptability in any depth, or feasibility with this patient population. The SF-36 is a popular core in some of the epilepsy-specific measures reviewed, and many indicators built on earlier scales, the views of expert panels, professionals and behavioural scientists. More information on how relevant and important the items are to patients themselves is needed.

Relatively few investigators examined the performance of measurement scales against each other. The SF-36 and SF-12 have been tested against the QOLIE scales, with some inconsistent results. More comparisons have been made between the epilepsy-specific measures and domain-specific measures (e.g. psychological mood). The lack of information on scale distribution and variation by patients' socio-demographic variables was also noticeable.

Recommendations

The evidence summarised here supports the use of the SF-36 as a generic tool, with patients with epilepsy. Indeed, popular epilepsy-specific instruments have included this instrument within their generic core. The most extensively used and tested epilepsy-specific instruments are the Epilepsy Surgery Inventory, the QOLIE-31 and QOLIE-89, the Washington Psychosocial Seizure Inventory, the Liverpool QoL Battery and Seizure Severity Scale. The measures which are recommended are the ESI-55 and the QOLIE-89, although they require further evidence, especially with European populations, and, for ESI-55 testing outside of the specific surgical context. It also should be noted that the QOLIE family of questionnaires was originally derived, in item content, from the SF-36. Given that it is often recommended that a disease-specific and generic measure be used in conjunction, it may not be sensible to combine the QOLIE with the SF-36 in this way.

REFERENCES

Abetz LN, Jacoby A, Baker GA, McNulty P. Patient-based assessments of quality of life in newly diagnosed epilepsy patients: validation of the NEWQOL. *Epilepsia* 2000; **41:**1119-28.

Baker GA, Smith DF, Dewey ME, Jacoby A, Chadwick DW. The initial development of a health-related quality of life model as an outcome measure in epilepsy. *Epilepsy Research* 1993; **16:**65-81.

Baker GA, Jacoby A, Smith DF, Dewey ME, Chadwick DW. Development of a novel scale to assess life fulfilment as part of the further refinement of a quality-of-life model for epilepsy. *Epilepsia* 1994; **35:**591-6.

Baker GA. Quality of life and epilepsy: the Liverpool experience. *Clinical Therapeutics* 1998; **20:**A2-A12.

Birbeck GL, Kim S, Hays RD, Vickrey BG. Quality of life measures in epilepsy: how well can they detect change over time? *Neurology* 2000; **54:**1822-7.

Breier JI, Fuchs KL, Brookshire BL, Wheless J, Thomas AB, Constantinou J *et al*. Quality of life perception in patients with intractable epilepsy or pseudoseizures. *Archives of Neurology* 1998; **55:**660-5.

Brown SW, Tomlinson LL. Anticonvulsant side-effects: a self report questionnaire for use in community surveys. *British Journal of Clinical Practice* 1982; **18** - symposium supplement: 147-9.

Buck D, Jacoby A, Baker GA, Ley H, Steen N. Cross-cultural differences in health-related quality of life of people with epilepsy: findings from a European study. *Quality of Life Research* 1999; **8:**675-85.

Buelow JM, Ferrans CE. Quality of life in epilepsy. In Kanner AM, ed. *Psychiatric issues in epilepsy: a practical guide to diagnosis and treatment. 400 pp*, pp 307-18. Philadelphia, PA, USA: Lippincott Williams & Wilkins Publishers, 2001.

Chadwick DW. Measuring anti-epileptic therapies: the patient vs. the physician viewpoint. *Neurology* 1994; **44:**S24-S28.

Chang CH, Gehlert S. The Washington Psychosocial Seizure Inventory (WPSI): psychometric evaluation and future applications. *Seizure* 2003; **12:**261-7.

Chaplin JE, Yepez R, Shorvon S, Floyd M. A quantitative approach to measuring the social effects of epilepsy. *Neuroepidemiology* 1990; **9:**151-8.

Cramer JA, Perrine KR, Devinsky O, Meador K. A brief questionnaire to screen for quality of life in epilepsy: the QOLIE-10. *Epilepsia* 1996; **37:**577-82.

Cramer JA, Perrine KR, Devinsky O, Bryant-Comstock L, Meador K, Hermann BP. Development and cross-cultural translations of a 31-item Quality of Life in Epilepsy Inventory. *Epilepsia* 1998; **39:**81-8.

Cramer JA, Arrigo C, Van Hammee G, Bromfield EB. Comparison between the QOLIE-31 and derived QOLIE-10 in a clinical trial of levetiracetam. *Epilepsy Research* 2000; **41**:29-38.

Cramer JA, French J. Quantitative assessment of seizure severity for clinical trials: a review of approaches to seizure components. *Epilepsia* 2001; **42:**119-29.

Cramer JA, Hammer AE, Kustra RP. Improved mood states with lamotrigine in patients with epilepsy. *Epilepsy and Behavior* 2004; **5:**702-7.

Devinsky O,Vickrey BG. Quality of life: recent developments in the USA. In Trimble MR, Dodson WE, eds. *Epilepsy and quality of life,* New York, NY, USA: Raven Press, 1994.

Devinsky O, Vickrey BG, Cramer JA, Perrine KR, Hermann BP, Meador K *et al*. Development of the Quality of Life in Epilepsy Inventory. *Epilepsia* 1995; **36:**1089-104.

Dodrill CB, Batzel LW, Queisser HR, Temkin NR. An objective method for the assessment of psychological and social problems among epileptics. *Epilepsia* 1980; **21:**123-35.

Dodrill CB, Batzel LW. The Washington Psychosocial Seizure Inventory and quality of life in epilepsy. In Trimble MR, Dodson WE, eds. *Epilepsy and quality of life*, New York, NY, USA: Raven Press, 1994.

Dodrill CB, Batzel LW. Epilepsy. *Archives of Neurology* 1996; **53:**476-7.

Espie CA, Paul A, Graham M, Sterrick M, Foley J, McGarvey C. The Epilepsy Outcome Scale: the development of a measure for use with carers of people with epilepsy plus intellectual disability. *Journal of Intellectual Disability Research* 1998; **42:**90-6.

Espie CA, Watkins J, Duncan R, Espie A, Sterrick M, Brodie MJ *et al*. Development and validation of the Glasgow Epilepsy Outcome Scale/GEOS: a new instrument for measuring concerns about epilepsy in people with mental retardation. *Epilepsia* 2001; **42:**1043-51.

Fargo JD, Schefft BK, Szaflarski JP, Dulay MF, Testa SM, Privitera MD *et al*. Accuracy of self-reported neuropsychological functioning in individuals with epileptic or psychogenic nonepileptic seizures. *Epilepsy and Behavior* 2004; **5:**143-50.

Ferrans CE, Powers MJ. Quality of Life Index: development and psychometric properties. *Advances in Nursing Science* 1985; **8**:15-24.

Ferrans CE, Powers MJ. Psychometric assessment of the Quality of Life Index. *Research in Nursing and Health* 1992; **15**:29-38.

Gillham R, Baker GA, Thompson P, Birbeck K, McGuire A, Tomlinson L *et al*. Standardisation of a self-report questionnaire for use in evaluating cognitive,

affective, and behavioural side-effects of anti-epileptic drug treatments. *Epilepsy Research* 1996; **24:**47-55.

Gillham R, Bryant-Comstock L, Kane K. Validation of the Side-Effect and Life Satisfaction/SEALS inventory. *Seizure* 2000; **9:**458-63.

Gunter MJ, Brixner D, Von Worley A, Carter S, Gregory C. Impact of a seizure disorder disease management program on patient-reported quality of life. *Disease Management* 2004; **7:**333-47.

Hays RD, Vickrey BG, Hermann BP, Perrine KR, Cramer JA, Meador K *et al*. Agreement between self reports and proxy reports of quality of life in epilepsy patients. *Quality of Life Research* 1995; **4:**159-68.

Hermann BP, Wyler AR, Somes G. Preoperative psychological adjustment and surgical outcome are determinants of psychosocial status after anterior temporal lobectomy. *Journal of Neurology, Neurosurgery, and Psychiatry* 1992; **55:**491-6.

Hermann BP. The evolution of health-related quality of life assessment in epilepsy. *Quality of Life Research* 1995; **4:**87-100.

Hermann BP, Vickrey BG, Hays RD, Cramer JA, Devinsky O, Meador K *et al*. A comparison of health-related quality of life in patients with epilepsy, diabetes and multiple sclerosis. *Epilepsy Research* 1996; **25:**113-8.

Jacoby A, Baker GA, Smith DF, Dewey ME, Chadwick DW. Measuring the impact of epilepsy: the development of a novel scale. *Epilepsy Research* 1993; **16:**83-8.

Jacoby A. Assessing quality of life in patients with epilepsy. *Pharmacoeconomics* 1996; **9:**399-416.

Jacoby A, Baker GA, Steen N, Buck D. The SF-36 as a health status measure for epilepsy: a psychometric assessment. *Quality of Life Research* 1999; **8**:351-64.

Katz MM, Lyerly SB. Methods for measuring adjustment and social behavior in the community: I. Rationale, description, discriminative validity and scale development. *Psychological Reports* 1963; **13:**503-35.

Kendrick AM, Trimble MR. Repertory grid in the assessment of quality of life in patients with epilepsy: the Quality of Life Assessment Schedule. In Trimble MR, Dodson WE, eds. *Epilepsy and quality of life,* New York, NY, USA: Raven Press, 1994.

Kim S, Hays RD, Birbeck GL, Vickrey BG. Responsiveness of the Quality of Life in Epilepsy/QOLIE-89 in an anti-epileptic drug trial. *Quality of Life Research* 2003; **12:**147-55.

Langfitt JT, Wiebe S. Cost-effectiveness of epilepsy therapy: how should treatment effects be measured? *Epilepsia* 2002; **43**:17-24.

Langfitt JT. Comparison of the psychometric characteristics of three quality of life measures in intractable epilepsy. *Quality of Life Research* 1995; **4:**101-14.

Leidy NK, Rentz AM, Grace EM. Evaluating health-related quality of life outcomes in clinical trials of anti-epileptic drug therapy. *Epilepsia* 1998; **39:**965-77.

Leidy NK, Elixhauser A, Vickrey BG, Means E, Willian MK. Seizure frequency and the health-related quality of life of adults with epilepsy. *Neurology* 1999a; **53:**162.

Leidy NK, Elixhauser A, Rentz AM, Beach R, Pellock J, Schachter S *et al*. Telephone validation of the Quality of Life in Epilepsy Inventory-89/QOLIE-89. *Epilepsia* 1999b; **40**:97-106.

Leppik IE. Quality of life of people with epilepsy in the United States. *Clinical Therapeutics* 1998; **20:**A13-A18.

Loring DW, Meador KJ, Lee GP. Determinants of quality of life in epilepsy. *Epilepsy and Behavior* 2004; **5:**976-80.

Loring DW, Larrabee GJ, Meador KJ, Lee GP. Dimensions of the Epilepsy Foundation Concerns Index. *Epilepsy and Behavior* 2005; **6:**348-52.

McLachlan RS, Rose KJ, Derry PA, Bonnar C, Blume WT, Girvin JP. Health-related quality of life and seizure control in temporal lobe epilepsy. *Annals of Neurology* 1997; **41:**482-9.

O'Donoghue MF, Duncan JS, Sander JWA. The subjective handicap of epilepsy: a new approach to measuring treatment outcome. *Brain* 1998; **121:**317-34.

Perrine KR. A new quality of life inventory for epilepsy patients: interim results. *Epilepsia* 1993; **34:**S28-S33.

Perrine KR, Hermann BP, Meador KJ, Vickrey BG, Cramer JA, Hays RD *et al*. The relationship of neuropsychological functioning to quality of life in epilepsy. *Archives of Neurology* 1995; **52:**997-1003.

Rapp S, Shumaker SA, Smith T, Gibson P, Berzon RA, Hoffman R. Adaptation and evaluation of the Liverpool Seizure Severity Scale and Liverpool Quality of Life battery for American epilepsy patients. *Quality of Life Research* 1998; **7:**467-77.

Remák E, Hutton J, Selai CE, Trimble MR, Price MJ. A cost-utility analysis of adjunctive treatment with newer antiepileptic drugs in the UK. *Journal of Drug Assessment* 2004; **7:**109-120.

Schwartz CE, Cole BF, Vickrey BG, Gelber RD. The Q-TWIST approach to assessing health-related quality of life in epilepsy. *Quality of Life Research* 1995; **4:**135-41.

Selai CE, Trimble MR. Quality of life assessment in epilepsy: the state of the art. *Journal of Epilepsy* 1995; **8:**332-7.

Selai CE, Elstner K, Trimble MR. Quality of life pre- and post-epilepsy surgery. *Epilepsy Research* 2000; **38:**67-74.

Smith DF, Baker GA, Dewey ME, Jacoby A, Chadwick DW. Seizure frequency, patient-perceived seizure severity and the psychosocial consequences of intractable epilepsy. *Epilepsy Research* 1991; **9:**231-41.

Smith DF, Baker GA, Davies G, Dewey ME, Chadwick DW. Outcomes of add-on treatment with lamotrigine in partial epilepsy. *Epilepsia* 1993; **34:**312-22.

Trimble MR, Dodson WE eds. *Epilepsy and quality of life,* New York, NY, USA: Raven Press, 1994.

Trueman P, Duthie T. Use of the Hospital Anxiety and Depression Scale (HADS) in a large, general population study of epilepsy. *Quality of Life Newsletter* 1998; **19:**9-10.

Vickrey BG, Hays RD, Graber J, Rausch R, Engel JJ, Brook RH. A health-related quality of life instrument for patients evaluated for epilepsy surgery. *Medical Care* 1992a; **30:**299-319.

Vickrey BG, Hays RD, Brook RH, Rausch R. Reliability and validity of the Katz Adjustment Scales in an epilepsy sample. *Quality of Life Research* 1992b; **1:**63-72.

Vickrey BG, Hays RD, Engel JJ, Spritzer KL, Rogers WH, Rausch R *et al*. Outcome assessment for epilepsy surgery: the impact of measuring health-related quality of life. *Annals of Neurology* 1995; **37:**158-66.

Vickrey BG, Berg AT, Sperling MR, Shinnar S, Langfitt JT, Bazil CW *et al*. Relationships between seizure severity and health-related quality of life in refractory localization-related epilepsy. The Multicenter Epilepsy Surgery Study. *Epilepsia* 2000; **41:**760-4.

Wagner AK, Keller SD, Kosinski M, Baker GA, Jacoby A, Hsu MA *et al*. Advances in methods for assessing the impact of epilepsy and anti-epileptic drug therapy on patients' health-related quality of life. *Quality of Life Research* 1995; **4:**115-34.

Wagner AK, Bungay KM, Kosinski M, Bromfield EB, Ehrenberg BL. The health status of adults with epilepsy compared with that of people without chronic conditions. *Pharmacotherapy* 1996; **16**:1-9.

Wagner AK, Ehrenberg BL, Tran TA, Bungay KM, Cynn DJ, Rogers WH. Patient-based health status measurement in clinical practice: a study of its impact on epilepsy patients' care. *Quality of Life Research* 1997; **6:**329-41.

Wiebe S, Rose K, Derry PA, McLachlan R. Outcome assessment in epilepsy: comparative responsiveness of quality of life and psychosocial instruments. *Epilepsia* 1997; **38:**430-8.

Wiebe S, Eliasziw M, Matijevic S. Changes in quality of life in epilepsy: how large must they be to be real? *Epilepsia* 2001; **42:**113-8.

Wiebe S, Matijevic S, Eliasziw M, Derry PA. Clinically important change in quality of life in epilepsy. *Journal of Neurology, Neurosurgery and Psychiatry* 2002; **73:**116-20.

Chapter 8: Patient-reported Health Instruments used for people with heart failure

Heart failure is a common clinical syndrome resulting from cardiac disease. It is recognised by a constellation of symptoms and signs due to a failing heart, including dyspnoea, raspy breathing/wheezing, persistent coughing, blood-tinged sputum, weight gain due to fluid retention, swollen feet, ankles, legs, abdomen, sleeplessness, fatigue, listlessness, poor effort tolerance. The most common cause of heart failure in the developed world is coronary heart disease (CHD), although hypertension often co-exists.

Heart failure results in high levels of ill-health, disability and mortality, and is a heavy burden on health services. Quality of life, physical ability and prognosis for heart failure are poor, and less than half survive one year after first diagnosis. It is an area where rapid investigation, confirmation of diagnosis and prescribing of appropriate treatment is essential. Given the nature of the symptoms, and their potential impact on people's lives, both physically, socially and psychologically, it is an area where patient-based outcome assessments are important. However, the high mortality rate makes longer-term, patient-based outcome assessment difficult.

Search terms and results: identification of articles

At the time of the review, the PHI database contained 12,562 records (up to June 2005). Record abstracts and titles were initially searched using the terms 'heart failure *or* heart disease *or* cardiac disease *or* cardiac failure'; a further search was run using the terms 'cardiac *or* cardiovascular'. Finally, a search was made of the PHI keywords field using the subject classification keyword 'cardiovascular'. These searches generated 821 records, as shown in Table 8.1. All records were reviewed. When assessed against the review inclusion criteria, 173 articles were retrieved and reviewed in full. Of these, 89 articles were included in the review.

Table 8.1 Number of articles identified by the literature review

Source	***Results of search***	***No. of articles considered eligible***	***Number of articles included in review***
PHI database: original search (up to June 2005) **Total number = 12,562**	799	130	64
Additional PHI database search (July-December 2005) **Total number = 4021**	22	12	6
Supplementary searching		31	19
TOTAL	821	173	**89**

Supplementary searches included scanning the reference lists of key articles, checking instrument websites, where found, and drawing on other bibliographic resources. All titles of issues of the following journals published between January and September 2006 were scanned:

- Heart and Lung
- Journal of Cardiopulmonary Rehabilitation
- Journal of Cardiac Failure

- Health and Quality of Life Outcomes
- Medical Care
- Quality of Life Research

Identification of patient-reported health instruments
Five generic and four heart failure-specific instruments were included in the review. Instruments where there was no evidence that an English-language version had been tested were excluded. The developmental and evaluative studies relating to generic instruments are shown in Tables 8.2 to 8.5; those for heart failure-specific instruments are shown in Table 8.6 to 8.11. Table 8.12 provides an overview of other records of heart failure-specific instruments and generic cardiovascular disease (CVD) instruments used with heart failure patients.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Five generic instruments were identified which were evaluated with patients with heart failure. For full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36
b) SF-12
c) SIP
d) EQ-5D
e) Other utility measures

a) SF-36

21 studies assessed the SF-36 and eight examined the SF-12 in relation to with adults with heart failure. Most of these reported construct validity. The studies were based on population surveys and clinical samples of patients.

Reliability
Relatively few of the included studies assessed the reliability of the SF-36 when used with patients with heart failure. Internal consistency has been reported to be good for all the SF-36 sub-scales (Cronbach's alpha > 0.80) except for social functioning and general health perceptions (Wolinsky et al., 1998). Green et al. (2000) also reported social functioning to have poor reproducibility.

Validity
Jenkinson et al. (1997a) reported that the SF-36 was able to discriminate between patients aged 60 years and over with chronic heart failure and people aged 65 years and over, who reported no chronic illness. Hobbs et al. (2002) found that, in their population screening survey, people with heart failure had more severe physical impairment with the SF-36 than those with chronic lung disease or arthritis. Analyses of self-reports of chronic conditions in international surveys showed that arthritis, chronic lung disease and congestive heart failure (CHF) were the conditions with the greatest differences in physical component summary scores (Alonso et al., 2004). These differences were consistent across all SF-36 scales.

In a study of heart failure (HF) clinic patients, Havranek et al. (1999) reported significant correlations between the MOS Rand SF-36 physical component score, the Minnesota Living with Heart Failure Questionnaire (MLHFQ), timed walking, a visual analogue scale rating health status, and time-trade-off techniques. But there was no significant correlation between the MOS Rand SF-36 mental health component and 6-minute timed walking. Lalonde et al. (1999) found that the SF-36 Physical Component scale, but not the General Health Perceptions scale, was able to discriminate between CHD patients with various levels of physical disability). It correlated significantly with the Beck Depression Index, Hospital Anxiety and Depression Scale (HADS) and the Cardiac Depression Scale in a cardiac population (including HF patients) (Birks et al., 2004).

Socio-demographic variables

The SF-36 has been administered mainly to HF clinic patients or population samples, of both sexes, with ages ranging from 28-87 years, where given. Distributions or variations by socio-demographic characteristics were not given.

Heart failure-specific patient-reported health instruments

The general health perception scale of the Rand SF-36 has been reported to correlate significantly but modestly (r = 0.45) with the QoL domain of the Kansas City Cardiomyopathy Questionnaire (KCCQ), while the SF-36 social limitation scale correlated more highly with the KCCQ social limitation domain (r = 0.62) (Green et al., 2000). Oldridge et al. (2002) reported highly significant Pearson correlation coefficients between the SF-36, the Seattle Angina Questionnaire, the MacNew and the MLHFQ (ranging from 0.63 to 0.78 for the SF-36 and these instruments). However, the SF-36 was not able to discriminate between patients with heart failure, angina or myocardial infarction (MI). Apart from the physical functioning subscale; the SF-36 was also significantly associated with anxiety and depression (measured with the HADS). Dempster et al. (2004) examined the MacNew and the SF-36 and showed that the range of domain correlations between the instruments ranged from low to high (r = 0.18 to 0.85), although the highest were achieved for correlations between similar domains (r = 0.52 to 0.85).

Sneed et al. (2001) compared the SF-36 with the MLHFQ with a small sample of HF clinic attendees; the SF-36 was better able to differentiate physical and emotional aspects of QoL. However, Wolinsky et al. (1998) tested the SF-36 and the Chronic Heart Failure Questionnaire (CHQ), slightly adapted for use with coronary artery disease (CAD) and HF patients, among outpatients with CAD or chronic HF. While the SF-36 was more comprehensive in its coverage of health status domains, the CHQ was more psychometrically sound and had fewer problems with floor and ceiling effects, and was more reproducible and internally consistent.

Measures of HF Function

Arterburn et al. (2004) reported the SF-36 was associated with body mass index. While Hobbs et al. (2002) reported the SF-36 to be significantly correlated with New York Heart Association (NYHA) status, Havranek et al. (1999) reported significant correlations between the MOS Rand SF-36 physical component score and timed walking, and no significant correlation between the MOS Rand SF-36 mental health component and 6-minute timed walking.

Generic heath status
Lalonde et al. (1999) reported low to moderate correlations (between 0.12 and 0.51) between the SF-36 and rating scale, time trade-off and standard gamble health utility measures (although, technically, these are preference measures rather than comparable general health status instruments); significance levels were not reported.

Responsiveness
Gwadry-Sridhar et al. (2005), in an RCT of an educational intervention with HF inpatients, found no significant effect of time by intervention, or treatment intervention, in either of the SF-36 mental or physical summary scores, in contrast to the more sensitive MLHFQ. Green et al. (2000) reported that the Rand SF-36 was less responsive to important clinical change in HF patients than the MLHFQ. Wyrwich et al. (1999) compared change in the SF-36 and CHQ, and examined standard errors in detail. Both measures compared well at follow-up assessments of change. The physical component summary score has been shown to be predictive of decline in HF patients over four years (Bayliss et al., 2004); and the general health and physical role sub-scales were sensitive to changes in depression in patients with heart failure (Sullivan et al., 2004).

Interpretation
Expert consensus
Wyrwich et al. (2005) used Delphi and consensus panel techniques with expert panels of physicians to examine clinically important differences for the SF-36 and a modified CHQ. They reported on panel-derived thresholds for change over time.

Precision
Large ceiling effects (> 15%) have been found for the SF-36 role-physical, social functioning and role-emotional subscales, which potentially mask patient improvement or deterioration, and reduce scale sensitivity (Wyrwich et al., 1999). Ceiling effects were confirmed by Wolinsky et al. (1998).

Acceptability
In a study by Gwadry-Sridhar et al. (2005), 12 out of 134 patients were reported to find the questionnaire battery (which included the MLHFQ and SF-36) cumbersome and did not respond. High non-response to follow-up was reported in a study by Lalonde et al. (1999) with 75 (36%) participants refusing to come back for the second interview and 41% (20%) missing the second interview for various reasons. A study by Wolinsky et al. (1998) reported a 79% completion rate.

Feasibility
The SF-36 took 10-15 minutes to complete depending on whether self-completed or interviewer-read in a study by Sneed et al. (2001).

b) SF-12

Reliability
The internal consistency of the SF-12 is reportedly good, with all coefficients exceeding 0.70 in HF patients (Bennett et al., 2002), and in general CVD populations (Lim and Fisher, 1999), although Bennett et al. (2003) reported that the SF-12 was less reliable than the CHQ or the MLHFQ.

Validity
The construct validity was supported by associations with reported hospital admissions among mixed CVD populations (Lim and Fisher, 1999). Analyses have generally supported the convergent and discriminant validity of the SF-12 with HF patients (Bennett et al., 2002). Jenkinson and Layte (1997) have reported on the construction of the physical and mental component scales.

Socio-demographic variables
Lim and Fisher (1999) reported that the instrument discriminated between men and women, and being older in their study of mixed heart and stroke patients.

Heart failure-specific patient-reported health instruments
Correlations between the SF-12 and the CHQ and MLHFQ are moderate to high, being lower among the physical component summary score of the SF-12 and the CHQ and MLHFQ, than the mental component summary score of the SF-12 and the CHQ and MLHFQ (Bennett et al., 2002).

Measures of HF Function
The SF-12 physical component scale, but not the mental component scale, has correlated significantly with NYHA status in HF patients (Bennett et al., 2002).

Responsiveness
Bennett et al. (2003) reported that the SF-12 was less responsive to change in patients' condition than the CHQ or the MLHFQ. Ni et al. (2000) found the SF-12 to be more responsive to change in mental health, but less responsive to change in physical health at follow-up, than the MLHFQ. Ni et al. (2000) also found that the MLHFQ performed better than the SF-12 in ability to distinguish differences in perceived global health transition, and concluded that the SF-12 alone should not be used to measure changes in QoL of patients with HF. Spertus et al. (2005) found that the SF-12 and the EQ-5D did not exhibit much sensitivity to the magnitude of observed clinical change, unlike the KCCQ which demonstrated the highest discriminative abilities. Jenkinson et al. (1997b) reported that the SF-36 and SF-12 physical and mental component summary scores indicated the same magnitude of change over time.

Precision
Floor and ceiling effects with HF patients have been reported to be non-existent (Bennett et al., 2002; Ni et al., 2000).

Acceptability
SF-12 physical component and mental component sub-scales were missing for 13% of patients in the Bennett et al. study (2002) of clinic patients. Lim and Fisher (1999) reported a 22% non-completion rate in a mixed CVD population sample.

Feasibility
No specific evidence was found.

Table 8.2: Evaluative studies relating to the SF-36 when completed by patients with heart failure

Study/ Country	**Population (N) Age (years) Method of administration Setting**	**Measurement and Practical properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Alonso et al. (2004) 8-country survey of general adult population (IQOLA)	Different country population samples examining prevalence of chronic conditions and SF-36 scores (including HF) Age: mean 44.4 Postal *Descriptive data only presented*						
Arterburn et al. (2004) USA	Study of male veterans enrolled in general internal medical clinics (30,921) Age: mean by body mass index ranged from 57 to 66 Postal		Construct ✓				
Bayliss et al. (2004) USA	Medical Outcomes Study - longitudinal (1574 patients, including HF; n with HF unspecified) Age: mean 57.6 Self-administration			✓			
Birks et al. (2004) UK	Cardiac support group patients, including HF patients (396) Age: range 37-90, mean 67 Postal		Construct ✓				
Cunningham et al. (2003) USA	Patients (including HF) receiving care across 48 physician groups (5701) Age: 45% aged 50+ Postal		Construct ✓				

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Dempster et al. (2004) UK	IHD patients (mixed group, including 'other' unspecified) (117) Age: mean 60.61 Inpatients Interview		Construct ✓ Concurrent ✓				
Green et al. (2000) USA	Two patient cohorts with stable or decompensated CHF with LVEF < 40% (129) Age: mean 64.3 Outpatients Postal	Internal consistency ✓ Test-retest✓	Construct ✓ Concurrent ✓	✓			
Gwadry-Sridhar et al. (2005) Canada	Clinically diagnosed HF patients and LVEF < 40%, RCT educational intervention (134) Age: mean 67 intervention group, 65 control group Interview and telephone interview follow-up Inpatients		Construct ✓	✓		✓	
Havranek et al. (1999) USA	HF clinic patients (50) Age: mean 52.5 Outpatients Interview		Construct ✓				
Hobbs et al. (2002) UK	Patients in screening study of prevalence of HF and LV systolic dysfunction (5961) Age: 45+ Population sample Self-administration		Construct ✓				

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Jenkinson et al. (1997a) UK	HF patients (61) Age: range 60-92, mean 81 Outpatients Self-administration		Construct ✓	✓			
Jenkinson et al. (1997b) UK	Patients treated for HF (61), sleep apnoea, inguinal hernia Age: HF patients range 60-92, mean 82 Outpatients Self-completion			✓			
Lalonde et al. (1999) Canada	Outpatients with CHD (including HF), their ('healthy') friends and family, and hospital staff (878) Age: mean 55 Interviews	Test re-test ✓ Internal consistency ✓	Construct ✓ Concurrent ✓				
O'Leary and Jones (2000) UK	Patients with chronic LV dysfunction, including LVEF =/< 50% (60) Age: mean 60 Outpatients Self-administration		Construct validity ✓ Concurrent ✓				
Oldridge et al. (2002) USA	HF patients sampled from electronic medical records as having MI (161), angina or heart failure Age: MI patients mean 69 Postal		Construct ✓ Concurrent ✓				
Sidorov et al. (2003) USA	HF patients in disease management programme (268) Age: means 75.2 Outpatients, inpatients, community patients referred for discharge planning Self-administration		Construct ✓	✓			

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Sneed et al. (2001) USA	Patients attending HF clinic (30) Age: mean 57, range 35-88 Self-completed or interview if patient unable to read Outpatients		Construct ✓				✓
Sullivan et al. (2004) USA	Elderly outpatients with heart failure diagnosed in primary care and confirmed with 'chart review' (139, plus 80 spouses) Age: mean 75, 83% female Outpatients Self-administration		Construct ✓	✓			
Wolinsky et al. (1998) USA	Outpatients with CAD or chronic HF (560) Age: not given Telephone interview	Internal consistency ✓	Construct ✓		✓	✓	
Wyrwich et al. (1999) USA	Patients with a history of cardiac problems - CAD/CHF/both, participating in RCT of computerised medication reminders to physicians (605) Age not given Outpatients Interview			✓	✓		
Wyrwich et al. (2005) USA	Expert consensus panels of physicians (3); further details of numbers not given			✓			

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-12		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bennett et al. (2002) USA	HF clinic patients sampled from electronic medical records for diagnostic data (211) Age: 72% < 65 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓ Concurrent ✓		✓	✓	
Bennett et al. (2003) USA	Convenience sample of HF patients (211) Age: mean 57 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓	✓			✓
Conard et al. (2006) USA	HF patients, 13 centres, LVEF < 40% (539) Age: 59.5 burdened economically, 62.1 not burdened Outpatients Self-administration		Construct ✓				
Jenkinson et al. (1997b) UK	Three longitudinal datasets of patients treated for HF, sleep apnoea, and inguinal hernia (61 HF patients) Age: HF patients range 60-92, mean 82 Outpatients Self-completion			✓			
Jenkinson and Layte (1997) UK	Population survey, comparison of patient groups including HF (9332); construction of SF-12 summary scores Age: HF group mean 81, range 60-92 Postal		Construct ✓				

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
SF-12		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Lim and Fisher (1999) Australia	Population survey: AMI, IHD, 'other heart conditions' - not specified (1831) Age: 61% aged 65+ Postal	Internal consistency ✓	Construct ✓			✓	
Ni et al. (2000) USA	Clinic attendees with chronic and symptomatic HF (87) Age: 24% aged 60+ Outpatients Mode of administration: not given		Construct ✓	✓	✓		
Spertus et al. (2005) USA	Clinic attendees in 14 centres (476) Age: mean 61 Self-administration Outpatients		Construct ✓	✓			

c) Sickness Impact Profile (SIP)

Grady et al. (2003a), in a study of post-left ventricular assist device (LVAD) implantation and heart transplantation patients, reported that the SIP showed improvement for work and home management disability after heart transplantation. Mobility, self-care ability, physical ability and overall functional ability improved after LVAD implant and after heart transplant. Grady et al. (2003b), also reporting on LVAD implantation patients, found that functional disability measured with the SIP decreased post-discharge. Janz et al. (2004) used the emotional behaviour domain of the SIP, alongside generic domain specific measures, in an intervention trial of a disease management programme in 457 older women with heart disease (including HF). Women in the intervention arm were more likely to have improvements on the SIP, compared to controls. Avis et al. (1996) provided evidence supporting the construct validity and reliability of the SIP subscales for cognitive functioning, social functioning and productivity.

Table 8.3: Developmental and evaluation studies relating to the Sickness Impact Profile

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Sickness Impact Profile (SIP)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Avis et al. (1996) USA	Clinic patients and healthy patients (129 CVD patients) Age: mean 63 Interview *Selected SIP subscales only*	Internal consistency ✓ Test-retest ✓	Construct ✓			✓	
Grady et al. (2003a) USA and Australia	Post LVAD and heart transplant patients from medical centres in two countries (40) Age: mean 51.1 Inpatients Self-administration		Construct ✓	✓			
Grady et al. (2003b) USA and Australia	Post LVAD (62) Age: not discharged mean 52.8, discharged mean 50.2 Inpatients Self-administration		Construct ✓	✓			
Janz et al. (2004) USA	CHD patients, including heart failure, participating in psychological stress intervention trial Age: mean 73 (intervention), 72.1 (control); 100% female Outpatients and 'physician's offices' Telephone interviews (457) *Physical subscale only; Emotional behaviour subscale only used as outcome variable*		Construct ✓	✓			

d) EuroQoL/EQ-5D

There were five reports of the use of the EQ-5D in heart failure.

Reliability

No specific evidence was found.

Validity

The EQ-5D was found to correlate significantly with the MLHFQ, the NYHA functional status, and age-group, and has been reported to have higher response rates, reflecting its brevity (Calvert et al., 2005).

Responsiveness

Sullivan et al. (2004), in a study of older people with heart failure, reported that the EQ-5D thermometer scale was sensitive to independent measures of depression over time. The VAS scale of the EQ-5D, along with the KCCQ, was reportedly sensitive to variability in the health status of advanced HF (Hauptman et al., 2004). Spertus et al. (2005) found the EQ-5D and the SF-12 did not show much sensitivity to the magnitude of observed clinical change, unlike the KCCQ which demonstrated the highest sensitivity. Feldman et al. (2005) reported that EQ-5D scores improved for patients in a basic home health-care intervention arm, but not in the augmented intervention group, compared with controls. In contrast, the KCCQ mean summary scores improved for both intervention arms, compared with controls.

Precision

No specific evidence was found.

Acceptability

No specific evidence was found.

Feasibility

No specific evidence was found.

Table 8.4: Developmental and evaluation studies relating to the EQ-5D

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
EQ-5D		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Calvert et al. (2005) UK and 11 other countries	HF patients enrolled in study of cardiac re-synchronisation in HF (813) Age: mean 65 Self-administered		Construct ✓ Concurrent ✓			✓	
Feldman et al. (2005) USA	Patients with diagnosed HF in home health-care intervention trial (628) Age: 71.2 usual care, 72.4 basic care, 71.8 augmented care Home care patients Mode of administration not given		Construct ✓	✓			
Hauptman et al. (2004) USA	HF patients in multi-centre cohort study with LVEF < 40 (547) Age: mean 61 (advanced HF), 61 (non-advanced HF) Outpatients Mode of administration: not given		Construct ✓	✓			
Spertus et al. (2005) USA	Clinic attendees in 14 centres (476) Age: mean 61 Self-administration Outpatients		Construct ✓	✓			
Sullivan et al. (2004) USA	Elderly outpatients with heart failure diagnosed in primary care and confirmed with 'chart review' (139, plus 80 spouses) Age: mean 75, 83% female Outpatients Self-administration		Construct ✓	✓			

e) Other utility measures

Havranek et al. (1999) reported significant correlations between time-trade-off techniques and the MOS Rand SF-36 physical component score, the MLHFQ, a 6-minute walking test and a visual analogue scale rating health status. Utilities did not vary by age, sex or ethnicity of the patient. In a study of HF clinic patients, Havranek et al. (1999) reported significant correlations between the MOS Rand SF-36 physical component score, the MLHFQ, timed walking, a visual analogue scale rating health status, and time-trade-off techniques. Havranek et al. (2004) reported that the DASI correlated significantly with utility scores (time trade-off).

Kirsch and McGuire (2000) examined the feasibility of developing a QALY from the NYHA classification of heart failure, and concluded that constant proportionality did not hold across more severe health states, questioning the use of QALYs as representing cardinal preference structures. Lalonde et al. (1999) compared preference-based (rating scale, time trade-off, and standard gamble) and non-preference-based (SF-36) measures of HRQoL in CHD patients (including HF) and healthy people. While all measures were stable over 3-6 weeks, in contrast to SF-36 subscales, the utility measures were less able to discriminate between patients with various levels of disability. A large proportion of respondents also refused to return for the second interview, suggesting this battery of instruments (i.e. administered together) was not acceptable.

Table 8.5: Developmental and evaluation studies relating to Time Trade-off

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
TTO		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Havranek et al. (1999) USA	HF clinic patients (50) Age: mean 52.5 Outpatients Interview	Test-retest ✓	Construct ✓				
Havranek et al. (2004) USA and Canada	Patients in multi-site drug trial (153) Age: mean 68.3 In- or outpatients not specified Postal and telephone interview		Construct ✓				
Lalonde et al. (1999) Canada	Outpatients with CHD (including HF), their ('healthy') friends and family, and hospital staff (878) Age: mean 55 Interviews	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓			✓	

RESULTS: HEART FAILURE-SPECIFIC PATIENT-REPORTED HEALTH INSTRUMENTS

Four heart failure-specific instruments were identified which were evaluated with patients with various cardiovascular conditions resulting in heart failure. Full details of the development, domains and scoring methods are detailed in Tables 8.6 and 8.7.

Measurement properties are reported for the following instruments:

a) Chronic Heart Failure Questionnaire
b) Kansas City Cardiomyopathy Questionnaire
c) MacNew (ex-QLMI: Quality of Life after Myocardial Infarction Questionnaire)
d) Minnesota Living with Heart Failure Questionnaire

a) Chronic Heart Failure Questionnaire (CHQ) (Guyatt et al., 1989)

This 16-item instrument aims to measure subjective health status in heart failure patients, and is complex to administer as open-ended questions are used to yield score weights. It covers dyspnoea, fatigue, and emotional functions; it has a time recall period of two weeks. It was developed by presenting 123 items to a sample of 88 patients, who rated their importance. Item selection was based on frequency and importance ratings. A section of the CHQ is individualised, and patients are asked to nominate those activities associated with shortness of breath and that affect them most often/importantly. It requires a trained interviewer. Administration takes 10-20 minutes.

b) Kansas City Cardiomyopathy Questionnaire (KCCQ) Green et al., 2000)

This instrument aims to describe HRQoL over the previous two weeks in patients with congestive heart failure (CHF). It contains 23 items, covering physical function, clinical symptoms, social function, self-efficacy and knowledge and QoL ('enjoyment'), each with different Likert scaling wording, including limitations, frequency, bother, change in condition, understanding, levels of enjoyment and satisfaction. It is self-administered. A change of 5 points on the scale scores, either as a group mean or an intra-individual change is regarded as clinically important (Rumsfeld et al., 2003).

c) MacNew (ex-QLMI: Quality of Life after Myocardial Infarction Questionnaire) (Lim et al., 1993; Valenti et al., 1996)

While not solely heart failure-specific, MacNew measures HRQOL in heart disease (myocardial infarction, coronary disease and heart failure) in the previous two weeks. This instrument is a modification of the earlier Quality of Life after Myocardial Infarction (QLMI) Questionnaire, which had questionable validity (see review by Hofer et al., 2004). MacNew contains 27 items in three domains (Emotional, Physical, and Social). It takes up to 10 minutes to complete, and respondent burden is low.

d) Minnesota Living with Heart Failure Questionnaire (MLHF/MLHFQ/LHFQ/LiHFe) (Rector et al., 1987)

This contains 21 items that ask about patients' perceptions of the effects of heart failure and its treatment on physical, socioeconomic and psychological aspects of their life, rated on a 6-point Likert scale. Subscale scores for emotional and physical domains can be obtained. It is easy to administer by self-administration or interview. The items were drawn from the SIP. Patients with congestive heart failure were asked to select 21 items from the SIP, and these formed the MLHFQ. Some concern has been expressed about its content validity and whether all relevant items have been included (Dunderdale et al., 2005; O'Leary and Jones, 2000).

HEART FAILURE-SPECIFIC INSTRUMENTS:

Table 8.6: Details of Heart failure-specific patient-reported health instruments

Instrument	***Domains (no. items)***	***Response options***	***Score***	***Administration Completion time***
Chronic Heart Failure Questionnaire (CHQ) **(Guyatt et al., 1989)**	*16 items in 3 domains:* Dyspnoea (5) Fatigue (4) Emotional function (7) *Plus open-ended probes (3) for most important activities causing symptoms*	1-7 response scales of frequency or severity	Summed to yield subscale scores Weighting based on open-ended responses Minimum (worse function) to maximum (best function) scores in the 3 domains are: dyspnoea 5-35; fatigue 4-28; emotional 7-49	Interview 10-20 mins
Kansas City Cardiomyopathy Questionnaire (KCCQ) **(Green et al., 2000)**	*23 items in 5 domains* 1. Physical limitation (6) 2. Symptoms (8) 3. Self-efficacy and knowledge (2) 4. QoL/mood (3) 5. Social limitation (4)	6-point Likert scales, including severity and frequency	Summation of physical limitation, symptoms, social limitation and QoL domains. 0-100, higher scores represent fewer symptoms/better function/better QoL	Self-administered 4-6 mins
MacNew (ex-QLMI – Quality of Life after Myocardial Infarction) **(Lim et al., 1993)**	*23- 27 items in 3 overlapping domains:* Emotional Physical Social *In previous 2 weeks*	Item scores 1 = poor to 7 = high	Summation; domain scores calculated by taking the average of responses to items in each domain; averaging all items gives a global score.	Self-administered (modification of original interviewer-administered QLMI instrument) 5-10 minutes to complete
Minnesota Living with Heart Failure Questionnaire (MLHFQ) **(Rector et al., 1987)**	*21 items on impact of heart failure on:* Physical aspects of daily life (9) Emotional/psychological (5) Social/economic (7) *In previous 4 weeks.*	6-point Likert scales (0 = not at all, to 5 = very much)	Summation; range 0 (best) to 105 (worst QoL). Physical and emotional domains can also be summed.	Self-administered or interview

Table 8.7: Summary of heart failure-specific instruments: health status domains ***(after Fitzpatrick et al., 1998)***

Instrument	*Instrument domains*								
	Physical function	**Symptoms**	**Global judgement of health**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**	**Treatment satisfaction**
Chronic Heart Failure Questionnaire (CHQ)		x		x					
Kansas City Cardiomyopathy Questionnaire (KCCQ)	x	x		x	x		x	x	
MacNew (ex QLMI – Quality of Life after Myocardial Infarction)	x	x		x	x		x	x	
Minnesota Living with Heart Failure Questionnaire (MLHFQ)	x	x		x	x		x		

RESULTS: HEART FAILURE-SPECIFIC PATIENT-REPORTED HEALTH INSTRUMENTS

a) Chronic Heart Failure Questionnaire (CHQ)

The CHQ has been used with a wide range of HF patients, and age groups, including both sexes. Ten studies examined this instrument.

Reliability
Bennett et al. (2003) reported that the CHQ was more reliable, and more responsive to change in patients' condition, than the SF-12. It had comparable Cronbach's alphas as the MLHFQ (totals: 0.93 and 0.95, respectively; range for subscales 0.86-0.92 and 0.89-0.94, respectively). Bennett et al. (2002) supported the high internal consistency of the instrument in HF patients (all coefficients exceeding 0.70). The measure also had satisfactory reproducibility (Guyatt et al., 1989; O'Keefe et al., 1998).

Validity
Analyses have supported the convergent and discriminant validity of the instrument with HF patients, and the factor structure has been supported (Bennett et al., 2002). The measure can distinguish between patients receiving medication (digoxin) or placebo Jaeschke et al. (1990).

Socio-demographic variables
Age, sex, ethnic status, and income have been reported to be associated with higher scores on some of the subscales, in expected directions (Clark et al., 2003).

Heart Failure-specific patient-reported health instruments
The CHQ and LHFQ have been shown to be significantly and highly correlated (Bennett et al. 2002), and the CHQ and KCCQ have significant, but low to high, correlations (Clark et al., 2003).

Measures of Heart Function
Guyatt (1993) reviewed the psychometric properties of the measure during its development, including its significant correlations with change in walking test scores; the instrument's dyspnoea score also correlated significantly with heart failure scores. The instrument also correlates significantly with NHYA status in HF patients (Bennett et al., 2002). However, the measure was not found to be associated with left ventricular ejection fraction (LVEF) and co-morbidity in one cross-sectional study of HF patients (Clark et al., 2003).

Generic heath status
Wolinsky et al. (1998) tested the SF-36 and the CHQ (slightly adapted for use with CAD and HF patients) among outpatients with CAD or chronic HF. While the SF-36 was more comprehensive in its coverage of health status domains, the CHQ was more psychometrically sound, having fewer problems with floor and ceiling effects, and was more reproducible and internally consistent. The CHQ correlates moderately to highly, although significantly, with the SF-12 subscales (Bennett et al., 2002).

Responsiveness
Bennett et al. (2003) reported that the CHQ was more reliable, and more responsive to change in patients' condition than the SF-12. It correlates moderately highly, and significantly, with change in dyspnoea (0.65), change in walking test score (0.60) and change in heart failure scores (0.42) (Guyatt et al., 1989). Reviews by Guyatt (1993, 1994) of the measure's development also reported that the CHQ dyspnoea score was sensitive to improvements in patients' condition over time. O'Keeffe et al. (1998) found it was responsive to change at clinical re-assessment 3-8 weeks post-baseline assessment, and effect sizes for detecting deterioration were greater than those for detecting improvement.

Wyrwich et al. (1999) compared change in the SF-36 and CHQ, and examined standard errors in detail. They reported that both measures compared well at follow-up assessments of change. Jaesche et al. (1989) reported minimal clinically important differences. Wyrwich et al. (2005) used Delphi and consensus panel techniques with expert panels of physicians to examine clinically important differences for the SF-36 and a modified CHQ and reported on panel- derived thresholds for change over time.

Precision
The CHQ has fewer respondents at the floor and ceiling end of the scale than the MLHFQ (Bennett et al., 2002).

Acceptability
81% answered all CHQ questions in a study by Wolinsky et al. (1998).

Feasibility
No specific evidence was found.

Table 8.8: Developmental and evaluation studies relating to the Chronic Heart Failure Questionnaire (Guyatt et al., 1989)

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Chronic Heart Failure Questionnaire (CHQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bennett et al. (2003) USA	Convenience sample of HF patients (211) Age: mean 57 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓	✓			✓
Bennett et al. (2002) USA	HF clinic patients sampled from electronic medical records for diagnostic data (211) Age: 72% < 65 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓ Concurrent ✓		✓	✓	
Clark et al., (2003) USA	HF patients in medication adherence study ((212) Age: mean 63 Interview		Construct ✓ Concurrent ✓				
Guyatt et al. (1989) Canada	HF patients participating in drug trial (20) Age: 69.1 Interview In- or outpatients not specified	Test-retest ✓	Construct ✓	✓			
Jaeschke et al. (1989) Canada	HF patients from 3 studies (75) Age: not given In- or outpatients not specified Mode of administration: not given			✓			
Jaeschke et al. (1990) Canada	HF patients in drug trial (20) Age: not given Interview In- or outpatients not specified		Construct ✓				
O'Keeffe et al. (1988) UK	HF clinic patients (60) Age: mean 82 Outpatients Interview	Internal consistency ✓		✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Chronic Heart Failure Questionnaire (Guyatt et al., 1989)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Wolinsky et al. (1998) USA	Outpatients with CAD or chronic HF (560) Age: not given Telephone interview	Internal consistency ✓	Construct ✓		✓		
Wyrwich et al. (1999) USA	Patients with a history of cardiac problems – CAD/CHF/both, participating in RCT of computerised medication reminders to physicians (605) Age not given Outpatients Interview			✓	✓		
Wyrwich et al. (2005) USA	Expert consensus panels of physicians (3); further details of numbers not given			✓			

b). Kansas City Cardiomyopathy Questionnaire (KCCQ) (Green et al., 2000)

The KCCQ has been used with a wide range of heart failure patients. 13 studies examined this instrument.

Reliability
The development research for the instrument reported Cronbach's alphas for the sub-scales to range between 0.62 (self-efficacy) and 0.93 (functional status); the scale was shown to be reproducible at 3.3 months (mean duration of follow-up) (Green et al., 2000).

Validity
The development research for the KCCQ also indicated that the instrument had good validity overall (Green et al., 2000). Morgan et al. (2006) found that patients with difficulty taking medication had significantly worse HF symptoms, more social limitations, less self-efficacy and poorer QoL with the KCCQ, than patients with no difficulty taking their medications.

Socio-demographic variables
Age, sex, ethnic status, and income have been found to be associated with higher scores on some of the subscales, in expected directions (Clark et al., 2003).

Heart failure-specific patient-reported health instruments
Clark et al. (2003) reported correlations between the GHFQ and the KCCQ questionnaire of 0.16 to 0.37.

Measures of HF- Function
The QoL domain, the social limitation domain, functional status score and clinical summary score of the KCCQ all correlated significantly with NYHA class during the developmental testing of the measure (Green et al., 2000). However, Subramanian et al. (2005), in their longitudinal survey of older adults, reported only slight agreement between the instrument and clinician-reported NYHA functional classifications. The measure was not found to be associated with left ventricular ejection fraction or co-morbidity in a cross-sectional study of heart failure patients (Clark et al., 2003). It has also been reported that, while the KCCQ correlated significantly with the NHYA, it was not associated with B-type natriuretic peptide (BNPO) levels, regardless of the threshold used to define a clinically meaningful BNP change (Luther et al., 2005). Myers et al. (2006) found that only the QoL component of the KCCQ was significantly associated with peak VO_2; however, the physical limitation component and clinical summary score were significantly associated with 6 minute walk test.

Generic heath status
The QoL domain of the instrument correlated significantly but modestly ($r = 0.45$) with the general health perception scale of the Rand SF-36; it correlated more highly, and significantly, with the emotional domain of the MLHFQ ($r = 0.62$); the KCCQ social limitation domain was significantly correlated with the SF-36 social limitation scale ($r = 0.62$) (Green et al., 2000).

Responsiveness
Spertus et al. (2005) found that the KCCQ demonstrated the highest sensitivity to the magnitude of observed clinical change (cardiologists' assessments), compared with the SF-12, and the EQ-5D. Green et al. (2000) also reported it was more responsive to important clinical change in HF patients than the Rand SF-36 and the MLHFQ. The KCCQ, along with the VAS scale of the EQ-5D, was sensitive to variability in the health status of advanced heart failure (Hauptman et al., 2004). Rumsfeld et al. (2003) showed it was sensitive to changes in symptoms of depression in these patients. In a home health-care trial, mean KCCQ summary scores improved for both basic and augmented care intervention arms, compared with controls, while EuroQoL scores improved for patients in only the basic home health-care intervention arm (Feldman et al., 2005). In a study of older people with heart failure, Sullivan et al. (2004) reported that it was sensitive to independent measures of depression over time.

Precision
No specific evidence was found.

Acceptability
No specific evidence was found.

Feasibility
No specific evidence was found.

Table 8.9: Developmental and evaluation studies relating to the Kansas City Cardiomyopathy Questionnaire (Green et al., 2000)

Study/ Country	**Population (N)** **Age** **Method of administration** **Setting**	**Measurement properties**					
Kansas City Cardiomyopathy Questionnaire (KCCQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Clark et al. (2003) USA	HF patients in medication adherence study ((212) Age: mean 63 Interview		Construct ✓ Concurrent ✓				
Conard et al. (2006) USA	HF patients, 13 centres, LVEF < 40% (539) Age:59.5 (burdened economically), 62.1 (not burdened) Outpatients Self-administration		Construct ✓				
Feldman et al. (2005) USA	Patients with diagnosed HF in home health care intervention trial (628) Age: 71.2 usual care, 72.4 basic care, 71.8 augmented care Home care patients Mode of administration: not given		Construct ✓	✓			
Green et al. (2000) USA	Two patients cohorts with stable or decompensated CHF with LVEF < 40% (129) Age: mean 64.3 Outpatients Postal	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓	✓			
Hauptman et al. (2004) USA	HF patients in multi-centre cohort study with LVEF < 40% (547) Age: mean 61 advanced HF, 61 non-advanced HF Outpatients Mode of administration: not given			✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Kansas City Cardiomyopathy Questionnaire (KCCQ) (Green et al., 2000)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Luther et al. (2005) USA	Patients with systolic HF in 14 centres (342) Age: mean 60.4 Outpatients Mode of administration not given		Construct ✓				
Morgan et al. (2006) USA	HF patients with LVEF < 40% (522) Age: mean 58.1 (difficulty taking medications), 61.4 (no difficulty taking medications) Outpatients Self-administration		Construct ✓				
Rumsfeld et al. (2003) USA	HF patients 7with LVEF < 40% (460), in depressed and non-depressed groupings, multi-centre study Age: 57.3 and 62.6, respectively Outpatients Self-completion		Construct ✓	✓			
Myers et al. (2006) USA	HF patients (41) Age: mean 68 Outpatients Self-administration		Construct ✓				
Prasun et al. (2005) USA	Heart failure patients with LVEF =/< 40%, participating in RCT of patient-directed flexible diuretic protocol (66) Age: mean 65 intervention, 70 control group Outpatients Mode of administration: self-completion		Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Kansas City Cardiomyopathy Questionnaire (KCCQ) (Green et al., 2000)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Spertus et al. (2005) USA	Clinic attenders in 14 centres (476) Age: mean 61 Self-administration Outpatients		Construct ✓	✓			
Subramanian et al. (2005) USA	Longitudinal study of older adults (156 with complete 6-month follow-up data) Age: mean 63 Face-to-face or telephone interview		Construct ✓				
Sullivan et al. (2004) USA	Elderly outpatients with heart failure diagnosed in primary care and confirmed with 'chart review' (139, plus 80 spouses) Age: mean 75, 83% female Outpatients Self-administration		Construct ✓	✓			

c) MacNew (ex-QLMI - Quality of Life after Myocardial Infarction)

This instrument is a modification of the earlier Quality of Life after Myocardial Infarction (QLMI) Questionnaire. Just three studies were identified which used the MacNew.

Reliability
None of the included papers reported on reliability, although the instrument has been reviewed with favourable conclusions for its internal consistency by Höfer et al. (2004).

Validity
The review by Höfer et al. (2004) also reported favourable results for construct validity confirmation of the instrument's factor structure. However, Dempster et al. (2004) reported that a five-factor solution was more appropriate than the three factors reported for it.

HF-specific patient-reported health instruments
Oldridge et al. (2002) reported highly significant Pearson correlation coefficients between the MacNew, the Seattle Angina Questionnaire and the MLHFQ (ranging from 0.624 to 0.904).

Generic heath status
Oldridge et al. (2002) reported a highly significant Pearson correlation coefficient between the MacNew physical limitations domain and the SF-36 physical component summary at 0.63 for HF patients. The correlation between the SF-36 mental component summary and the MacNew was 0.70 for HF patients. Dempster et al. (2004) showed that the range of domain correlations between the MacNew and the SF-36 ranged from low to high ($r = 0.18$ to 0.85), although the highest were achieved for correlations between similar domains ($r = 0.52$ to 0.85).

Responsiveness
Dixon et al. (2002) reported the MacNew scores of HF patients to be significantly lower than those of other heart patients at four-month follow-up. Their change data suggested that a value of 0.5 may be a useful indicator of the minimal clinically important difference. The review by Höfer et al. (2004) also reported good results for responsiveness and sensitivity to changes post-intervention.

Precision
No specific evidence was found.

Acceptability
The review by Höfer et al. (2004) review reported favourable results for acceptability (high response rates).

Feasibility
No specific evidence was found.

Table 8.10: Developmental and evaluation studies relating to the MacNew (ex-QLMI - Quality of Life after Myocardial Infarction) instrument (Lim et al., 1993)

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
MacNew (ex-QLMI – Quality of Life after Myocardial Infarction)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Dempster et al. (2004) UK	IHD patients (mixed group, including 'other' unspecified) (117) Age: mean 60.61 Inpatients Interview		Construct ✓ Internal ✓				
Dixon et al. (2002) Australia	Discharged hospital patients with acute MI, HF, and angina taking part in longitudinal QoL study (1506) Age: mean 66.2 at baseline, 67.1 at follow-up Postal			✓			
Oldridge et al. (2002) USA	Study of HF patients sampled from electronic medical records as having MI (161) angina or heart failure Age: MI patients - mean 69 Postal		Construct ✓				

d) Minnesota Living with Heart Failure Questionnaire (MLHFQ/MLHFQ/LHFQ/LiHFe) (Rector et al., 1987)

This is the most popular heart failure-specific-instrument, and has been used with in- and outpatients of both sexes, with ages (where reported) ranging from 25-87 years. Items for the MLHFQ were drawn from the SIP. 28 studies which examined the MLHFQ were identified.

Reliability

Initial studies indicated good results for the reliability of the instrument (Rector et al., 1993a, 1993b). Cronbach's alphas are high, with studies reporting them to be between 0.80 and 0.94 (Gorkin et al., 1993; Heo, 2005). Bennett et al. (2003) reported that the MLHFQ was more reliable than the SF-12, and was comparable with the CHQ (Cronbach's alpha totals: 0.95 and 0.93, respectively; range for subscales 0.89-0.94 and 0.69-0.92, respectively, for the MLHFQ and the CHQ). Bennett et al. (2002) and O'Leary and Jones (2000) supported the high internal consistency of the instrument. Test-retest reliability is good, with correlation coefficients between r = 0.87 and 0.93 (Rector and Cohn, 1992; Rector et al., 1993b).

Validity

Early studies by Rector et al. (1987) found that the instrument correlated highly with patients' global assessments of restrictions on their lives (r = 0.80). A review of medication trials of beta blockers with heart failure patients by Reddy and Dunn

(2000) [not shown in Table, as review data], reported inconsistent results for the effect of beta-blockers on MLHFQ scores. Item and factor analyses indicate that some items need to be removed and others reworded (Heo, 2005).

Socio-demographic variables
This instrument has been used with a wide range of patient groups, with documented ages ranging from 28-87 years. The MLHFQ is apparently sensitive to age, independently of symptoms (Rector et al., 2006), although O'Leary and Jones (2000) reported no significant associations between the MLHFQ and age or sex.

Heart failure-specific patient-reported health instruments
Oldridge et al. (2002) reported highly significant Pearson correlation coefficients between the MLHFQ, the MacNew and the Seattle Angina Questionnaire (ranging from 0.624 to 0.904). The CHQ and MLHFQ are also highly correlated (Bennett et al., 2002).

Measures of HF Function
Havranek et al. (1999) reported significant correlations between the MLHFQ and a 6-minute walking test. O'Leary and Jones (2000) found moderate, significant correlations between MLHFQ scores and exercise capacity (VO_2 max) and duration (0.49 and 0.38, respectively). No significant association was found with echocardiographic measurements; significant associations were found with NYHA classes, though not between classes II and IV.

Zambroski et al. (2005) reported that the MLHFQ was sensitive to symptom prevalence and burden, and NYHA functional classification. The latter finding was supported by Rector et al. (1987, 2006; Gorkin et al., 1993; Calvert et al., 2005). Bennett et al. (2002) found that only the MLHFQ physical subscale differentiated between patients with NYHA class III and IV, although it discriminated between the other classes. A secondary analysis of a 'convenience sample' of nine experimental or quasi-experimental studies in the USA [not shown in Table, as secondary review analysis] showed mixed results for associations between the MLHFQ and NYHA classes, and it was unable to discriminate between LVEF values (Riegel et al., 2002). The MLHFQ was also reported to be insensitive to clinical indicators of cardiac function and symptoms in a study of outpatients by Carels (2004).

Generic heath status
In a study of heart failure clinic patients, Havranek et al. (1999) reported significant correlations between the MLHFQ and the MOS Rand SF-36 physical component score and time-trade-off techniques. Oldridge et al. (2002) also found highly significant Pearson correlation coefficient between the MLHFQ physical limitations domain and the SF-36 physical component summary at 0.63; the correlation between the SF-36 mental component summary and the MLHFQ was 0.72. O'Leary and Jones (2000) reported moderate to high significant correlations between the MLHFQ and all eight Rand MOS SF-36 domains ($r = -0.46$ to -0.75). The MLHF also correlates significantly with the EQ-5D (Calvert et al., 2005).

Responsiveness
While the MLHFQ has been reported to be sensitive to change in patients' condition over time (Aranda et al., 2004; Gary et al., 2004b; Prasun et al., 2005; Park et al.,

2005; Rector and Cohn, 1992; Rector et al., 1993a, 1993b), others have reported that the MLHFQ is not sensitive to trial interventions (Feldman et al., 2004). Not all investigators used independent measures of change in patients' condition, and it is possible that, in some cases, lack of change could be due to the insensitivity of the MLHFQ. Doughty et al. (2002) found that only the physical dimension was sensitive to heart failure management interventions, compared with controls.

The instrument is reportedly responsive to changes due to exercise therapy (Chang et al., 2005; Gary et al., 2004a). Gwadry-Sridhar et al. (2005), in an RCT of an educational intervention with HF inpatients, reported a significant effect of both time and treatment intervention with the MLHFQ, but not with the SF-36. Bennett et al. (2003) also found that the MLHFQ performed better than the SF-12 with regard to responsiveness to change. Ni et al. (2000) found the SF-12 to be more responsive to changes in mental health, but less responsive to change in physical health at follow-up, than the MLHFQ; the MLHFQ also performed better than the SF-12 in ability to distinguish differences in perceived global health transition. Green et al. (2000) reported that the MLHFQ was less responsive to important clinical change in HF patients than the KCCQ.

Sethares and Elliott (2004), in their RCT of tailored message intervention, assessed heart failure patients in hospital and post-discharge. For both treatment and control groups, there were significant differences in their MLHFQ scores between baseline and follow-up assessments, indicating improved QoL. Although no clinical evidence of improvement was provided, this could suggest some support for the instrument's responsiveness to expected clinical improvement post-discharge. However, there were no differences detected between groups, which indicated either that the intervention had no effect or that the measure was insensitive. Rector and Cohn (1992) reported that changes in total and physical MLHFQ scores were significantly, but weakly, associated with changes treadmill exercise tests at follow-up, but more strongly associated with patients' own assessments of changes in dyspnoea and fatigue.

Precision
Results for floor and ceiling effects are mixed. While Bennett et al. (2002) found that the MLHFQ had more respondents at the floor and ceiling end of the scale than the CHQ, Ni et al. (2000) reported that no respondents had the highest or lowest possible scores, excluding an obvious floor or ceiling effect. O'Leary and Jones (2000) reported that the instrument had no floor effects, and a very small ceiling effect (4% of respondents scored at the ceiling).

Acceptability
Two MLHFQ items were reported to be missing for large numbers of respondents in one study: difficulty working to earn a living (27%) and difficulty with sexual activities (22%) (Bennett et al., 2002). In a study by Gwadry-Sridhar et al. (2005), 12 out of 134 patients reported finding the questionnaire battery (which included the MLHFQ and SF-36) cumbersome and did not respond.

Feasibility
No specific evidence was found.

Table 8.11: Developmental and evaluation studies relating to the Minnesota Living with Heart Failure Questionnaire (Rector et al., 1987)

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Minnesota Living with Heart Failure Questionnaire (MLHFQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Aranda et al. (2004) USA	HF patients participating in trial (313) Age: means 63.5, 66.3, 66.8 in different treatment groups In- or outpatients not specified Mode of administration not given		Construct ✓	✓			
Arena et al. (2002a) USA	Patients with compensated HF (31) Age: mean 52.8 In- or outpatients not specified Mode of administration not given		Construct ✓				
Bennett et al. (2002) USA	HF clinic patients sampled from electronic medical records for diagnostic data (211) Age: 72% < 65 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓ Concurrent ✓		✓	✓	
Bennett et al. (2003) USA	Convenience sample of HF patients (211) Age: mean 57 Outpatients Telephone interviews	Internal consistency ✓	Construct ✓	✓			✓
Calvert et al. (2005) UK + 11 other countries	HF patients enrolled in cardiac resynchronisation in HF study (813) Age: mean 65 Self-administered		Construct ✓ Concurrent ✓				
Carels (2004) USA	HF clinic patients (58) Age: mean 67.7 Outpatients Self-administered		Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Minnesota Living with Heart Failure Questionnaire (MLHFQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Doughty et al. (2002) New Zealand	HF patients participating in RCT of HF management (197) Age: 72.5 (intervention), 73.5 (control) Post-discharge patients Mode of administration not given			✓			
Chang et al. (2005) USA	HF patients trial of relaxation therapy, LVEF =/< 40% (95) Age: means 69.7, 68.7 and 69.2 in different study arms Outpatients/primary care patients Self-administered and postal		Construct ✓	✓			
Feldman et al. (2004) USA	Women with LVEF < 45% plus symptoms (32) in exercise RCT Age: mean 67 (intervention group), 69 (control group) Mode of administration at baseline not specified; telephone interview follow-up		Construct ✓	✓			
Gary et al. (2004b) USA	Exercise trial, females with heart failure, LVEF > 45%, symptoms, NYHA class II & III (32) Age: 67 (intervention group), 69 (control group) Heart clinic and local practice patients Mode of administration at baseline not given; telephone follow-up		Construct ✓	✓			
Gorkin et al. (1993) USA	Patients enrolled in quality of life study (158). White males only with LVEF =/< 35%, and no MI within previous 30 days. Age: mean 59.6 (NYHA Class I), 61.9 (NYHA Class II or III) Mode of administration not given, battery administered during assessment visit.	Internal consistency ✓	Construct ✓				

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Minnesota Living with Heart Failure Questionnaire (MLHFQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Green et al. (2000) USA	Two patient cohorts with stable or decompensated CHF with LVEF < 40 (129) Age: mean 64.3 Outpatients Postal	Internal consistency ✓ Test-retest✓	Construct ✓	✓			
Gwadry-Sridhar et al. (2005) Canada	Clinically diagnosed HF patients and LVEF < 40%, RCT educational intervention (134) Age: mean 67 (intervention group), 65 (control group) Inpatients Interview with telephone interview follow-up		Construct ✓	✓			
Havranek et al. (1999) USA	HF clinic patients (50) Age: mean 52.5 Outpatients Interviews		Construct ✓ Concurrent ✓				
Heo (2005) USA	HF patients (638) enrolled in 4 separate studies Age: not given Inpatients Mode of administration: not given	Internal consistency ✓	Construct ✓				
Ni et al. (2000) USA	Clinic attendees with chronic and symptomatic HF (87) Age: 24% aged 60+ Outpatients Mode of administration: not given		Construct ✓	✓	✓		
O'Leary and Jones (2006) UK	Cardiac clinic patients with chronic LV dysfunction, including LVEF =/< 50% (60) Age: mean 60 Outpatients Self-administration	Internal consistency ✓ Test re-test ✓	Construct ✓ Concurrent ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Minnesota Living with Heart Failure Questionnaire (MLHFQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Oldridge et al. (2002) USA	Study of HF patients sampled from electronic medical records as having MI (161), angina or heart failure Age: mean 69 (MI patients) Postal		Construct ✓ Concurrent✓				
Park et al. (2005) USA	RCT of LVADs with end stage HF patients NYHA class IV, symptoms 90 days plus (129) Age: mean 68 (medical management group), 66 (LVAD group) In- or outpatients not specified Mode of administration not specified		Construct ✓	✓			
Rector et al. (1987) USA	Patients with LV dysfunction participating in several studies (83; 84% males) Age: mean 61 In- or outpatients not specified Self-administration	Internal consistency ✓	Construct ✓				
Rector and Cohn (1992) USA	HF patients enrolled in multi-site drug trial (198; 78% males) Age: mean 58 In- or outpatients not specified Self-administration	Test re-test ✓	Construct ✓	✓			
Rector et al. (1993a) USA	Patients in preventative drug trial without symptoms of HF (172) and patients with HF (77; 86% males) Age: mean 63 In- or outpatients not specified Mode of administration not given	Test re-test ✓	Construct ✓	✓			
Rector et al. (1993b) USA	Patients enrolled in multi-site veterans centre drug trial (804) Age: mean 61 Ambulatory patients Self-administration	Test re-test ✓	Construct ✓	✓			

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Minnesota Living with Heart Failure Questionnaire (MLHFQ)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Rector et al. (1995) USA	Clinic patients with HF (101) Age: range 50-75, mean 56 Interview		Construct ✓				
Rector et al. (2006) USA	White, male patients enrolled in heart failure drug trial (1651; 77% males) Age: median 62 Outpatients Mode of administration: not specified		Construct ✓				
Sethares and Elliott (2004) USA	Primary diagnosis of chronic HF (70) RCT of tailored message intervention Inpatients and post-discharge follow-up Age: mean 75.70 (treatment group), 76.84 (control group) Interview		Construct ✓	✓			
Sneed et al. (2001) USA	Patients attending HF clinic (30) Age: mean 57 Outpatients Self-administered or interview if patient unable to read		Construct ✓				
Zambroski et al. (2005) USA	Convenience sample HF clinic patients (53) Age: mean 55.5 Self-administered		Construct ✓				

Other heart failure-specific instruments identified from the review

Table 8.12: Overview of other records of heart failure-specific instruments and generic CVD instruments used with heart failure patients.

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Heart failure-specific instruments								
Heart Failure Symptom Checklist **Grady et al. (2003a)** **USA and Australia**	Post LVAD implantation and heart transplantation patients from medical centres in 2 countries (40) Age: mean 51.1 Inpatients Self-administration		Construct ✓	✓				90 post-LVAD (89 post-transplant) items covering: somatic sensation, psychological state. 0 = not bothered to 3 = very bothered. 6 subscales: cardiopulmonary, gastrointestinal, genitourinary, neurological, dermatological, and psychological. Sensitive to change post LVAD
Grady et al. (2003b) **USA and Australia**	Post LVAD implantation patients from medical centres in 2 countries (62) Age: mean 52.8 (not discharged), 50.2 (discharged) Inpatients Self-administration		Construct ✓	✓				HFSC not sensitive to change post-discharge LVAD patients, in contrast to changes detected by functional disability (SIP) stress (LVAD stressor scale), coping (Jalowiec Coping Scale), global ratings, and the Quality of Life Index.
Heart Failure Symptom Scale (HFSS) **Baker et al. (2005)** **USA**	HF patients from 7 sites: hospital clinics, health plan and physician groups (781) Age: 62% aged 65+ Telephone interview	Internal consistency ✓	Construct ✓ Concurrent ✓					7 items on symptom severity and frequency (5-point response scales) HFSS correlated with SF-12 PCS (0.63), and MCS (0.54). Single factor reported, Cronbach's alpha 0.88

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Left ventricular dysfunction questionnaire (LVD-36) **O'Leary and Jones (2000)** **UK**	Cardiac clinic patients with chronic LVD, including LVEF =/< 50% (60) Age: mean 60 Outpatients Self-administration	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓	✓				Newly developed measure; takes 5 minutes to complete 36-item questionnaire, responses are dichotomous and summed, and expressed as a percentage so 100 = worst possible score and 0 = best possible score Tested against the Rand SF-36 and MLHFQ. High repeatability and internal consistency; no floor effects and very small ceiling effects; correlations with the SF-36 ranged from 0.46-0.75, and with the MLHFQ from 0.41 to 0.74; correlations with clinical tests were weak or non-significant, and moderate at best with exercise test; there was evidence of responsiveness to change in health status at follow-up
Memorial Symptom Assessment Scale–Heart Failure **Zambroski et al. (2005)** **USA**	Convenience sample HF clinic patients (53) Age: mean 55.5 Self-administered		Construct ✓					32 items on 3 symptom subscales: physical, emotional, HF. Summed to give Prevalence; Burden calculated as mean of frequency, severity and distress of each on 4- and 5-point scales Correlated with NYHA functional class, predicted worse Qol (MLHFQ)

Generic cardiovascular measures used with HF patients								
Instrument **Reference** **Country**	**Population (N)** **Age** **Method of administration** **Setting**	**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**	**Comments** **No other records identified unless stated**
Cardiac Depression Scale (CDS) **Hare and Davis (1996)** **Birks et al. (2004)** **UK**	Cardiac support group patients (396) Age: range 37-90, mean 67 Postal	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓					26 item, self-administration, disease-specific depression scale, originally developed in Australia. In the UK population Cronbach's alphas 0.91 and 0.86. Two domains highly correlated (0.649). Test-retest 0.79. CDS correlated significantly with SF-36, Beck Depression Inventory & Hospital Anxiety and Depression Scale.
Cardiac Quality of Life Index ***scale developers and modifiers:*** **Padilla and Grant (1983, 1985); Rukholm and McGirr (1994); Rukholm et al. (1998)** **USA**	Cardiac patients (222) Age: range 32-65+ Interview-administered	Test-retest ✓	Construct ✓	✓				A generic cardiovascular instrument; 30 items. Domains: psychological state, physical and occupational function, social interaction: 0 = very dissatisfied to 100 = very satisfied High levels of test re-test reliability. Discriminated healthy and cardiac patients. Strong correlation with Spitzer QLI

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Duke Activity Status Index (DASI) **Hlatky et al. (1989)** **Arena et al. (2002b)** **USA**	Patients with 'past medical history significant for HF' (33) Age: mean 52.8 Self-administration In- or outpatients not specified	Test-rest ✓	Construct ✓					A generic cardiovascular tool, aiming to measure functional capacity and QoL. It contains 12 items on functional capability: personal care, ambulation, household tasks, sexual function and recreational activities. Yes/no response formats for ability. Each item has a weighted value of 1.75-8.0; the DASI is the sum of these: 0 = worst, 58.2 = best
Gary et al. (2004a) **USA**	Patients with diagnosed diastolic HF or diastolic dysfunction participating in exercise trial Age: mean 67 (intervention), 69 (controls); range 51 – 86 In- or outpatients not specified Mode of administration not given		Construct ✓	✓				

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
(DASI continued) **Gary et al. (2004b)** **USA**	Exercise trial, females with heart failure, LVEF 45% or over, symptoms, NYHA class II & III (32) Age: 67 (intervention group), 69 (control group) Heart clinic and local practice patients Mode of administration at baseline not given; telephone follow-up							Results for DASI not clearly presented
Havranek et al. (2004) **USA and Canada**	Patients in multi-site drug trial (153) Age: mean 68.3 In- or outpatients not specified Postal and telephone interview		Construct ✓					
Myers et al. (2006*)* **USA**	HF patients (41) Age: mean 68 Outpatients Self-administration		Construct ✓ Concurrent ✓					

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Multidimensional Index of Life Quality (MILQ) **Avis et al. (1996)** **USA**	Clinic patients and healthy patients (129 CVD patients) Age: mean 63 Interview	Internal consistency ✓ Test-retest ✓	Construct ✓ Concurrent ✓		✓			Qualitative interviews with CVD patients and healthy people identified 9 domains: mental health, physical health, physical functioning, cognitive functioning, social functioning, intimacy, productivity, financial status, relationship with health professionals. Initial tests on hospital and clinic patients. 7-point satisfaction and 4-point importance response scales. High Cronbach's alphas, exceeding 0.70. Good test-retest (0.73 or higher). Correlated significantly with scales of depression, anxiety, MOS indicators of mental health, physical functioning, health perceptions and SIP domains of mobility, social interaction, work.
NYHA Functional Classification (*The Criteria Committee of NYHA, 1994)* **Subramanian et al. (2005)** **USA**	Longitudinal study of older adults (156 with complete 6-month follow-up data) Age: mean 63 Face-to-face or telephone interview		Construct ✓					Most widely used and well-tested of the cardiovascular classification scales. There is a large general CVD literature. Re. HF-specific measures: associated with MSAS-HF, but only slight agreement between patient-based KCCQ and clinician-reported NYHA functional class.

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Quality of Life in Severe Heart Failure (QLF-SHF) **Wiklund et al. (1987)** **Reddy and Dunn (2000)** **USA**	Review of medication trial data and comparison with MLHFQ			✓				27 items in 4 domains, designed for self-completion. Includes physical activity (7), somatic symptoms (7), psychological (7), life satisfaction (5), improvement rating (1), all on 6-point Likert scales. Domain scores summed for global score. Inconsistent results. Lack of consistent effect of beta blockers on QoL of HF patients. Little use of the scale in the literature.
Quality of Life Index-Cardiac version (Ferrans and Powers 1985) **Grady et al. (2003a, 2003b)** **USA and Australia**	Post LVAD implantation and heart transplantation patients from medical centres in 2 countries (40) Age: mean 51.1 Inpatients Self-administration		Construct ✓	✓				Greater satisfaction (QLI) at three months noted for heart transplantation versus LVAD implantation
Prasun et al. (2005) **USA**	Heart failure patients with LVEF =/< 40%, participating in RCT of patient-directed flexible diuretic protocol (66) Age: mean 65 (intervention group), 70 (control group) Outpatients Mode of administration: self-completion		Construct ✓	✓				Generic cardiovascular instrument. 30 items; domains: psychological state, physical and occupational function, social interaction: 1 = very dissatisfied to 6 = very satisfied

Instrument Reference Country	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Utility Based Quality of Life-Heart (UBQ-H) Questionnaire (*cardiovascular extension to the Health Measurement Questionnaire*) **Martin et al. (1999)** **Australia**	Cardiac patients (including HF) (322) Age: mean 60 Outpatients Postal	Internal consistency ✓ Test-retest ✓	Construct ✓	✓				UBQ-H: psychological distress 16 items (response categories 0-10) : self-care 4 items (1-3); social/usual activities 5 items (1-4); physical ability 4 items (1-4.5); overall QoL (1-4/0.0-1.0//0-100) (3 items); Rosser Index (1-7/1-4/-1.486-1) (3 items) Within the QoL domain, this measure includes self-rated health, a time trade-off rating scale, anchored by full health and death, a rating scale; the Rosser Index is a separate domain. Response and item-completion high; high Cronbach's alpha: 0.79-0.91 and test-retest: 0.65-0.82. Correlated significantly with the General Health Questionnaire-30 (symptoms of general psychological morbidity, mainly depression/anxiety).

NB Insufficient information was found on the following:

- 19-item Cardiac Health Profile
- 11-item Heart Condition Assessment Questionnaire
- generic 35-item Multidimensional Index of Quality of Life
- Quality of Life at the End of Life measure/QUAL-E (this is a palliative care instrument, and not heart failure-specific)

SUMMARY - GENERIC INSTRUMENTS

The most frequently used generic measure in patients with heart failure is the SF-36. 21 studies assessed the SF-36, and eight examined the SF-12, in relation to adults with heart failure. Most of these reported construct validity. The studies were based on population surveys and clinical samples of patients. There is less published information about reliability, compared with validity, when used with this population. Of those that examined reliability, internal consistency was good for all the SF-36 scales, except for social functioning (Wolinsky et al., 1998; Green et al., 2000).

Evidence was more mixed for the discriminative ability of the SF-36 in relation to different diagnostic groups of CVD patients. The instrument correlated well with disease-specific measures, including the MacNew and the MLHFQ. The SF-36 was found by some investigators to be less responsive to important clinical change in HF patients than disease-specific instruments such as the MLHFQ (Green et al., 2000). Floor and/or ceiling effects have been found for the SF-36 role-physical, social functioning and role-emotional subscales, which potentially mask patient improvement or deterioration, and reduce scale sensitivity (Wolinsky et al., 1998; Wyrwich et al., 1999).

The internal consistency of the SF-12 was also good reportedly good, although Bennett et al. (2003) reported that the SF-12 was less reliable than the CHQ or the MLHFQ. Results for validity were generally good. Correlations between the SF-12 and the CHQ and MLHFQ were moderate to high, although results for responsiveness were more mixed. No floor or ceiling effects were reported.

Other generic measures used in heart failure, but in a smaller number of studies, included the SIP, which had generally good results for validity. The EuroQoL/EQ-5D was used in a very small number of studies, and with mixed results for responsiveness, although it correlated significantly with disease-specific indicators.

Overall, the SF-36 is the most widely used and well-tested generic instrument for use with patients with heart failure, although, unlike the shorter SF-12, it has floor and ceiling effects. More evidence of the reliability of these instruments is needed when used with this group of patients.

SUMMARY – HEART FAILURE-SPECIFIC INSTRUMENTS

While several general cardiovascular disease instruments were identified, four main heart failure-specific instruments were identified: three papers examined the MacNew (ex-QLMI – Quality of Life after Myocardial Infarction) (Lim et al., 1993); ten papers examined the Chronic Heart Failure Questionnaire (CHQ) (Guyatt et al., 1989); 13 examined the Kansas City Cardiomyopathy Questionnaire (KCCQ) (Green et al., 2000); and most, 28, examined the Minnesota Living with Heart Failure Questionnaire (MLHFQ) (Rector et al., 1987). The Duke Activity Status Index (DASI - see Table 8.12) (Hlatky et al., 1989) is a short but narrow measure of activity performance. It does have some attractive features with promising measurement properties.

The MLHFQ, followed by the KCCQ, then, was the most commonly used and tested disease-specific instrument. The psychometric properties of the KCCQ were generally

good but more mixed than those for the MLHFQ. The KCCQ was more responsive to clinical change than generic measures.

The MLHFQ has been shown to have high internal consistency, and satisfactory results for test-retest reliability. It correlates highly with other heart failure-specific measures (MacNew, CHQ) (Bennett et al., 2002; Oldridge et al., 2002), and most studies supported its sensitivity to NYHA classes, and responsiveness to change. However, not all results were consistent or good. Some items on the MLHFQ have lower item-response: difficulty with sexual activities, difficulty with recreational pastimes, sports or hobbies (Bennett et al., 2002), indicating that further revision is needed.

DISCUSSION AND RECOMMENDATIONS

This review was based on generic and heart failure-specific instruments. Most studies were carried out in North America. Johansson et al. (2004), from their systematic review, identified 32 different generic, disease-specific and domain-specific HRQoL questionnaires used in 33 articles on HF. However, the number of commonly used generic and heart failure-specific measures is relatively small, comprising mainly the SF-36; the SF-12; the Chronic Heart Failure Questionnaire (CHQ); the Kansas City Cardiomyopathy Questionnaire (KCCQ); the MacNew (ex -QLMI – Quality of Life after Myocardial Infarction); and the Minnesota Living with Heart Failure Questionnaire (MLHFQ). The most popularly used measure is the MLHFQ. There was relatively little use of utility measures.

The generic instruments reviewed were multidimensional indicators of health status or health-related quality of life. Acceptability was examined only for the SF-12, and no study examined feasibility. For the disease-specific instruments, feasibility was not assessed. While the MLHFQ is the most popular and well-tested disease-specific instrument, two MLHFQ items were reported to be missing for large numbers of respondents in one study: difficulty working to earn a living (27%) and difficulty with sexual activities (22%) (Bennett et al., 2002).

The SF-36 and SF-12 have been tested against heart failure-specific questionnaires, with overall moderate to high significant correlations, although with some inconsistency between studies. The heart failure-specific instruments have been tested for concurrent validity, and most results show that the different instruments correlate highly, except between the GHFQ and the KCCQ where correlations were weak to modest. There was relatively little information on the distribution of scale items, or variation, by socio-demographic characteristics.

Recommendations

This review supports the use of the SF-36 and the SF-12 as generic instruments, and the MLHFQ as a heart failure-specific instrument with people with heart failure. However, there was some indication that the wording of some of the MLHFQ items needs revision and retesting, and the narrow scope of the instrument suggests that it may not have full content validity. The development and use of broader, generic and disease-specific patient-based indicators has been slow in the field of cardiovascular disease compared to other areas (e.g. respiratory conditions), and this is reflected in the range and development of instruments included here.

REFERENCES

Alonso J, Ferrer M, Gandek B, Ware JEJ, Aaronson NK, Mosconi P *et al*. Health-related quality of life associated with chronic conditions in eight countries: results from the International Quality of Life Assessment (IQOLA) Project. *Quality of Life Research* 2004;**13**:283-98.

Aranda JMJ, Conti JB, Johnson JW, Petersen SS, Curtis AB. Cardiac resynchronization therapy in patients with heart failure and conduction abnormalities other than left bundle-branch block: analysis of the Multicenter InSync Randomized Clinical Evaluation (MIRACLE). *Clinical Cardiology* 2004;**27**:678-82.

Arena R, Humphrey R, Peberdy MA. Relationship between the Minnesota Living with Heart Failure Questionnaire and key ventilatory expired gas measures during exercise testing in patients with heart failure. *Journal of Cardiopulmonary Rehabilitation* 2002a;**22**:273-7.

Arena R, Humphrey R, Peberdy MA. Using the Duke Activity Status Index in heart failure. *Journal of Cardiopulmonary Rehabilitation* 2002b;**22**:93-5.

Arterburn DE, McDonell MB, Hedrick SC, Diehr P, Fihn SD. Association of body weight with condition-specific quality of life in male veterans. *American Journal of Medicine* 2004;**117**:738-46.

Avis NE, Smith KW, Hambleton RK, Feldman HA, Selwyn A, Jacobs A. Development of the Multidimensional Index of Life Quality: a quality of life measure for cardiovascular disease. *Medical Care* 1996;**34**:1102-20.

Baker DW, Brown J, Chan KS, Dracup KA, Keeler EB. A telephone survey to measure communication, education, self-management, and health status for patients with heart failure: the Improving Chronic Illness Care Evaluation (ICICE). *Journal of Cardiac Failure* 2005;**11**:36-42.

Bayliss EA, Bayliss MS, Ware JEJ, Steiner JF. Predicting declines in physical function in persons with multiple chronic medical conditions: what we can learn from the medical problem list. *Health and Quality of Life Outcomes* 2004;**2**:47.

Bennett SJ, Oldridge NB, Eckert GJ, Embree JL, Browning S, Hou N *et al*. Discriminant properties of commonly used quality of life measures in heart failure. *Quality of Life Research* 2002;**11**:349-59.

Bennett SJ, Oldridge NB, Eckert GJ, Embree JL, Browning S, Hou N *et al*. Comparison of quality of life measures in heart failure. *Nursing Research* 2003;**52**:207-16.

Birks Y, Roebuck A, Thompson DR. A validation study of the Cardiac Depression Scale/CDS in a UK population. *British Journal of Health Psychology* 2004;**9**:15-24.

Calvert MJ, Freemantle N, Cleland JGF. The impact of chronic heart failure on health-related quality of life data acquired in the baseline phase of the CARE-HF study. *European Journal of Heart Failure* 2005;**7**:243-51.

Carels RA. The association between disease severity, functional status, depression and daily quality of life in congestive heart failure patients. *Quality of Life Research* 2004;**13**:63-72.

Chang B, Hendricks A, Zhao Y, Rothendler JA, LoCastro JS, Slawsky MT. A relaxation response randomized trial on patients with chronic heart failure. *Journal of Cardiopulmonary Rehabilitation* 2005;**25**:149-57.

Clark DO, Tu W, Weiner M, Murray MD. Correlates of health-related quality of life among lower-income, urban adults with congestive heart failure. *Heart and Lung* 2003;**32**:391-401.

Conard MW, Heidenreich P, Rumsfeld JS, Weintraub WS, Spertus J. Patient-reported economic burden and the health status of heart failure patients. *Journal of Cardiac Failure* 2006;**12**:369-74.

Cunningham WE, Nakazono TT, Tsai KL, Hays RD. Do differences in methods for constructing SF-36 physical and mental health summary measures change their associations with chronic medical conditions and utilization? *Quality of Life Research* 2003;**12**:1029-35.

Dempster M, Donnelly M, O'Loughlin C. The validity of the MacNew Quality of Life in heart disease questionnaire. *Health and Quality of Life Outcomes* 2004;**2**:6.

Dixon T, Lim LLY, Oldridge NB. The MacNew Heart Disease Health-related Quality of life instrument: reference data for users. *Quality of Life Research* 2002;**11**:173-83.

Doughty RN, Wright SP, Pearl A, Walsh HJ, Muncaster S, Whalley GA *et al*. Randomized, controlled trial of integrated heart failure management: The Auckland Heart Failure Management Study. *European Heart Journal* 2002;**23**:139-46.

Dunderdale K, Thompson DR, Miles JNV, Beer SF, Furze G. Quality-of-life measurement in chronic heart failure: do we take account of the patient perspective? *European Journal of Heart Failure* 2005;**7**:572-82.

Feldman PH, Peng TR, Murtaugh CM, Kelleher C, Donelson SM, McCann ME *et al*. A randomized intervention to improve heart failure outcomes in community-based home health care. *Home Health-Care Services Quarterly* 2004;**23**:1-23.

Feldman PH, Murtaugh CM, Pezzin LE, McDonald MV, Peng TR. Just-in-time evidence-based e-mail 'reminders' in home health-care: impact on patient outcomes. *Health Services Research* 2005;**40**:865-85.

Ferrans CE, Powers MJ. Quality of Life Index: development and psychometric properties. *Advances in Nursing Science* 1985; **8**:15-24.

Gary RA, Sueta CA, Dougherty M, Rosenberg B, Cheek D, Preisser J *et al*. Home-based exercise improves functional performance and quality of life in women with diastolic heart failure. *Heart and Lung* 2004a;**33**:210-8.

Gary RA, Sueta CA, Rosenberg B, Cheek D. Use of the 6-minute walk test for women with diastolic heart failure. *Journal of Cardiopulmonary Rehabilitation* 2004b;**24**:264-8.

Gorkin L, Norvell NK, Rosen RC, Charles E, Shumaker SA, McIntyre KM *et al*. Assessment of quality of life as observed from the baseline data of the Studies of Left Ventricular Dysfunction (SOLVD) trial quality of life substudy. *American Journal of Cardiology* 1993;**71**:1069-1073.

Grady KL, Meyer PM, Mattea A, Dressler D, Ormaza S, White-Williams C *et al*. Change in quality of life from before to after discharge following left ventricular assist device implantation. *Journal of Heart and Lung Transplantation* 2003a;**22**:322-33.

Grady KL, Meyer PM, Dressler D, White-Williams C, Kaan A, Mattea A *et al*. Change in quality of life from after left ventricular assist device implantation to after heart transplantation. *Journal of Heart and Lung Transplantation* 2003b;**22**:1254-67.

Green CP, Porter CB, Bresnahan DR, Spertus JA. Development and evaluation of the Kansas City cardiomyopathy questionnaire: a new health status measure for heart failure. *Journal of the American College of Cardiology* 2000;**35**:1245-55.

Guyatt GH, Nogradi S, Halcrow S, Singer J, Sullivan MJJ, Fallen EL. Development and testing of a new measure of health status for clinical trials in heart failure. *Journal of General Internal Medicine* 1989; **4**:101-7.

Guyatt GH. Measurement of health-related quality of life in heart failure (1993). *Journal of the American College of Cardiology* 1993;**22**:185A-91A.

Guyatt GH. Measurement of health-related quality of life in heart failure (1994). *Irish Journal of Psychology* 1994;**15**:148-63.

Gwadry-Sridhar FH, Arnold JM, Zhang Y, Brown JE, Marchiori G, Guyatt G. Pilot study to determine the impact of a multidisciplinary educational intervention in patients hospitalized with heart failure. *Am.Heart J* 2005;**150**:982.

Hare DL, Davis CR. Cardiac Depression Scale: validation of a new depression scale for cardiac patients. *Journal of Psychosomatic Research* 1996;**40**:379-86.

Hauptman PJ, Masoudi FA, Weintraub WS, Pina I, Jones PG, Spertus JA. Variability in the clinical status of patients with advanced heart failure. *Journal of Cardiac Failure* 2004;**10**:397-402.

Havranek EP, McGovern KM, Weinberger J, Brocato A, Lowes BD, Abraham WT. Patient preferences for heart failure treatment: utilities are valid measures of health-related quality of life in heart failure. *Journal of Cardiac Failure* 1999; **5**:85-91.

Havranek EP, Simon TA, L'Italien G, Smitten A, Brett HA, Chen R *et al*. The relationship between health perception and utility in heart failure patients in a clinical trial: results from an OVERTURE substudy. *Journal of Cardiac Failure* 2004;**10**:339-43.

Heo S, Moser DK, Riegel B, Hall LA, Christman N. Testing the psychometric properties of the Minnesota Living With Heart Failure Questionnaire. *Nursing Research* 2005;**54**:265-72.

Hlatky MA, Boineau RE, Higginbotham MB, Lee KL, Mark DB, Califf RM *et al*. A brief self-administered questionnaire to determine functional capacity, the Duke Activity Status Index. *American Journal of Cardiology* 1989;**64**:651-4.

Hobbs FD, Kenkre JE, Roalfe AK, Davis RC, Hare R, Davies MK. Impact of heart failure and left ventricular systolic dysfunction on quality of life: a cross-sectional study comparing common chronic cardiac and medical disorders and a representative adult population. *European Heart Journal* 2002;**23**:1867-76.

Höfer S, Lim L, Guyatt GH, Oldridge NB. The MacNew Heart Disease health-related quality of life instrument: a summary. *Health and Quality of Life Outcomes* 2004; **2**:3.

Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Controlled Clinical Trials* 1989;**10**:407-15.

Jaeschke R, Singer J, Guyatt GH. A comparison of seven-point and visual analogue scales. Data from a randomized trial. *Controlled Clinical Trials* 1990;**11**:43-51.

Janz NK, Dodge JA, Janevic MR, Lin X, Donaldson AE, Clark NM. Understanding and reducing stress and psychological distress in older women with heart disease. *Journal of Women and Aging* 2004;**16**:19-38.

Jenkinson CP, Jenkinson D, Shepperd S, Layte R, Petersen S. Evaluation of treatment for congestive heart failure in patients aged 60 years and older using generic measures of health status (SF-36 and COOP charts). *Age and Ageing* 1997a;**26**:7-13.

Jenkinson CP, Layte R, Jenkinson D, Lawrence KC, Petersen S, Paice C *et al*. A shorter form health survey: can the SF-12 replicate results from the SF-36 in longitudinal studies? *Journal of Public Health Medicine* 1997b;**19**:179-86.

Jenkinson CP, Layte R. Development and testing of the UK SF-12. *Journal of Health Services Research and Policy* 1997; **2**:14-8.

Johansson P, Agnebrink M, Dahlstrom U, Brostrom A. Measurement of health-related quality of life in chronic heart failure, from a nursing perspective: a review of the literature. *European Journal of Cardiovascular Nursing* 2004; **3**:7-20.

Kirsch J, McGuire A. Establishing health state valuations for disease-specific states: an example from heart disease. *Health Economics* 2000; **9**:149-58.

Lalonde L, Clarke AE, Joseph L, MacKenzie T, Grover SA, Cassidy LE *et al*. Comparing the psychometric properties of preference-based and non-preference-based health-related quality of life in coronary heart disease. *Quality of Life Research* 1999; **8**:399-409.

Lim LLY, Valenti LA, Knapp JC, Dobson AJ, Plotnikoff R, Higginbotham N *et al*. A self-administered quality-of-life questionnaire after acute myocardial infarction. *Journal of Clinical Epidemiology* 1993;**46**:1249-56.

Lim LLY, Fisher JD. Use of the 12-item Short-Form (SF-12) Health Survey in an Australian heart and stroke population. *Quality of Life Research* 1999; **8**:1-8.

Luther SA, McCullough PA, Havranek EP, Rumsfeld JS, Jones PG, Heidenreich PA *et al*. The relationship between B-type natriuretic peptide and health status in patients with heart failure. *Journal of Cardiac Failure* 2005;**11**:414-21.

Martin AJ, Glasziou PP, Simes RJ. A cardiovascular extension of the Health Measurement Questionnaire. *Journal of Epidemiology and Community Health* 1999;**53**:548-57.

Morgan AL, Masoudi FA, Havranek EP, Jones PG, Peterson PN, Krumholz HM *et al*. Difficulty taking medications, depression, and health status in heart failure patients. *Journal of Cardiac Failure* 2006;**12**:54-60.

Myers J, Zaheer N, Quaglietti S, Madhavan R, Froelicher V, Heidenreich P. Association of functional and health status measures in heart failure. *Journal of Cardiac Failure* 2006;**12**:439-45.

Ni H, Toy W, Burgess D, Wise K, Nauman DJ, Crispell K *et al*. Comparative responsiveness of Short-Form 12 and Minnesota Living with Heart Failure questionnaire in patients with heart failure. *Journal of Cardiac Failure* 2000; **6**:83-91.

O'Keeffe ST, Lye M, Donnellan C, Carmichael DN. Reproducibility and responsiveness of quality of life assessment and six minute walk test in elderly heart failure patients. *Heart* 1998;**80**:377-82.

O'Leary CJ, Jones PW. The Left Ventricular Dysfunction questionnaire/LVD-36: reliability, validity, and responsiveness. *Heart* 2000;**83**:634-40.

Oldridge NB, Perkins A, Hodes Z. Comparison of three heart disease-specific health-related quality of life instruments. *Monaldi Archives for Chest Disease* 2002;**58**:10-8.

Padilla GV, Presant C, Grant MM, Metter G, Lipsett J, Heide F. Quality of life index for patients with cancer. *Research in Nursing and Health* 1983; **6**:117-26.

Padilla GV, Grant MM. Quality of life as a cancer nursing outcome variable. *Advances in Nursing Science* 1985; **8**:45-60.

Park SJ, Tector A, Piccioni W, Raines E, Gelijns A, Moskowitz A *et al*. Left ventricular assist devices as destination therapy: a new look at survival. *Journal of Thoracic and Cardiovascular Surgery* 2005;**129**:9-17.

Prasun MA, Kocheril AG, Klass PH, Dunlap SH, Piano MR. The effects of a sliding scale diuretic titration protocol in patients with heart failure. *Journal of Cardiovascular Nursing* 2005;**20**:62-70.

Rector TS, Kubo SH, Cohn JN. Patients' self-assessment of their congestive heart failure - Part 2: content, reliability and validity of a new measure, the Minnesota Living with Heart Failure Questionnaire. *Heart Failure* 1987;Oct-Nov:198-209.

Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Pimobendan Multicenter Research Group. *American Heart Journal* 1992;**124**:1017-25.

Rector TS, Kubo SH, Cohn JN. Validity of the Minnesota Living with Heart Failure Questionnaire as a measure of therapeutic response to enalapril or placebo. *American Journal of Cardiology* 1993a;**71**:1106-7.

Rector TS, Johnson G, Dunkman WB, Daniels G, Farrell L, Henrick A *et al*. Evaluation by patients with heart failure of the effects of enalapril compared with hydralazine plus isosorbide dinitrate on quality of life. V-HeFT II. The V-HeFT VA Cooperative Studies Group. *Circulation* 1993b;**87**:VI 1-VI 7.

Rector TS, Tschumperlin LK, Kubo SH, Bank AJ, Francis GS, McDonald KM *et al*. Use of the Living With Heart Failure questionnaire to ascertain patients' perspectives on improvement in quality of life versus risk of drug-induced death. *Journal of Cardiac Failure* 1995; **1**:201-6.

Rector TS, Anand IS, Cohn JN. Relationships between clinical assessments and patients' perceptions of the effects of heart failure on their quality of life. *Journal of Cardiac Failure* 2006;**12**;87-92.

Reddy P,.Dunn AB. The effect of beta-blockers on health-related quality of life in patients with heart failure. *Pharmacotherapy* 2000;**20**:679-89.

Riegel B, Moser DK, Glaser D, Carlson B, Deaton C, Armola R *et al*. The Minnesota Living With Heart Failure Questionnaire: sensitivity to differences and responsiveness to intervention intensity in a clinical population. *Nursing Research* 2002;**51**:209-18.

Rodriguez-Artalejo F, Guallar-Castillon P, Pascual CR, Otero CM, Montes AO, Garcia AN *et al*. Health-related quality of life as a predictor of hospital readmission and death among patients with heart failure. *Archives of Internal Medicine* 2005;**165**:1274-9.

Rukholm E,.McGirr M. A quality-of-life index for clients with ischemic heart disease: establishing reliability and validity. *Rehabilitation Nursing* 1994;**19**:12-6.

Rukholm E, McGirr M, Potts J. Measuring quality of life in cardiac rehabilitation clients. *International Journal of Nursing Studies* 1998;**35**:210-6.

Rumsfeld JS, Havranek E, Masoudi FA, Peterson ED, Jones P, Tooley JF *et al*. Depressive symptoms are the strongest predictors of short-term declines in health status in patients with heart failure. *Journal of the American College of Cardiology* 2003;**42**:1811-7.

Sethares KA, Elliott K. The effect of a tailored message intervention on heart failure readmission rates, quality of life, and benefit and barrier beliefs in persons with heart failure. *Heart and Lung* 2004;**33**:249-60.

Sidorov J, Shull RD, Girolami S, Mensch D. Use of the Short Form 36 in a primary care-based disease management program for patients with congestive heart failure. *Disease Management* 2003; **6**:111-7.

Sneed NV, Paul S, Michel Y, Van Bakel A, Hendrix G. Evaluation of three quality of life measurement tools in patients with chronic heart failure. *Heart and Lung* 2001;**30**:332-40.

Spertus J, Peterson E, Conard MW, Heidenreich PA, Krumholz HM, Jones P *et al*. Monitoring clinical changes in patients with heart failure: a comparison of methods. *American Heart Journal* 2005;**150**:707-15.

Subramanian U, Weiner M, Gradus PG, Wu J, Tu W, Murray MD. Patient perception and provider assessment of severity of heart failure as predictors of hospitalization. *Heart and Lung* 2005;**34**:89-98.

Sullivan MD, Newton J, Hecht J, Russo JE, Spertus JA. Depression and health status in elderly patients with heart failure: a six-month prospective study in primary care. *American Journal of Geriatric Cardiology* 2004;**13**:252-60.

The Criteria Committee of the New York Heart Association. Nomenclature and criteria for diagnosis of diseases of the heart and great vessels. pp 253-6. Boston, Massachusetts, USA: Brown & Co, 1994.

Valenti L, Lim L, Heller RF, Knapp J. An improved questionnaire for assessing quality of life after acute myocardial infarction. *Quality of Life Research* 1996; **5**:151-61.

Wiklund IK, Lindvall K, Swedberg K, Zupkis RV. Self-assessment of quality of life in severe heart failure: an instrument for clinical use. *Scandinavian Journal of Psychology* 1987;**28**:220-5.

Wolinsky FD, Wyrwich KW, Nienaber NA, Tierney WM. Generic versus disease-specific health status measures: an example using coronary artery disease and congestive heart failure patients. *Evaluation and the Health Professions* 1998;**21**:216-43.

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care* 1999;**37**:469-78.

Wyrwich KW, Tierney WM, Babu AN, Kroenke K, Wolinsky FD. A comparison of clinically important differences in health-related quality of life for patients with chronic lung disease, asthma, or heart disease. *Health Services Research* 2005;**40**:577-91.

Zambroski CH, Moser DK, Bhat G, Ziegler C. Impact of symptom prevalence and symptom burden on quality of life in patients with heart failure. *European Journal of Cardiovascular Nursing* 2005; **4**:198-206.

Chapter 9: Patient-reported Health Instruments used for people with Stroke

There are two types of stroke: ischaemic, where there is either a cerebral thrombosis or embolism obstructing the blood supply; or haemorragic, either intracerebral or subarachnoid. There are several problems or disabilities stroke survivors may face in the first few weeks after having a stroke. Most of these will improve over time as the brain recovers. In severe cases, they may cause long-term disability. Hemiplegia is the most common symptom of a stroke, usually happening in one side of the body. The weakness or paralysis results in unsteady gait and stiffness or spasticity of the muscles and joints. There are many other problems associated with having a stroke including functional aspects such as difficulty in swallowing (dysphagia), speaking and understanding (dysphasia); impaired mobility; and increased need for assistance with activities of daily living. The emotional impact of having a stroke combined with the inability to communicate effectively causes further burden to the patient and carers. Recovery can be slow and full functioning may never return to pre-stroke status.

The following review provides current information available on the patient-reported health questionnaires used to measure health-related quality of life in patients with stroke.

Search terms and results: identification of articles

At the time of the review, the PHI database contained 12,000+ records (up to June 2005). The primary search strategy, using the terms 'stroke,' generated 277 records, as shown in Table 9.1. All abstracts were reviewed. When assessed against the review inclusion criteria, 114 articles were retrieved and reviewed in full. Of these, 54 articles were included in the review.

Table.9.1 Number of articles identified by the literature review

Source	*Results of search*	*No. of articles considered eligible*	*Number of articles included in review*
PHI database: original search (up to June 2005) **Total number= 12,562**	277	114	44
Supplementary search	-	-	10
TOTAL	-	-	**54**

Supplementary searches included hand-searching of titles from 2004 to 2006 of the following key journals:
-Clinical Rehabilitation
-Health and Quality of Life Outcomes
-Medical Care
-Quality of Life Research
-Stroke

Further searches were conducted within the bibliography and using PubMed per instrument up to September 2006.

Identification of patient-reported health instruments
Six generic and 10 stroke-specific instruments were included in the review. The developmental and evaluative studies relating to the generic instruments reviewed are listed in Tables 9.2 to 9.5. Those relating to stroke-specific instruments are shown in Tables 9.8 to 9.14.

RESULTS: GENERIC PATIENT-REPORTED HEALTH INSTRUMENTS

Six generic instruments were identified which were evaluated with patients with stroke. Full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36
b) SF-12
c) SF-6D
d) EQ-5D
e) Health Utilities Index
f) Nottingham Health Profile

a) SF-36:

The SF-36 is the most widely validated measure of subjective health status in stroke. Fifteen papers were found evaluating the SF-36 in stroke, of which six were based on data available from the UK.

Reliability
Internal consistency reliability was examined in four studies (Anderson et al., 1996; Dorman et al., 1998; Hagen et al., 2003; Hobart et al., 2002). Internal consistency reliability coefficients (Cronbach's alpha) were found to satisfy Nunnally's criterion of 0.7 (Nunnally, 1978) in most instances. Internal consistency reliability coefficients generally fell between 0.80 to 0.95 (Anderson et al., 1996; Dorman et al., 1998). However, although not an issue raised in the papers reviewed here, one possibility for high alpha (over 0.90) statistics in the role functioning domains may be due to floor or ceiling effects as these dimensions have dichotomous response sets in Version 1 of the SF-36 (the version reviewed in these papers). None-the-less, results in general were good, and unlikely to be purely caused by floor or ceiling effects on items. However, in an interview based survey of 90 respondents in Australia the internal consistency reliability of the Vitality dimension was found to be low (alpha=0.6) and below the threshold cited by Nunnally (Anderson et al., 1996). The internal reliability of the SF-36 was assessed in early post stroke patients and found to be generally acceptable (Hagen et al., 2003). However, Vitality (alpha = 0.68) at one month post-stroke and General Health (alpha=0.67) three months post-stroke, fell below the accepted criteria of 0.70. Similarly, Hobart et al., (2002) found alpha coefficients to be lower for the General Health dimension (alpha=0.68) in a study of 177 patients.

Reproducibility of the SF-36 was assessed in a study of UK patients randomly selected from the International Stroke Trial (Dorman et al., 1999). SF36 domains were generally found to be acceptable, except for the Mental Health domain (ICC=0.30 when completed by patients; ICC=0.24 when patient assisted by a relative or friend). However, 95% CIs for the mean differences between scores between test and re-test were substantial across dimensions. Reproducibility was higher when patients completed instruments alone than when they were proxy rated.

Item-total correlations were reported to be good by Hagen et al., (2003), although the worst were for the item 'I expect my health to get worse' in the General Health scale. This item may well seem irrelevant after a stroke.

Validity

Construct validity was assessed by Hobart et al., (2002) who reported item-total correlations in excess of 0.4 for all items in their respective dimensions, except for two items in the General Health scale.

Dorman et al., (1999) found that the domains of Physical Functioning, Social Functioning, Bodily Pain and General Health as measured by the SF-36 and EuroQol instruments were strongly correlated. However, Mental Health as measured on the SF-36 was poorly correlated with the Psychological Functioning domain of the EuroQol (rho =0.21, $p<0.001$). The authors suggest that this may be due to the fact that the domains are measuring somewhat different constructs. However, the authors also suggest that the SF-36 Mental Health domain may have poor measurement properties in stroke, and they suggest poor reproducibility (ICC=0.28) as evidence of this. However, as they failed to ask recipients if any aspect of their health had changed since baseline it is difficult to know if these results reflect change or poor measurement properties on the Mental Health dimension of the SF-36. They also suggest that as many questionnaires were completed by proxies, and proxy report is unreliable when assessing mental health, this could be a major cause of the Mental Health domain's apparent measurement problems.

Hackett et al., (2004) compared SF-36 scores of stroke patients who had experienced a stroke six years previously with age-sex standardised normative data. They found scores were worse for stroke patients on six of the eight domains, but that Mental Health and Pain scores did not differ from the controls.

Anderson et al., (1996) assessed the construct validity by comparing the SF-36 scores to those on the patient completed Barthel Index (a measure of physical disability) and the 28-item General Health Questionnaire (a measure of non-psychotic psychiatric disturbance), controlling for age and sex in multiple regression analyses. Significant associations were found between the Physical Functioning scale on the SF-36, the Barthel Index (Beta=-0.55, $p<0.001$), and the Role Limitations-Emotional and Social Functioning scales and the GHQ-28 total score (Beta=-0.41, $p<0.001$). Furthermore, Anderson et al., (1996) did not find evidence of ceiling effects on any of the dimension scores which, they claim, is a common problem on many disability scales. However, they were critical of the measurement properties of the Social Functioning scale. They utilised an instrument called the Adelaide Activities Profile (AAP) (Clark et al., 1995) to assess daily activities, including domestic chores, household maintenance, service to others and social functioning. They found no association

between AAP scores and scores on the Social Functioning scale and concluded that the Social Functioning domain did not assess in a way that was meaningful to stroke patients.

Hagen et al., (2003) found significant correlations between SF-36 domains and a self completed version of the Barthel Index and the Canadian Neurological Scale (Cote et al., 1986) over three administrations. The Physical Functioning domain was very highly correlated with the Barthel Index and the Canadian Neurological Scale. Duncan et al., (2002) found moderate correlations between physical measures of health status on the Stroke Impact Scale and the Physical Dimension of the SF-36.

Patel et al., (2006) found a graded positive relationship between all SF-36 domains and the Barthel Index and Frenchay Activities Index.

Scores on the Physical Function domain of the SF-36 were found to be highly related to the modified Rankin Stroke Outcome Scale (a measure of disability assessed by clinicians) in a survey of 459 stroke patients in the USA (Duncan et al., 2000).

The developers of the SF-36 suggest a method of calculating two summary scores from the results gained on the eight dimensions. Hobart et al., (2002) tried to replicate this work using higher order factor analytic techniques on their data. They found the hypothesised two factor solution, but as it accounted for only 60% of the variance they argue that a substantial amount of the information is lost by reporting these summary statistics alone.

Williams et al., (1999b) found, in a regression analysis that the SF-36 did not predict overall self-reported quality of life, and suggest the instrument may be insufficiently sensitive to quality of life changes after stroke.

Responsiveness

Responsiveness was assessed in patients at one, three and six months post-stroke (Hagen et al., 2003). The authors suggest that low sensitivity to change was found on three SF-36 scales: Bodily Pain, General Health and Mental Health, and for other subscales sensitivity to change was comparable to the Barthel Index. However, the evidence presented in their paper suggests that the Barthel Index indicated far greater change (SRM=0.51) than that found on any of the dimensions on the SF-36 (indeed, the SRMs only get close to this on two dimensions: Social Functioning=0.39; Role Physical = 0.33).

Precision

Floor and ceiling effects were reported in a number of studies. Role Physical was found to have substantial floor effects (70%) in the study reported by Hagen et al., (2003). Other 'end' effects reported in that study included 35% ceiling effects for the Pain dimension, 23% scores exhibiting a floor effect for Physical Functioning and 27% floor and 16% ceiling effects for Social Functioning. Hobart et al., (2002) also found serious floor effects (59.1%) for the Role Physical dimension and ceiling effects for Social Functioning (29.9%), Bodily Pain (25.6%) and Role Emotional (63.1%) domains. Somewhat different results were reported by Anderson et al., (1996), who found considerable ceiling effects for the Role Physical, Role Limitations, Social Functioning and Bodily Pain dimensions of the SF-36, and

Hamedani et al., (2001) who also found ceiling effects for Physical Functioning, Role Physical, Bodily Pain, Social Functioning and Role Emotional. O'Mahony et al., (1998) report ceiling effects on the Role Emotional, Role Physical, Social Functioning, Mental Health, and Bodily Pain dimensions in a small scale survey of older stroke patients. They also report floor effects on Role Physical, Role Emotional, Mental Health and Physical Functioning domains. Pickard et al., (2005) report ceiling effects on the Role Emotional domain and floor effects on the Physical Functioning, Role Physical and Role Emotional domains.

Lai et al., (2004) compared results gained from the SIS Participation Domain and the SIS-16 (measuring Physical Function) with the SF-36 Social Functioning and Physical Functioning Domains. Rasch analyses indicated that both the SIS-16 and SF-36 Physical Functioning domain both showed a good spread of item difficulty, but the SIS-16 incorporates easier items that are capable of measuring lower levels of physical functioning in patients with severe stroke. Similar analyses of the SIS Participation domain and the SF-36 Physical Functioning domain indicated that the SIS measure has widespread item difficulty, whereas the SF-36 domain does not. The Social Functioning domain of the SF-36 contains only two items, measuring the same level of item difficulty, leading to severe ceiling effects and consequently an inability to discriminate among more active patients.

Acceptability

Anderson et al., (1996) reported that of the 124 patients approached to undertake an interview administration of the SF-36, 13 were unable to communicate sufficiently well to complete the instrument. Dorman et al., (1999) randomly selected UK patients from the International Stroke Trial. An initial survey was undertaken in which patients completed either the EuroQol questionnaire or the SF-36. Respondents to the EuroQol were then mailed a copy of the SF-36 at a three week follow-up (n=272), and respondents to the EuroQol were mailed a copy of the SF-36 at follow-up (n=505). Ninety-one percent sent the EuroQol at follow-up replied, whilst 85% of those sent the SF-36 at follow-up responded.

O'Mahony et al., (1998) reported poor completion rates on the SF-36 and consequently difficulties calculating dimension scores in older age group stroke patients. Similarly, O'Mahony et al., (1998) claimed that completion rates for some items were as low as 66%. Hagen et al., (2003) simply reported that some of the patients in their study 'encountered some problems' completing the SF-36.

Dorman et al., (1997, 1999) randomised all patients who had been entered by UK centres to the International Stroke Trial between March 1992 and May 1995, who were not known to have died, to either the EuroQol or SF-36 instruments. The acceptability of the EuroQol appeared superior with a 5% difference in returns between the two measures, whilst missing data was found on returned SF-36 forms in 45% cases and 34% on the EuroQol.

Feasibility

Segal and Schall (1994) evaluated the feasibility of using reports by carers to complete the SF-36 (which they refer to in the paper as the Health Status Questionnaire - HSQ). Proxy agreement with patient evaluations was low, and the authors claim that the instrument is an inadequate outcome measure in stroke. They

find high levels of association between patient and carer completion of functional assessment measures (the Functional Independence Measure - FIM and Frenchay Activities Index - FAI), which they suggest indicates their superior measurement properties. This view could be criticised as the FAI and FIM were designed for completion by observers of the patient, whereas the SF-36 was designed to tap subjective experience, which is not always readily observed.

b) SF-12

The SF-12 contains a sub-set of the items included in the SF-36, and was initially designed to reduce patient burden and provide the summary Mental Health and Physical Health Component Scores. The instrument was assessed in six papers, none of which were based on data collected in the UK.

Reliability

Bohannon et al., (2004a) evaluated the internal consistency of the twelve items of the SF-12 using the alpha statistic, and found internal consistency reliability for the measure as a whole to be high at three different times, following stroke, and three months and twelve months after stroke (alpha values = 0.83, 0.88 and 0.89, respectively). Bohannon et al., (2004b), in a separate study, evaluated the test-retest reliability of the SF-36 in a small telephone interview based survey of 31 stroke patients. The SF-12 was administered at two occasions 16.2 +/- 5 days apart. The mean difference between the two administrations was less than 1.5 points on both the Physical Component Scale (PCS) and Mental Component Scale (MCS) scores. The authors claim that ICC's for both summary scores of 0.80 $p<0.001$, are good, though more realistically this result may be judged as promising or satisfactory. In part, ICC's may not be higher as the authors did not indicate that they had removed any respondent who reported their health had changed during the period between the two administrations of the instrument.

Internal consistency reliability of the SF-12 was found to be high by King et al., (2005), with a reported alpha of 0.76. Similarly, internal consistency reliability is reported by Lim and Fisher (1999), in a study of heart disease and stroke patients, for both 'the items' of the PCS and MCS. This appears to suggest that the authors have used different items to calculate the PCS and MCS scores. This is not the method by which SF-12 scores are calculated, as the developers suggest that they are created by differentially weighting the same items: consequently, it is difficult to interpret the results reported by Lim and Fisher (1999). However, their results would tend to suggest that they may have incorrectly calculated the PCS and MCS, or they have calculated alpha coefficients on weighted items.

Validity

The developers of the SF-12 claim that it can be used to measure two distinct domains, the Physical and Mental Component scores. Consequently, Bohannon et al., (2004) used principal components analysis with varimax rotation to determine if the hypothesised scales existed in the SF-12 for a stroke sample. The analysis resulted in a two factor solution, which, the authors state, 'presumably' reflects the hypothesised dimensions.

Lim and Harris (1999) report that trends for PCS and MCS scores were worse for those who were older and women who had longer hospital stays. However, the statistically significant results are possibly due to multiple comparisons on a large dataset. Furthermore, the data are aggregated heart disease and stroke patients.

Rubenach et al., (2000) in a small scale telephone survey found that PCS scores were able to discriminate patients classified as dependent from those classified as independent in activities of daily living as indicated on the Barthel Index. They also found poorer PCS and MCS scores observed in patients with high GHQ-28 scores. They suggest this provides evidence that the SF-12 provides a 'valid indicator of health-related quality of life among patients with stroke'. However the association of GHQ-28 scores with the PCS may be seen as evidence against such a claim. The authors counter such a potential criticism by claiming that the GHQ-28 'may reflect questions with a somatic emphasis'. This is indeed true (the GHQ-28 can provide a Somatic sub-scale score), but scores can be calculated from the GHQ-28 to overcome this (i.e. scores for the Severe Depression and Anxiety sub-scales). However, Rubenach et al., (2000) do not undertake such analyses.

King et al., (2005) found a relationship between Hospital and Anxiety Scale (HAD) Anxiety scores and the MCS on the SF-12. However, no such relationship was found between MCS and the HAD Depression scale. The authors suggest that the HAD Depression scale is measuring a somewhat different aspect of mental health than the MCS, which is assessing general mental health status. They also report high levels of association between functional status measures (including the self-report Barthel Index and Glasgow Outcome Scale) and the PCS.

Pickard et al., (1999) compared SF-12 Physical Mobility Component Scores (PCS-12) to SF-36 Physical Mobility Component Scores (PCS-36) and found them be highly correlated (intra-correlation coefficient 0.95). Similarly, Mental Health Component Scores (MCS-12) were compared to SF-36 Mental Health Component Scores (MCS-36) and found to be highly correlated (ICC = 0.97). Mean scores between the two measures were separated by only a few points (Pickard et al., 1999). However, such small differences can be meaningful (Jenkinson 1998), and could mean the SF-12 is not exactly replicating SF-36 results in this patient group.

Responsiveness

Bohannon et al., (2004) report results from a longitudinal survey over a period of twelve months, and report that PCS scores drop three months after stroke, and then improve at follow-up twelve months later. MCS scores did not change over the 12 months of the study. The authors claim that these results suggest that the SF-12 is sensitive to changes in health as a result of stroke, but provide no evidence that the changes are either accurate or meaningful.

Precision

No data available.

Acceptability

Lim and Harris claim that over 50% of respondents (heart disease and stroke patients) omitted at least one item which, unless a data substitution algorithm is used, would suggest over half of the PCS and MCS scores could not be calculated. Rubenach et

al., (2000), however, claim that the SF-12 may be an appropriate instrument to use in postal surveys. In a small scale survey (n=45) by telephone they claimed that '85% of patients who were able to be interviewed fully completed the SF-12'.

Feasibility
No data available.

c) SF-6D
The SF-6D index score can be calculated from six items of the SF-36. It is only included in this review because any data set containing the SF-36 is amenable to such analyses. It is a preference/utility based measure intended for providing an index intended for use in QALY calculations.

Reliability
No data available.

Validity
QALY estimates based on the SF-6D were half as large as those calculated when using the HUI3 or EQ-5D Index (Pickard et al., 2005), which may cast some doubt as to the appropriateness of this instrument in stroke.

Responsiveness
The SF-6D was found to be more responsive to change than the EuroQol (Pickard et al., 2005). Pickard et al., (2005) also report that change scores were found to be highly correlated with EQ-VAS, EQ-5D Index and HUI3.

Precision
No data available.

Acceptability
No data available.

Feasibility
No data available.

a) SF-36; SF12; SF-6D

Table 9.2: Developmental and evaluation studies relating to the SF-36, SF-12 and SF-6D in stroke

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Anderson C, et al., 1996 Australia	Stroke patients (90) Age: mean 72 Out-patients Interview administered	Internal consistency ✓	Construct ✓			✓	
Dorman P et al., 1997 UK	Stroke patients (2253) Age: mean not specified RCT of measures Postal administration		Construct			✓	
Dorman P et al., 1998 UK	Stroke patients (SF-36 n=253; EuroQol=271) Age: mean not specified Out-patients Postal administration	Internal consistency ✓ Test-retest ✓	Construct ✓			✓	
Dorman P et al., 1999 UK	Stroke patients (2253) Age: mean not specified RCT of measures Postal administration		Construct ✓			✓	
Duncan P, et al., 2000 America	Stroke patients (459) Age: 70+/- 11.4 years Out-patients Method of administration not specified		Construct ✓				
Duncan et al., 2002 America	Stroke patients (125) Mean age = 68.1 Telephone administration of SF-36		Construct ✓				
Hackett et al., 2000 New Zealand	639 stroke patients and 310 controls 76% of cases aged 65 or over Stroke patients interviewed 6 years after stroke		Construct ✓				
Hamedani et al., 2002UK	111 stroke patients (40 interviewed, 71 sent questionnaire) Patients aged between 18 and 49 Open ended interviews and telephone administered questionnaire interviews				✓		

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hobart et al., 2002 UK	177 Stroke patients Mean age=62 Face to face interview administration of SF-36	Internal consistency ✓	Construct ✓		✓		
Lai et al., 2003 USA	278 individuals with stroke Age (mean) 72.5 Interview survey				✓		
O'Mahony et al., 1998	Stroke patients (73) Age: impossible to determine.				✓	✓	
Patel et al., 2006 UK	Stroke patients (490) Age: SF-36 collected by interview		Construct ✓				
Pickard et al., 2005 Canada	Stroke patients (n=124) Age (mean) 67 Self completion but 53% of respondents required assistance. Longitudinal survey				✓		
Segall and Schall, 1994 USA	Stroke patients (38) and their carers Age mean 65 (patients) and 54 (carers)					✓	Proxy versus self - report
Williams et al., 1999a	Stroke patients (n=71) Age mean 61 Interview administered survey of patients in three hospitals		Construct ✓				
SF-12		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bohannon et al., 2004a USA	Stroke patients (90) Age: mean 70.4 In-patients Interview administered	Internal consistency ✓	Construct ✓				
Bohannon et al., 2004b USA	Stroke patients (31) Age: mean 66.5 In-patients Interview administered, by telephone	Test re-test ✓	Construct				
King et al., 2005 USA	Stroke patients (n=170) Mean age = 53.5 years SF-12 self completed	Internal consistency ✓	Construct ✓				
Lim and Fisher 1999 Australia	(2341 respondents of which 62% diagnosed with stroke) Age: mean 66.5 Postal survey		Construct ✓			✓	

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
SF-12		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Pickard et al., 1999 Canada	Stroke patients (n=161, of which 53 proxy completed) Age (mean) 72.11 Self completion but 32.92% proxy completed)		Construct ✓				
Rubenach et al., 2000 Australia	Stroke patients (40) Age: not specified Telephone interview	Internal consistency ✓	Construct ✓			✓	
SF-6D							
Pickard et al., 2005 Canada	Stroke patients (n=124) Age (mean) 67 Self completion but 53% of respondents required assistance. Longitudinal survey		Construct ✓				

d) EuroQol- EQ-5D

The EuroQol was evaluated in six papers, two of which were based on data gained from the UK (Dorman et al., 1997, 1999). Results from the UK papers were based upon the same dataset.

Reliability
Test-retest reliability was found to be good for the EuroQol in a study of UK patients randomly selected from the International Stroke Trial (Dorman et al., 1997, 1999). The authors claim the EuroQol overall score has greater reproducibility than individual items whether weighted by utility weights or not. Unweighted Kappa values for the utility weighted EuroQol were 0.83 for questionnaires completed by the patient alone and 0.81 for those completed by proxy.

Validity
The construct validity of the EuroQol was assessed by Dorman et al., (1999) who compared results on the measure with those gained from the SF-36. Measures assessing Physical Functioning, Social Functioning, Bodily Pain and Overall Health were highly correlated, but this was not the case for Mental Health which was poorly correlated on the two measures. They suggest this may be due to the instruments tapping different aspects of mental health or possible measurement error on the SF-36. However, no firm conclusion is drawn from this finding.

McPherson et al., (2004) compared population based valuation estimates for the EuroQol with those elicited from patients. Valuations provided by stroke patients were significantly different from population-based ratings and correlations between EuroQol Index calculations based on the two weighting schemes were poor. Population based ratings of health are systematically lower than ratings gained from patients with stroke. Additionally, the magnitude of this difference depends on health status in a curvilinear way, increasing as health state severity increases but decreasing in the most severe states. The authors conclude that the valuations used in any given survey could have considerable effects on the results, and this has important implications for interpreting shifts in health status valuations following interventions.

Polsky et al., (2001) examined the health status of patients enrolled in a clinical trial for a new drug for treating aneurysmal subarachnoid haemorrhage. These assessments were made using the EuroQol classification and weighting system, and also the visual analogue 'thermometer'. They developed a model for predicting responses to the thermometer and derived scoring weights for the EuroQol health state classification that met convergent validity criterion of having higher predicted scores for better outcomes and lower scores for worse outcomes. They suggest the scoring rule they developed could be used to impute health valuations in clinical trials when self-rating for health states is not possible. Additionally, they found differences on scores gained from stroke patients than from the general public, with the general public rating higher (i.e. better) levels of function more favourably than stroke patients, yet worse levels of function less favourably than stroke patients.

Responsiveness
Change scores EQ-VAS and EuroQol EQ-5D Index have been found to be highly correlated with results from other utility measures (SF-6D and HUI3), as well as clinically assessed Barthel Index change scores (Pickard et al., 2005).

Precision
Poissant et al., (2004) reported that in a 'high functioning' stroke population the EuroQol EQ-5D exhibited an end effect with many patients scoring as 'perfect health' on the utility index but not on the EQ-VAS.

Acceptability
The acceptability of the EuroQol has been evaluated in a study in which the EuroQol and the SF-36 were randomly allocated to patients taking part in the International Stroke Trial (Dorman et al., 1997, 1998). One thousand one hundred and twenty five (1125) patients were randomly selected to receive the EuroQol and 1128 to receive the SF-36. The response frequency was found to be statistically significantly higher for the EuroQol (80% versus 75%). Patients returning the questionnaire were then sent another copy 'within approximately three weeks' (-sic) to assess response rate and test-retest reliability. A similar proportion responded for each questionnaire (86% for the EuroQol versus 83% for the SF-36). Respondents were asked if they required help completing the instruments and 52% requested help with the EuroQol and 51% with the SF-36.

Table 9.3: Developmental and evaluation studies relating to the EuroQol in stroke

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
EuroQol		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Dorman et al., 1999 UK	Stroke patients (2253) Age: mean not specified RCT of measures Postal administration	Test re-test ✓	Construct ✓			✓	
Dorman P et al., 1997 UK	Stroke patients (2253) Age: mean not specified RCT of measures Postal administration		Construct			✓	
McPherson et al., 2004 New Zealand	Stroke patients (585) age 79% aged 60 or over; 54% aged 70 or over Postal survey		Construct ✓				
Pickard et al.,.2005 Canada	Stroke patients (n=124) Age (mean) 67 Self completion Longitudinal survey		Construct	✓			
Poissant et al., 2003 Canada	Stroke patient (n=91) Age (mean) 69 Six months post stroke Self completion		Construct		✓		
Polsky et al., 2000 USA	Stroke patients (649, aneurysmal subarachnoid haemorrhage) Age: mean 50 Interview administration		Construct ✓				

e) Health Utilities Index (HUI)

The Health Utility Index is a preference/utility instrument designed for use in economic analyses. The measure has been updated and is currently in its third version (HUI-3). Three papers were found that report on the evaluation of the HUI-3 in the USA and Canada.

Reliability
Goldstein et al., (2002) found reasonable test-retest results for most dimensions of the HUI-3. However, the Speech dimension on the HUI-3 showed very poor test-retest reliability (ICC=0.28).

Goldstein et al., (2002) found no significant differences between mean scores for patient and carer pairs when completing the HUI-3. However, this may be due to the small sample size (n=73 pairs at two time periods) and high degrees of missing data (see below) as correlation coefficients were variable, ranging from a low 0.24 to a high 0.88. The fact that the data was pooled (i.e. patient and carers completed the measures at two time periods) may also artificially raise the level of correlation.

Validity
The construct validity of the HUI-3 was assessed by Grootendorst et al., (2000) in respondents reporting having had a stroke or arthritis in the Ontario Health Survey. Subjects with stroke (n=173) or arthritis (n=7,751) had substantially lower health related quality of life than those not reporting such conditions (referred to as the 'reference group'). Respondents with stroke reported worse health on the Global utility Index than either arthritis patients or the reference group (n=53,838). Furthermore, stroke patients had lower (i.e. worse) scores on all eight dimension scores on the eight single attribute scores.

HUI-3 scores were compared to known groups defined by the Barthel Index. Scores on the HUI measures were found to distinguish between mild and moderate/severe cases as defined by the Barthel Index, but did not distinguish between moderate and severe groups.

Responsiveness
Change scores for HUI-3 have been found to be highly correlated with results from other utility measures (EQ-5D and SF-6D), as well as clinically assessed Barthel Index change scores (Pickard et al., 2005). Pickard et al., (2005) also report that the HUI-3 was found to be more responsive than the HUI-2 and the VAS on the EuroQol.

Precision
No data available.

Acceptability
Goldstein et al., (2002) report that the percentage of missing data on the HUI-3 was ‘surprisingly high’ with at least one item of assessment missing in over 70% of cases. They argue that the high proportion of missing data would limit the usefulness of the HUI-3 in the context of stroke trials.

Feasibility
No data available.

Table 9.4: Developmental and evaluation studies relating to the HUI-3 in stroke

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
HUI		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Goldstein LB et al., 2002 USA	Stroke patients (73) Age: 64 years population survey Method of administration Telephone Interview survey	Test-retest ✓	Construct ✓				
Grootendorst P, et al., 2000 Canada	Stroke patients (173) Age: 63 years population survey Method of administration postal survey		Construct ✓				
Pickard et al., 2005 Canada	Stroke patients (n=124) Age (mean) 67 Self completion but 53% of respondents required assistance. Longitudinal survey			✓			

f) Nottingham Health Profile (NHP)

Reliability
The test re-test reliability of the NHP was assessed in a postal study of 21 stroke patients. Questionnaires were sent to patients on the North East Thames Outcome Study six months after a stroke, and then a further questionnaire was sent two weeks later. The authors report significant variation in scores between the two administrations, and poor coefficients of repeatability (Trigg and Wood, 2000).

Validity
The construct validity of the NHP was indirectly assessed in a survey evaluating the Subjective Index of Physical and Social Outcome (SIPSO) (Trigg and Wood, 2000). The NHP domain of Mobility was highly correlated with scores on the Physical subscale of the SIPSO, and the NHP domains of Emotional Health and Social Functioning were highly correlated with the Social subscale of the SIPSO.

Responsiveness
No information available.

Precision
No information available.

Acceptability
No information available.

Feasibility
No information available.

Table 9.5: Developmental and evaluation studies relating to the NHP in stroke

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
NHP		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Trigg and Wood, 2000	Stroke patients (157) Age not specified Postal survey		Construct ✓				
Gompertz et al., 1993 UK	Stroke patients (21) Age (mean) 69 Longitudinal postal survey	Test-retest ✓					

RESULTS: STROKE-SPECIFIC PATIENT REPORTED HEALTH INSTRUMENTS:

Seven Stroke -specific instruments were identified which were evaluated with patients with COPD. Full details of the development, domains and scoring methods are detailed in Tables 9.6 and 9.7.

The following instruments measurement properties are reported:

a) Stroke Impact Scale
b) Stroke Specific Quality of Life Scale
c) Subjective Index of Physical and Social Outcomes
d) The Barthel Index
e) Frenchay Activities Index
f) Nottingham Extended Activities of Daily Living (ADL) Scale
g) London Handicap Scale

a) Stroke Impact Scale Versions 2 and 3

The developers of the Stroke Impact Scale noted that many instruments, such as the SIP and SF-36, exhibited ceiling and floor effects in stroke populations. Consequently, these measures had limited ability to evaluate stroke outcomes over time. Consequently, they decided to develop a stroke specific measure that may overcome such problems (Lai et al., 2003). The instrument content was derived from input from stroke patients, caregivers and health professionals with experience in the field of stroke. It contains 59 items across eight domains (Strength, Hand Function, ADL/IADL, Mobility, Emotion, Memory, Communication and Social Participation). A related measure to the SIS is the SIS-16 which was designed to assess physical functioning and be more sensitive to differences than existing measures of physical function. The SIS-16 contains 16 items from the SIS measuring ADL/IADL, mobility and hand function (Edwards and O'Connell, 2003). The SIS Version 3 contains minor modifications but consists of the same items and domains as the SIS Version 2. Note: Version 1 of the SIS is reported only in unpublished literature.

b) Stroke Specific Quality of Life Scale (SS-QOL)

At the time of the development of the SS-QOL the authors argued that there was no stroke specific health related quality of life measures available. Consequently, Williams et al., (1999a) set about devising a stroke specific QOL measure developed from interviews with patients. Thirty four survivors of ischemic stroke were interviewed to identify common themes that affect stroke patients' quality of life. Subjects included in the interviews were identified from stroke clinics one to six months after stroke and with no significant cognitive or language impairment. Patients were asked to identify three areas most affected by their stroke. Twelve commonly affected domains were identified: energy, family roles, language, mobility, mood, personality, self-care, social roles, thinking, upper extremity function, vision, and work/productivity. The final instrument contains 49 items measuring these concepts.

c) Subjective Index of Physical and Social Outcomes (SIPSO)

The SIPSO is an outcome tool that was designed to measure people's social integration rather than their abilities per se. It contains 10 items giving an overall score as well as Physical and Social Component scores.

The definition of social integration used in the work initially incorporated environment, activities as well as social integration. However, during test development the items relating to environment were omitted as they failed to fulfil the criteria necessary for inclusion. The authors claim that the main aim of rehabilitation should be to reintegrate the patient into as normal a lifestyle as possible. Interviews with patients and carers were undertaken covering three aspects of their life: (1) pre-stroke, (2) life since stroke and (3) perceptions of change since stroke. Content analysis was undertaken on this data (Trigg et al., 1999). On the basis of the interviews a questionnaire was developed and tested (Trigg and Wood, 2000). The authors claim that the SIPSO measures the ability of an individual to reintegrate to his or her own satisfaction.

An overall score can be calculated together with Physical and Social subscale scores.

d) The Barthel Index

The Barthel Index was originally developed for use in clinical practice as a means of assessing the degree of independence in patients with neurological and neuromuscular limitations. Strictly speaking, the instrument is neither stroke specific nor developed for completion by patients. However, it is widely used in the field of rehabilitation and patient completed versions of the instrument have been developed.

The original Barthel Index consists of ten items, each of which is rated in terms of the patient's ability to undertake the task. Patients are classified into one of dependent, performs task with help and independent. In the original index there were ten areas covered (Bowel control, Bladder control, Grooming, Toilet use, Feeding, Transfer (from bed to chair), Mobility, Dressing, Stairs, Bathing). There have been a number of modifications to this original formulation, including a version with fifteen areas covered called the Modified Barthel Index (Granger et al., 1979), and one developed by Wade and Collin (1988) which uses simplified scoring algorithms.

e) Frenchay Activities Index

The Frenchay Activities Index (FAI) was developed as a means of measuring social activities and lifestyle following stroke, to supplement the more basic functional activities of daily living assessed by measures such as the Barthel Index. The FAI was designed from the outset to be an instrument that would be administered by the clinician to the patient in the clinical interview (Holbrook and Skilbeck, 1983; Wade et al., 1985).

f) Nottingham Extended Activities of Daily Living (ADL) Scale

The Nottingham Extended ADL Scale was developed and evaluated as a questionnaire for postal use (Nouri and Lincoln, 1987). It assesses the ability to carry out functional tasks, such as using public transport, housework, social life and hobbies. Scores in four areas: mobility, kitchen tasks, domestic activities and leisure activities can be added to give a summary score out of 22. Respondents are asked

whether they do the activity rather than if they can do it, in order to assess level of activity rather than capability.

g) London Handicap Scale

The London Handicap Scale (LHS) was developed in response to the need for measures of morbidity to complement mortality statistics in the evaluation of health care interventions and services (Harwood, et al., 1994). Handicap is the disadvantage experienced by an individual patient because of ill-health. The developers adopt a definition of handicap developed by the World Health Organisation and claim that it can be classified according to disadvantages in each of six dimensions: mobility, physical independence, occupation, social integration, and economic self sufficiency. The LHS contains one item for each of these dimensions. A single index score is gained by summing and weighting responses to these items. The measure was designed for use in rehabilitation, hence its inclusion in this review as a stroke specific measure. However, although it has been primarily used in stroke patients it could be used in other serious illness where patients undergo rehabilitation.

STROKE-SPECIFIC INSTRUMENTS:

Table 9.6: Details of stroke-specific patient-reported health instruments

Instrument	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Stroke Impact Scale (SIS); Duncan et al., 1997, Wallace et al., 2002	Strength, Hand Function, ADL/IADL, Mobility, Emotion, Memory, Communication, Social Participation (SIS version 3 contains 59 items in total)	5 point scales	0-100 for all dimensions and aggregate 'Physical Domain'	Interview Self completion
Stroke Specific Quality of Life Scale; Williams et al., 1999	Energy (3), Family roles (3), Language (5), Mobility (6), Mood (5), Personality (3), Self-care (5), Social roles (5), Thinking (3), Upper extremity function (5), Vision (3), Work (3)	5 point scale	Unweighted averages of items per domain (0-5) Overall score 0-60	Interview
Subjective Index of Physical and Social Outcome (SIPSO); Trigg and Wood, 1999, 2000, 2003	Overall score (10) Physical component (5) Social component (5)	5 point scales	Mean score 0-40 (overall) 0-20 for Physical and Social Component scores	Self completion
Stroke adapted 30 item Sickness Impact Profile (30); Straten et al., 1997	Emotional Behaviour (4); Body care and movement (5); Household management (4); Mobility (3); Social Interaction (5); Ambulation (3); Alertness Behaviour (3); Communication (3); Physical component score (11); Psychosocial component score (15); Total score (30)	Dichotomous yes/no responses	0-100 for all dimensions and summary scores	Interview
Barthel Index (10); Mahoney and Barthel, 1965	Bowels (1) Bladder (1) Grooming (1) Toilet use (1) Feeding (1) Transfer (1); Mobility (1); Dressing (1); Stairs (1); Bathing (1)	Categorical: 2-4 options	0-100 (0-20 with simplified scoring)	Measure initially designed for completion by clinician, but interview and self completion versions have been developed
Modified Barthel Index (15); Granger et al., 1979	Drinking from a cup (1) Eating (1) Dressing - upper body (1); Dressing - lower body (1); Putting on brace or artificial limb (1); Grooming (1) Getting in and out of chair (1); Toilet use (1); Getting in and out of tub or shower; Walking 50 yards (1); Walking up/down one flight of stairs (1); If not walking: pushing a wheelchair	Categorical: 2-4 options	-2 - 100 Self care functions: -2 - 53 Mobility: 0-47.	Clinician, interview and self completion
Stroake and Aphasia Quality of Life scale	Language; Thinking; Personality; Energy; Mood; Family Roles; Social Roles; Work; Overall Score	5 point scales	0-5 for all dimensions and summary scores	Interview

Instrument	***Domains (no. items)***	***Response options***	***Score***	***Administration/ Completion (time)***
39 item Stroke and Aphasia Quality of Life Scale	Physical (17); Psychosocial (11); Communication (7); Energy (4) Overall score	5 point scales	0-5 for all dimensions and summary scores	Interview
London Handicap Scale (6)	Handicap (6)	6 options per question	Index of handicap	Interview or self
Frenchay Activities Index (FAI)	Single Index Scores (15) 15 items are Work; Driving, Hobby, Preparing meals, Local shopping, Reading books, Gardening, Washing up, Washing clothes, walking outside for longer than 15 minutes, Light housework, Heavy housework, Household/car maintenance, Social occasions, Travel outings	4 point scales	0 - 45 (or 15 to 60) point Index score Sub dimensions: Domestic Activities; Work and Leisure; Outdoors and Other	Interview Self/proxy completion
Nottingham Extended Activities of Daily Living Scale	Mobility (6); Kitchen Tasks (5); Domestic tasks (5); Leisure activities (6)	4 point scales	Total score, Mobility, Kitchen, Domestic and Leisure scores	Interview Self completion

Table 9.7: Summary of stroke-specific instruments: health status domains

Instrument	**Physical function**	**ADL/Self care**	**Emotions**	**Sleep**	**Social/Interpersonal**	**Cognitive functioning**	**Communication**	**Pain**	**Role Functioning**	**Fatigue**	**Vision**
NEWSQOL (56)	x	x	x	x	x	x	x	x		x	x
SIS	x	x			x		x				
SS-QOL	x	x	x		x		x		x	x	x
SIPSO	x				x						
Barthel Index	x	x									
FAI					x				x		
Nottingham Extended ADL Scale	x	x									
London Handicap Scale									x		
Reintegration to Normal Living Index		x									

STROKE-SPECIFIC PATIENT- REPORTED HEALTH INSTRUMENTS:

a) Stroke Impact Scale (SIS)

The SIS is a relatively recent addition to the battery of measures available to measure stroke outcomes. However, despite this it has been subject to a substantial amount of work evaluating its measurement properties. Seven papers documenting its development and use in the North American context, and one Australian study, were found for this review. To date, no work on the measure has been published in the UK.

Reliability

Internal consistency reliability of the SIS was assessed in a small scale interview survey of patients with mild and severe stroke (Duncan et al., 1999) and found to be high for all eight domains in both groups. This result was broadly substantiated in a larger interview study, except for the Strength subscale where an alpha of 0.63 was gained (n=216). Internal consistency reliability was also assessed for the SIS when administered by telephone and self completion and found to be high for all eight primary dimensions (alpha >0.75) (Duncan et al., 2005). Similarly, in a postal survey the SIS dimensions and the SIS-16 (a subset of items measuring functional ability) were found to have high internal consistency (Edwards and O'Connell, 2003).

Test-retest was undertaken on 25 stroke patients and found to be good (ICC's range 0.7 to 0.92) except for the Emotion dimension (ICC=0.57) (Duncan et al., 1999).

Duncan et al., (2003) evaluated the unidimensionality of dimensions on the SIS. They argued that domains that could not be shown to have unidimensionality would be difficult to interpret. Consequently, they decided to apply the Rasch model to each of the separate dimensions of the SIS. A total of 696 subjects completed the SIS at baseline and/or at follow-up (640 at baseline and 624 three months later). All 1264 SIS questionnaires were entered into the Rasch analysis. Rasch analysis assesses the extent to which items fit a unidimensional model: poor 'fit' statistics suggest items are not tapping a single underlying construct, and is therefore a good test of internal consistency reliability. Rasch analysis can be used to determine whether items fit a unidimensional model, and hence can indicate internal consistency reliability. Very few items were indicated not to 'fit' their proposed domains, one each from the memory, mobility and participation domains. Three items from composite physical domain (created by aggregating the domains of strength, Hand function, ADLs/IADLs and Mobility) had poor in fit statistics.

Edwards and O'Connell (2003) reported that item discriminant validity statistics (i.e. the number of correlations of items in own domains that were significantly higher than correlations with other domains) were adequate for most dimensions of the SIS and were excellent for Strength and Hand Function domains.

Validity
Convergent and discriminant validity of the SIS-16 was supported by correlations with the SIS and a general quality of life measure: the WHOQOL-Bref (WHOQOL Group, 1998).

Discriminant validity was assessed in an interview survey by comparison of SIS mean scores across groups defined by Rankin scores. Six of the eight domains showed significantly different results across scales (Duncan et al., 1999). The authors claim that 'criterion' validity was also assessed against existing measures and showed moderate to good associations with related dimensions on the SF-36, FIM and Barthel Index (Duncan et al., 1999).

In a telephone survey the SIS was found to have superior discrimination between Rankin Scores than either the SF-36V (a modified version of the SF-36) or Functional Independence Measure (Kwon et al., 2004).

In a postal survey of stroke patients the SIS Physical Domain scores and the Aggregate Physical Domain scores had fair to moderate correlations with data FIM Motor scores and the Physical Functioning dimension of the SF-36 gained via telephone interview (Duncan et al., 2002).

The developers also used Rasch analysis to assess the validity of the SIS (Duncan et al., 2003). One of the assumptions behind Rasch analysis is that items in a scale should form a hierarchy of difficulty. When measures are developed using a conceptual hierarchy then the ordering gained by Rasch analysis can be compared to that assumed when the items were initially chosen. Finally Rasch analysis produces an index that indicates the number of distinct strata of persons discerned within each domain: the larger the more distinct levels of functioning can be distinguished in the measure. A total of 696 subjects completed the SIS at baseline and/or at follow up (640 at baseline and 624 three months later). All 1264 SIS questionnaires were entered into the Rasch analysis. In each domain empirical ordering of items by difficulty was consistent with expectations regarding the theoretical ordering of task difficulty. This supports the construct validity of the SIS. Separation indices were calculated for each domain and results were generally good, although floor or ceiling effects were found on memory, emotion, communication and hand function domains.

Responsiveness
Data on change over time is reported in Duncan et al., (1999), and the authors claim that the instrument is responsive to 'ongoing recovery'. The authors suggest that differences of approximately 10-15 points would suggest meaningful change both clinically and subjectively.

Precision
Lai et al., (2003) compared results gained from the SIS Participation Domain and the SIS-16 (measuring physical function) with the SF-36 Social Functioning and Physical Functioning Domains. Rasch analyses indicated that both the SIS-16 and SF-36 Physical Functioning domain both showed a good spread of item difficulty, but the SIS-16 incorporates easier items that are capable of measuring lower levels of physical functioning in patients with severe stroke. Similar analyses of the SIS Participation domain and the SF-36 Physical Functioning domain indicated that the

SIS measures has a wide spread of item difficulty, whereas the SF-36 domain does not. The Social Functioning domain of the SF-36 contains only two items, measuring the same level of item difficulty, leading to severe ceiling effects and consequently an inability to discriminate among more active patients.

Floor or ceiling effects have, however, been found on the measure (Duncan et al., 1999, 2003). In one study floor effects were found for minor stroke patients on all dimensions except Hand Function, whilst only on Emotion was there a floor effect for severe stroke patients (Duncan et al., 1999). However, such results could be argued as supporting the construct validity of the instrument. In another study (Duncan et al., 2003) floor effects were found on the domains of Memory, Emotion and Communication. Floor effects suggest that stroke has had no effects in these areas, which is, of course, a possible explanation for the findings. However, ceiling effects were found on the Hand function dimension, and this suggests some potential measurement limitations on this domain for stroke patients, and the possibility that further 'more severe' items could meaningfully be added.

Acceptability

Duncan et al., (2005) evaluated results from self completion and telephone interviewer administered versions of the SIS in a randomised controlled trial of the two methods of administration. Response rates for mail and telephone were 45% and 69% respectively.

Missing data points were present in the mail version but not in the telephone version. In a mail survey Duncan et al., (2002) reported that non-responders to the SIS had more severe strokes and lower functional status than responders.

Feasibility

The cost of administering the questionnaire by telephone was found to be over twice that of self completion (Duncan et al., 2005). However, Kwon et al., (2004) suggest that such a method may be a practical method of measuring outcomes in community dwelling stroke survivors.

Table 9.8: Developmental and evaluation studies relating to the Stroke Impact Scale (SIS):

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Stroke Impact Scale		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Duncan et al., 1999 USA	33 individuals with minor stroke; 58 with major stroke Mean age 69.2 (minor stroke) and 71.9 years (major stroke) Interview administered	Internal consistency ✓ Test re-test ✓	Construct ✓	✓	✓		
Duncan et al., 2002 USA	125 individuals with stroke Mean age 68.1 years Postal interview		Construct ✓			✓	
Duncan et al., 2003 USA	696 individuals with stroke Age: Mean 68.6 Face to face interview	Internal consistency ✓	Construct ✓		✓		
Duncan et al., 2005 USA	190 individuals with stroke Age: Mean 68.6 RCT of either telephone interview or self completion versions	Internal consistency ✓ Test re-test ✓	Construct ✓			✓	
Edwards and O'Connell 2003 Australia	74 individuals with stroke Age: Mean 58.4 Postal questionnaire survey		Construct ✓				
Kwon et al., 2006 USA	136 individuals with stroke Age: Mean 68.0 Telephone survey		Construct ✓				
Lai et al., 2003 USA	278 individuals with stroke Age (mean) 72.5 Interview survey		Construct ✓		✓		
Nichols-Larsen et al., 2005 USA	213 individuals with stroke Age (mean) 62.1 Interview survey	Internal consistency ✓	Construct ✓				

b) Stroke Specific Quality of Life Scale

Three studies were identified which evaluated the SS-QOL, two based on data gained in North America and one based on UK data.

Reliability
The developers report high internal reliability in all dimensions of the SS-QOL (alpha >= 0.73) (Williams et al., 1999a).

Validity
Scores one month after stroke on the domains of Energy, Family Roles, Mobility, Mood, Personality, Self-care and Work domains were significantly linearly associated with the corresponding scores of the BI, BDI and subscales of the SF-36. However, scores on the Language and Thinking domains were not associated with clinician administered NIH Stroke Scale. The authors suggest this may be because the subjects in their study were largely unaffected by Language and Cognitive problems, though why this finding should not be replicated on the NIHSS is not fully explained. In a regression analysis overall self-reported health related quality of life was associated with SS-QOL domain scores, Barthel Index, NIH Stroke Scale and Beck Depression Index scores, but not with SF-36 scores (Williams et al., 1999b).

Responsiveness
No data available.

Precision
The developers found no evidence for ceiling and floor effects (Williams et al., (1999a).

Acceptability
Hilari and Byng (2001) evaluated the SS-QOL for stroke patients with aphasia as part of study of 80 people with long-term aphasia. They held two focus groups and, as a consequence, amended the form to be more easily completed by patients with aphasia. They amended the instrument so that it was interviewer administered, and simplified the wording of many of the items, and changed the response categories after pilot testing the instrument on 12 patients with aphasia. However, results from amending the SS-QOL to a more 'communicatively accessible' version are based on very small samples and are very preliminary.

Feasibility
No data available.

Table 9.9: Developmental and evaluation studies relating to the Stroke Specific Quality of Life Scale:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
SS-QOL		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Hilari K and Byng S., 2001 UK	Patients with aphasia as a consequence of stroke selected from focus groups (80). Age not specified		Construct ✓			✓	
Williams et al., 1999a USA	Stroke patients (n=32 interviews; n=72 survey) Age: interview sample - not specified; survey sample 61 years)	Internal consistency ✓	Construct ✓				
Williams et al., 1999b USA	Stroke patients (n=71) Mean age = 61 Patients in one of three hospitals		Construct ✓				

c) Subjective Index of Physical and Social Outcome (SIPSO)

The Subjective Index of Physical and Social Outcome is a measure developed and tested in the UK. Three papers were found outlining its development and validation and are included in this review.

Reliability

The developers report high item total correlations for this ten item scale (>0.6). Internal consistency validity was calculated for the overall scale (alpha =0.92) and the Physical Integration (alpha=0.94) and Social Integration (alpha=0.85) (Trigg and Wood, 2003). A small test-retest study (n=31) was undertaken by the developers and intraclass correlation coefficients were found to be high (>0.91 for all SIPSO measures). A further test-retest study (n=128) confirmed these results (Trigg and Wood, 2003).

Kersten et al., (2004) evaluated internal consistency reliability of the SIPSO in a survey of young adults with stroke and found to be very high (overall score alpha= 0.90; 0.92 for Physical Integration subscale and 0.82 for the Social Integration subscale. Test re-test was also found to be good with an intra-class correlation coefficient of 0.96 for the overall score, and 0.94 and 0.95 for the Physical Integration and Social Integration subscales.

Validity

The constructs used by the developers in validating the SIPSO were generated with respect to four other measures: the Barthel Index, the Frenchay Activities Index, the Wakefield Depression Inventory and the Nottingham Health Profile (NHP). It was hypothesised that the results of the SIPSO would correlate with each of these measures so that patients who were better integrated would be more able to perform

basic tasks, have better self assessed health and be less depressed. The SIPSO Physical Scale was most highly correlated with the Barthel Index, Frenchay Activities Index and Mobility on the NHP, suggesting it is tapping some aspect of physical ability. Indeed no significant correlations were found between the Physical Function scale of the SIPSO and dimensions of Emotion, Sleep and Social Functioning on the NHP. The Social Functioning scale of the SIPSO was found to be more highly correlated with the Wakefield Depression Inventory, Emotion and Social Functioning on the NHP (Trigg and Wood, 2003).

In a further validation paper of the SIPSO Trigg and Wood (2003) administered six dimensions of the FLP and SIPSO to122 patients. They hypothesised that the people who displayed better Physical and Social outcomes on the SIPSO would show better Ambulation, Mobility, Recreation, Social Interaction, Emotion and Communication scores on the FLP. All correlations between these SIPSO scores and FLP dimension scores were significant and none fell below 0.45.

Responders with poorer outcomes in terms of 'returning to work' and those reporting physical limitations and problems with their sex lives had poorer SIPSO scores. No associations were found for SIPSO scores and age or sex (Kersten et al., 2004).

Responsiveness
No data available.

Precision
The developers report that the measure shows ‘little ceiling or floor effect’ (Trigg and Wood, 2000). Scores range from 0 to 40 (i.e. across the score band) with an interquartile range of 15-32 a median of 24 and mode of 22 (Trigg and Wood, 2003).

Acceptability
Item completion was high, with missing data highest (7%) for the item 'Since your stroke how independent are you in your ability to move around your local neighbourhood?' (Kersten et al.,2004).

Feasibility
No data available.

Table 9.10: Developmental and evaluation studies relating to the Subjective Index of Physical and Social Outcome (SIPSO):

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
SIPSO		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Kersten et al., 2004 UK	390 individuals with stroke Age: Mean 57.7 Postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓			✓	
Trigg and Wood, 2000 UK	157 patients with stroke Age: not specified Postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓		✓		
Trigg and Wood, 2003 UK	268 patients with stroke Age <64 n=84, 65-74 n=80, Age > 75 n=97 Postal survey	Internal consistency ✓ Test re-test ✓	Construct ✓	✓			

d) The Barthel Index/Modified Barthel Index

The Barthel Index is typically completed by a clinician. However, a number of versions of the instrument exist which are suitable for patient completion. Only studies where the measure has been completed by the patient are included in this review. Five papers based on data from the UK and two papers from the USA are included.

Reliability

The Barthel Index when completed by patients (or by an unspecified number of proxy respondents) and internal reliability was found to be high (alpha=0.83, n=82) (Sadaria et al., 2001). Gompertz et al., (1993, 1994) undertook a small scale evaluation of the test re-test reliability of the Barthel Index in a postal survey (n=21). The mean difference in total score was -0.5 (SD 2.1) out of 20, with 95% CI of -4.6 to 3.6 corresponding to a change in dependence of up to two ADL items. The authors suggest that these results indicate a postal Barthel Index is both practicable and reliable.

Validity

Correlations between the FIM and the Barthel Index have been found to be high (rho=0.97, n=82) (Sadaria et al., 2001). A self completion version of the Barthel Index was found to correlate very highly with the Physical subscale of the Subjective Index of Physical and Social Outcome, r=0.82, p<0.01, n=43) (Trigg and Wood, 2000). King et al., (2005) also reported high levels of association between the SF-12 PCS and an interview administered version of the Barthel Index (rho=0.33, p<0.001).

In an interview based setting the Barthel Index was found to be highly correlated (r=0.76, p<0.001) with the London Handicap Scale, a measure of disadvantage experienced as a result of ill health (Jenkinson et al., 2000).

In another interview based study results on the Barthel Index (Shah modified version, Shah, 1994) were found to be highly correlated with the Nottingham Extended ADL Index both at discharge and follow up.

Responsiveness
Effect sizes indicating an instrument's ability to detect change were found to be high for both the Barthel Index and the FIM (2.2 and 2.4, respectively, n=82) (Sadaria, et al., 2001). However, Jacob-Lloyd (2005) found the Barthel Index (Shah modified version, Shah, et al., 1989).to be insensitive to changes over time in their study of 55 patients, whereas the Nottingham Extended ADL scale detected considerable change.

Precision
Jacob-Lloyd et al., (2005) claimed that the Barthel Index (Shah modified version, Shah et al., 1989) showed signs of ceiling effects in a study of 54 patients with complete data on the measure. However, only 2 respondents gained a score at the ceiling so this claims seems hard to justify.

Acceptability
Gompertz et al., (1994) evaluated a test re-test version of the BI on 21 patients. They do not explicitly state what number responded to the follow up, but claim that the measure is 'practical' for use via postal administered. Jacob-Lloyd (2005) report that 98% of stroke respondents in their survey completed the Barthel Index (Shah modified version, Shah et al., 1989).

Feasibility
No data available.

Table 9.11: Developmental and evaluation studies relating to the Barthel Index

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Barthel/Modified Barthel		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Gompertz et al., 1993 UK	Stroke (21) Mean age: 69 Postal survey	Test re-test ✓					
Gompertz P et al.,1994 UK	Stroke (21) Mean age: 69 Postal survey	Test re-test ✓	Construct ✓				
Jacob-Lloyd et al., 2005 UK	Stroke (55) Mean age 85% over 60 Interview		Construct ✓	✓			
Jenkinson et al., 2000 UK	Stroke (303) Mean age: 74 Interview survey		Construct ✓				
King et al., 2005 USA	Stroke (170) Mean age:53 SF-12 self completed		Construct ✓				
Sadaria KS, et al.,2001 USA	Stroke (82) Age: mean 70.8 Interview	Internal consistency ✓	Construct ✓	✓			
Trigg and Wood, 2000 UK	Stroke (157) Age: not specified		Construct ✓		✓		

e) Frenchay Activities Index

The Frenchay Activities Index was developed in the UK to assess social functioning, and developed for use in the clinical interview. It was always intended responses to the form should originate from the patient. It is widely used and reported in the literature on stroke, although few papers assess its psychometric and measurement characteristics. Five papers were judged suitable for inclusion in this review.

Reliability
Inter-rater reliability was assessed by Piercy et al., (2000). Moderate to high levels of agreement were found between the two raters both at the level of individual items. The Index score was very highly correlated between the two administrations (Spearman's rho=0.93, $p<0.001$, n=61). Similarly, Segall and Schall (1994) found high levels of agreement between two research assistants scoring a videotaped FAI (4 patients and 4 caregivers acting as proxy) (ICC=0.97). They also found good agreement between carers and patients on the FAI (ICC = 0.85 (CI 0.74 - 0.92) n=38). Intraclass correlations for the three subscales were found to be moderate (for the Work and Leisure Scale, ICC=0.59) to good (for the Domestic Activities and Outdoors/Other domains ICC=0.77).

Validity
Construct validity was assessed by Wade et al., (1985). They used to factor analytical techniques and found a high degree of communality for each item confirming the idea the items could be summed to a single score. They also found the FAI score to be highly correlated with Barthel Index score.

Whether calculated from either patient or proxy reports FAI total scores were found to be highly correlated with the Functional Independence Measure (FIM) (rho=0.80 for patients and 0.75 for proxies) (Segall and Schnall, 1994). FAI total score and domain scores had good agreement between patients and proxy assessment, and Segall and Schall suggest the instrument seems appropriate for use with relatives and friends who are primary caregivers for patients with cognitive impairment.

In an interview based setting the Barthel Index was found to be highly correlated ($r=0.73$, $p<0.001$) with the London Handicap Scale, a measure of disadvantage experienced as a result of ill health (Jenkinson et al., 2000).

Responsiveness
No data available.

Precision
No data available.

Acceptability
Segal and Schall (1994) report that the FAI can be completed in approximately 5 minutes either by interview or self completion.

Feasibility
Results on a postal version of the FAI were compared with those gained from interview. Item agreement varied considerably with items relating to social activities

having very low agreement, whilst items relating to work and driving having high levels of agreement. Kappa values (a statistic indicating level of agreement ranged from a low of 0.35 to a high of 1 (perfect agreement). Mean score differences for the two administrations of the Index were small, but masked substantial differences in some instances at the individual level (Carter et al., 1997).

In a small scale study Wade et al., (1985) assessed the extent that different interviewers may have on results from the FAI and found that whilst individual item scores varied considerably, the overall scores were highly correlated ($r=0.80$, $p<0.001$, $n=14$).

Table 9.12: Developmental and evaluation studies relating to the Frenchay Activities Index

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
FAI		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Carter et al., 1997 UK	Stroke patients (n=42) Mean age 71 Postal and interview (home visit)		Construct ✓				Interview versus postal completion
Jenkinson et al., 2000 UK	Stroke patients (n=303) Mean age = 74 Interview survey		Construct ✓				
Piercy et al., 2000	Stroke patients (n=35) and carers (n=24) Mean age (for both patients and carers) 71.1 Interview - home visit	Inter rater ✓	Construct ✓				
Segall and Schall, 1994 USA	Stroke patients (n=38 stroke patient and carer pairs) Mean age 68 (patients) and 54 years (carers) Interview (home visit)	Inter-rater ✓	Construct ✓				✓
Wade et al., 1985 UK	Stroke patients (n=581) Mean age = 72 Interview	Inter-rater ✓	Construct ✓				

f) Nottingham Extended ADL Scale

The Nottingham Extended ADL scale was developed in the UK as an instrument for postal use. Three papers that reported its use and evaluation are included in this review.

Reliability

Test re-test reliability of the Nottingham Extended ADL Scale was found to be high in a small scale (n=21) postal evaluation. Stroke patients were sent a questionnaire pack containing the Nottingham Extended ADL Scale six months after having a stroke and then again 2 weeks later. Results were found to be highly correlated, and the measure gained the best repeatability coefficient of all instruments assessed (including the Barthel Index and the Nottingham Health Profile). Item agreement was also found to be good (Gompertz et al., 1993).

Validity

Gompertz et al., (1994) evaluated the validity of the Extended ADL Scale in a longitudinal study. A total of 361 patients were recruited, but at follow-up only 191 questionnaires were returned at 6 months and 158 twelve months follow-up. High correlations were found between Barthel Score, NHP Physical Mobility, Energy and Pain Scores. However, the authors argue that gender, race and social class, which are independent of mobility, influence scores. Consequently, they suggest that results from the measure may be biased by such confounding variables.

Responsiveness

Gompertz et al., (1994) found that the Extended ADL Scale detected substantial changes between stroke and follow up at one month (effect size = 1.4), and moderate change between one month and six months (effect size = 0.6). However, the measure did not appear to be sensitive to changes between 6 and 12 months, which may indicate insensitivity on the measure or limited changes in patient health. Jacob-Lloyd et al., (2005) suggest that the Nottingham Extended ADL Scale was more sensitive to change that the Barthel Index in their study of 55 patients from discharge to first follow-up appointment. Indeed the measure suggested substantial change over time, whilst the Barthel hardly registered only very modest change, as assessed with the effect size statistic (ES= 0.63 and 0.17 respectively).

Precision

Jacob-Lloyd et al., (2005) suggest that the Nottingham Extended ADL Scale 'showed floor effects at discharge with 50/51 participants scoring below the midpoint and 3 on the minimum score.' However, these results do not seem to suggest serious floor effects, which are usually interpreted as a high proportion of scores at the very extreme range of the scale.

Acceptability

Jacob-Lloyd et al., (2005) found that 51 (98%) of stroke respondents completed the Nottingham Extended ADL suggesting the instrument is acceptable to patients.

Feasibility

No data available.

Table 9.13: Developmental and evaluation studies relating to the Nottingham Extended ADL Index

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Nottingham Extended ADL Scale		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Gompertz et al., 1993 UK	Stroke patients (n=21) Mean age = 69 Postal survey	Test re-test ✓	Construct ✓				
Gompertz et al., 1994 UK	Stroke patients (n=191) Mean age = not reported Postal survey		Construct ✓	✓			
Jacob-Lloyd et al., 2005 UK	Stroke patients (n=55) Age = 85% over 60 Interview survey		Construct ✓			✓	

g) London Handicap Scale

Only two papers evaluating the London Handicap Scale (LHS) were found which were suitable for inclusion in this review.

Reliability

Harwood et al., (1994) undertook a test-retest study on the LHS (n=37). They reported that 'the mean test-retest difference for the group was 0.01, standard deviation 0.09 (limits of agreement was 0.19) and the reliability coefficient was 0.91, implying reasonable agreement between replicate measurements.' Jenkinson et al., (2000) reported high levels of internal consistency reliability on the measure (alpha=0.98).

Validity

Harwood et al., (1994) found predicted high levels of correlation between LHS and the Barthel Index, the Nottingham Extended ADL Score and the NHP Physical Mobility subscale. Similarly Jenkinson et al., (2000) found high levels of correlation between the LHS and the Frenchay Activities Index and the Barthel Index.

Responsiveness

No data available.

Precision

No data available.

Acceptability

Harwood et al., (1994) reported that 71% of respondents to the LHS required help to complete the questionnaire.

Feasibility

Jenkinson et al., (2000) suggest that simple summation of items on the LHS is more straightforward to undertake and provides almost identical information to the more complex weighted scheme, devised by the developers.

Table 9.14: Developmental and evaluation studies relating to the London Handicap Scale

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Nottingham Extended ADL Scale		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Harwood et al., 1994	Stroke patients n=94 Mean age 71 Postal questionnaire	Test re-test ✓	Construct ✓			✓	
Jenkinson et al., 2000 UK	Stroke patients (n=303) Mean age = 74 Interview survey	Internal consistency ✓	Construct ✓				✓

Other instruments identified from the review.

The following table provides an overview of other instruments identified, of either newly developed instruments or single study reporting of measurement properties and/or evaluation.

Table 9.15

Instrument/ reference	*Population (N) Age Method of administration Setting*	*Reliability*	*Validity*	*Responsiveness*	*Precision*	*Acceptability*	*Feasibility*	*Comments* *No other records identified unless stated*
Newcastle Stroke Quality of Life measure (NEWSQOL) **Buck et al., 2004**	Stroke patients (106) Age:70 Interview at home	Internal consistency ✓ Test re-test ✓	Construct ✓		✓	✓		11 domains, 56 items Feelings (6) ADL/self care (8) Cognition (5) Mobility (9) Emotion (4) Sleep (6) Interpersonal relationships (6) Communication (4) ain/sensation (3) vision (2) Fatigue (3)
HSQuale for Young Haeorrhagic Stroke Patients **Hamedani et al., 2001**	Stroke patients (71) Age:44 (62% were 40 years old or less)	Internal consistency ✓ Test re-test ✓	Construct ✓		✓			7 domains, 54 items (not all items contribute to domain scores) List 4 ways stroke has changed your life (1) Overall quality of life (1) General outlook (9) Physical functioning (8) Cognitive functioning (8) Relationships (5) Social and leisure activities (6) Emotional well-being (6) Work and financial status (8) Overall summary question (1) What other ways has stroke affected your quality of life (1)

Instrument/ reference	*Population (N) Age Method of administration Setting*	*Reliability*	*Validity*	*Responsiveness*	*Precision*	*Acceptability*	*Feasibility*	*Comments* *No other records identified unless stated*
Stroke and Aphasia Quality of Life Scale (SAQOL-39) **Hilari et al., 2003**	Stroke patients (93) Age Mean 61.67	Internal consistency ✓ Test re-test ✓	Construct ✓			✓		4 domains, 39 items Physical (17) Psychosocial (11) Communication (7) Energy (4)
Continuity and Discontinuity Following Stroke Scale (CDSS) **Secrest and Zeller, 2003**	Stroke patients (n=55) Mean age 55	Internal consistency ✓ Test re-test	Construct ✓					Continuity (10) Discontinuity (10)
Burden of Stroke Scale (BOSS) **Doyle et al., 2004**	Stroke patients with and without communication disorders (n=135 and 146 respectively) Mean age=63.4	Internal consistency ✓ Test re-test	Construct ✓					Mobility (5) Mobility distress (3) Self-Care (5) Self Care Distress (3) Communication (7) Communication distress (3) Cognition (5) Cognition Distress (3) Swallowing (3) Swallowing distress (3) Social Relations (5) social Relations Distress (3) Energy and Sleep (4) energy and Sleep distress (3) Negative Mood (4) Domain restrictions (1) Positive Mood (4)

Instrument/ reference	*Population (N) Age Method of administration Setting*	*Reliability*	*Validity*	*Responsiveness*	*Precision*	*Acceptability*	*Feasibility*	*Comments* *No other records identified unless stated*
Preference Based Stroke Index (PBSI) **Poissant et al., 2004**	Stroke patients 1. Item generation: 493 patients interviewed six months post stroke 2. Item selection: 91 (mailed survey) 3. Pilot test: 68 (mailed survey) 4. Elicitation of weights: 32 interviews with stroke patients 5. Validation: 91 stroke patients at baseline and 6 months follow up.	Internal consistency ✓ Test re-test	Construct ✓					One item each for: Walking Stairs Physical Activities Recreational activities Work Driving Memory Speech Coping Self-esteem Produces a preference weighted cumulative index score
Schedule for the Evaluation of Individual Quality of Life - Direct Weight (SEIQoL-DW) LeVasseur et al., 2005	Stroke patients with and without communication disorders (n=46) Mean age=63.4	Internal consistency Test re-test	Construct ✓					Respondents nominate and weight their own areas of quality of life affected by their condition
Patient Generated Index Ahmed et al., 2005	Stroke patients with and without communication disorders (n=92) Mean age=63.4	Internal consistency Test re-test	Construct ✓				✓	Respondents nominate and weight their own areas of quality of life affected by their condition

Instrument/ reference	*Population (N) Age Method of administration Setting*	*Reliability*	*Validity*	*Responsiveness*	*Precision*	*Acceptability*	*Feasibility*	*Comments* *No other records identified unless stated*
Stroke and Aphasia Quality of Life instrument (SAQOL) Hilari et al., 2003 UK	n=83 Age mean=61.7 Interview survey	Internal consistency ✓ Test re-test (n=17) ✓	Construct ✓		✓	✓		Areas measured: Language; Thinking; Personality; Energy; Mood; Family Roles; Social Roles; Work; Overall Score
39 item Stroke and Aphasia Quality of Life instrument (SAQOL-39) Hilari et al., 2003 UK	n=83 Age mean=61.7 Interview survey	Internal consistency ✓ Test re-test (n=17) ✓	Construct ✓		✓	✓		Areas measured: Physical (17); Psychosocial (11); Communication (7); Energy (4) Overall score
Reintegration to Normal Living Index Daneski et al., 2003	76 stroke patients Age mean =67.1 Postal survey	Internal consistency ✓ Test re-test	Construct ✓					Total score Daily functioning score Perception of self score

SUMMARY - GENERIC INSTRUMENTS

Six generic instruments (SF-36, SF-12, SF-6D, EuroQol, HUI-3, and NHP) were identified in the review, which had been evaluated with people who have experienced a stroke. For only three of these was there sufficient data to make any informed decisions (SF-36, SF-12, and EuroQol).

The most frequently reported instrument evaluated was the SF-36 with evidence provided for all measurement selection criteria. The evidence for its use in stroke is generally, but not universally, good. For the most part studies reported the instrument domains to have good internal consistency reliability or test-retest reliability. However, there were exceptions, with the General Health dimension failing to fulfil the requirements for this attribute. Test-retest results were found to be acceptable on most dimensions but very low in one study for the Mental Health dimension. The validity of the SF-36 has been examined in concurrent evaluations with widely used rehabilitation measures (e.g. the Frenchay Activities Index and Barthel Index) and found to be good. Empirical evidence supports the internal structure and proposed health domains of the SF-36. There is evidence of responsiveness for the SF-36 domains but evidence suggests it may not perform as well as established instruments used in rehabilitation. Floor and ceiling effects were widely reported, and this may limit the use of the instrument in evaluative studies, especially in those where patients have serious ill health. That said, in order to score on the 'floor' of the domains on the SF-36 one has to have substantially compromised functioning and/or well-being, and any further ability to assess severity may not truly be necessary. Furthermore, a modified version of the SF-36, the SF-36v2, is now available and may reduce such problems in at least the Role Functioning domains, which have been altered to increase precision.

It is perhaps predictable that response rates in those with severe stroke are lower on the SF-36 than in shorter instruments, such as the EuroQol. Evidence for the accuracy of the measure by proxy (e.g. completed by carers, relatives etc) was not good. There was only a limited amount of evidence for the SF-12 in stroke. The two domain measurement model proposed by the developers of the instrument was supported in this patient group. Scores on the two dimensions were generally supported by concurrent evaluations with related measures. Indeed, the SF-12 can be evaluated in relation to a 'gold standard' (the SF-36) and scores between the two measures were found to be very highly correlated. However, there were differences, which could be meaningful, in terms of descriptive statistics and this could suggest inaccuracy in measurement and reduce the validity of the instrument in stroke.

The EuroQol EQ-5D was found to provide reproducible results, and was acceptable, in terms of completion, to more patients than the SF-36. It gave results comparable to other utility measures as well as the Barthel Index. There was some evidence of floor effects in patients defined as 'high functioning' stroke, and hence the instrument is likely to be less sensitive to changes in this group. There is debate as to how the EuroQol should be weighted (with different results gained from stroke patient valuations as to opposed societal valuations). However, as long as the same valuations are used across time and across studies results should be comparable, though whether they should be used in economic analyses remains a matter of debate.

Recommendations
Overall, the SF-36 is the most rigorously evaluated generic instrument although there is mixed evidence to support its application with patients with severe stroke. There is evidence to support the EuroQol as a brief, reasonably acceptable measure of general health in stroke, although both the amount and quality of evaluative material is limited.

SUMMARY – STROKE-SPECIFIC INSTRUMENTS

Fifteen disease specific questionnaires were included in this review, including a number of measures designed principally to assess the influence of rehabilitation. Two 'individualised' measures of outcome were also found to have been used in stroke, but limited information was available for them (PGI and SEIQoL). Consequently seven measures were found to have sufficient information available on their psychometric properties to warrant evaluation (Stroke Impact Scale (SIS), Stroke Specific Quality of Life Scale (SS-QOL), Subjective Index of Physical and social Outcomes (SIPSO), Barthel Index, Frenchay Activities Index, Nottingham Extended ADL scale, London Handicap Scale).

Well established rehabilitation measures fared reasonably in terms of their psychometric properties. The Barthel Index, Frenchay Activities Index and Nottingham Extended ADL Scale were all primarily designed to evaluation rehabilitation outcomes. They are not strictly multi dimensional health outcome/quality of life instruments, but all measure important aspects of health status. The Barthel Index is a measure of independence, and was not initially designed for self completion, but versions of the instrument exist that can be completed by patients. Self completion and interview versions of the instrument have been found to have good reliability and validity, although the sensitivity of the instrument to change is a matter of debate. The Frenchay Activities Index was designed for interview administration, and is a measure of social activities and lifestyle following stroke. The instrument is generally used in interview settings, and there is evidence that the interviewer agreement on items can vary, albeit not dramatically. Available evidence suggests the instrument has good validity, and is amongst the easier measures for stroke patients to complete. The Nottingham Extended ADL Scale has been found to be reliable, and valid in concurrent validation with other instruments. Furthermore it appears sensitive to changes, and appears acceptable to patients. Rehabilitation measures are widely used in the arena of stroke, are well understood by physicians and consequently provide useful and interpretable data. It is hard not to suggest a place for such instruments in evaluation of stroke. One potential criticism of such instruments is that they are typically designed on the basis of clinical judgement and may not reflect issues of importance to patients. Consequently, it seems that such instruments might reasonably be used in conjunction with other quality of life measures.

The London Handicap Scale is perhaps a rather domain specific measure, and the available data is too limited to recommend its widespread use. However, on-going validation of the measure is to be encouraged, although, within stroke at least, the instrument does not appear to be widely used.

In less than a decade there seems to have been an explosion of research into developing stroke specific measures of quality of life and health status. However, this research has not been well coordinated, and consequently a number of instruments exist but few have been subject to on-going evaluation. Researchers seem intent on developing new measures rather than testing existing instruments. Of ten measures documented in this review, only three had sufficient information to be included in the evaluation of instruments (SIS, SS-QOL, and SIPSO). The most data is available for the SIS, but none of the validation research undertaken on the instrument has originated in the UK. The measure has been found to have reasonably good psychometric properties. Internal consistency reliability of all the domains of the measure has been found to be high. Concurrent validations with other health status and rehabilitation instruments have supported the validity of the measure, although some floor and ceiling effects. Unsurprisingly, response rates to the questionnaire have been found to be adversely effected by severity of stroke. Nonetheless results thus far for the SIS are promising, but it does not seem appropriate to recommend such a measure for inclusion in surveys in the UK without there first being some data available on its measurement properties in a UK stroke sample. Similarly the SS-QOL was developed and validated in the USA. Initial assessment of the instrument has been undertaken in the UK, but there is insufficient information to recommend this instrument fully. The SIPSO has been developed in the UK and initial validation of this instrument is very promising. Internal reliability consistency and construct validity have been shown to be good. Furthermore item completion is good and there is little evidence of floor and ceiling effects. However, the instrument is primarily a measure of people's social integration rather than abilities per se, and as a consequence its focus may seem rather narrow. Furthermore, further research is needed to fully evaluate the measure across different levels of stroke severity.

Recommendations

At the present stage of development no single multi-dimensional outcome tool has sufficient information available to recommend it wholeheartedly. Both the SIS and the SIPSO seem highly promising but further data is required for both measures. It seems that, at least for the time being, interview and self completion versions of the Barthel Index, Frenchay Activities Index and Nottingham Extended ADL Scale would appear the most appropriate condition-specific instruments.

REFERENCES

Ahmed S, Mayo NE, Corbiere M, Wood-Dauphinee S, Hanley J, Cohen R. Change in quality of life of people with stroke over time: True change or response shift? *Quality-of-Life-Research:-An-International-Journal-of-Quality-of-Life-Aspects-of-Treatment,-Care-and-Rehabilitation* 2005;**14**:611-27.

Anderson CS, Laubscher S, Burns R. Validation of the Short Form 36 (SF-36) Health Survey questionnaire among stroke patients. *Stroke* 1996; **27**:1812-6.

Bohannon RW, Maljanian R, Lee N, Ahlquist M. Measurement properties of the Short Form (SF)-12 applied to patients with stroke. *International Journal of Rehabilitation Research* 2004; **27**:151-4.

Bohannon RW, Maljanian R, Landes M. Test-retest reliability of short form (SF)-12 component scores of patients with stroke. *International Journal of Rehabilitation Research* 2004; **27**:149-50.

Buck D, Jacoby A, Massey A, Steen N, Sharma A, Ford GA. Development and validation of NEWSQOLR, the newcastle stroke-specific quality of life measure. *Cerebrovascular-Diseases* 2004;**17**:143-52.

Carter J, Mant F, Mant JW, Wade DT, Winner S. Comparison of postal version of the Frenchay Activities Index with interviewer-administered version for use in people with stroke. *Clinical Rehabilitation* 1997; **11**:131-8.

Clark MS, Bond M, The Adelaide Activities Profile: a measure of the life-style activities of elderly people. Ageing Clin Exp Res 1995; **7**: 174-184.

Cote R, Hachinski V, Shurvell B, Norris J, Wolfson C. the Canadian Neurological Scale: A preliminary study in acute stroke. Stroke 1986; **17**: 731-737.

Daneski K, Coshall C, Tilling K, Wolfe CDA. Reliability and validity of a postal version of the Reintegration to Normal Living Index, modified for use with stroke patients. *Clinical Rehabilitation* 2003;**17**:835-9.

Dorman PJ, Waddell F, Slattery J, Dennis M, Sandercock P. Are proxy assessments of health status after stroke with the EuroQol questionnaire feasible, accurate, and unbiased? *Stroke* 1997; **28**:1883-7.

Dorman PJ, Waddell F, Slattery J, Dennis M, Sandercock P. Is the EuroQol a valid measure of health-related quality of life after stroke? *Stroke* 1997; **28**:1876-82.

Dorman PJ, Slattery J, Farrell B, Dennis M, Sandercock P. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. *Stroke* 1998; **29**:63-8.

Dorman PJ, Dennis M, Sandercock P. How do scores on the EuroQol relate to scores on the SF-36 after stroke? *Stroke* 1999; **30**:2146-51.

Doyle PJ, McNeil MR, Mikolic JM, Prieto L, Hula WD, Lustig AP *et al*. The Burden of Stroke Scale (BOSS) provides valid and reliable score estimates of functioning and well-being in stroke survivors with and without communication disorders. *Journal-of-Clinical-Epidemiology* 2004;**57**:997-1007.

Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The Stroke Impact Scale version 2.0: evaluation of reliability, validity, and sensitivity to change. *Stroke* 1999; **30**:2131-40.

Duncan PW, Lai SM, Keighley J. Defining post-stroke recovery: Implications for design and interpretation of drug trials. *Neuropharmacology* 2000; **39**:835-41.

Duncan PW, Reker DM, Horner RD, Samsa GP, Hoenig H, LaClair BJ *et al.,* Performance of a mail-administered version of a stroke-specific outcome measure, the Stroke Impact Scale. *Clinical Rehabilitation* 2002; **16**:493-505.

Duncan PW, Bode RK, Min-Lai S, Perera S. Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives of Physical Medicine and Rehabilitation* 2003; **84**:950-63.

Duncan P, Reker D, Kwon S, Lai S, Studenski S, Perera S et al.,. Measuring stroke impact with the Stroke Impact Scale: telephone versus mail administration in veterans with Stroke. *Medical-Care* 2005; **43**: 507-15

Edwards B, O'Connell B. Internal consistency and validity of the Stroke Impact Scale 2.0/SIS 2.0 and SIS-16 in an Australian sample. *Quality of Life Research* 2003; **12**:1127-35.

Goldstein LB, Lyden P, Mathias SD, Colman SS, Pasta DJ, Albers G *et al*. Telephone assessment of functioning and wellbeing following stroke: is it feasible? *Journal of Stroke and Cerebrovascular Diseases* 2002;**11**:80-7.

Gompertz P, Pound P, Ebrahim S. The reliability of stroke outcome measures. *Clinical Rehabilitation* 1993; **7**:290-6.

Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. *Clinical Rehabilitation* 1994; **8**:233-9.

Gompertz P, Pound P, Ebrahim S. Validity of the Extended Activities of Daily Living scale. *Clinical Rehabilitation* 1994; **8**:275-80.

Granger CV, Albrecht GL, Hamilton BB. (1979) Outcome of comprehensive medical rehabilitation: Measurement by PULSES Profile and the Barthel Index. Archives of Physical Medicine and Rehabilitation 1979; **60**: 145-54.

Grootendorst P, Feeny DH, Furlong WJ. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. *Medical Care* 2000; **38**:290-9.

Hackett ML, Anderson CS, House AO. Interventions for treating depression after stroke. *The Cochrane Library/Cochrane Database Systematic Review* 2004;CD003437.

Hagen S, Bugge C, Alexander H. Psychometric properties of the SF-36 in the early post-stroke phase. *Journal of Advanced Nursing* 2003; **44**:461-8.

Hamedani AG, Wells CK, Brass LM, Kernan WN, Viscoli CM, Maraire N *et al.,.,.* A quality of life instrument for young hemorrhagic stroke patients. *Stroke* 2001; **32**:687-95.

Harwood RH, Gompertz P, Ebrahim S. Handicap one year after a stroke: validity of a new scale. *Journal of Neurology, Neurosurgery and Psychiatry* 1994; **57**:825-9.

Harwood RH, Ebrahim S. The validity, reliability and responsiveness of the Nottingham Extended Activities of Daily Living scale in patients undergoing total hip replacement. *Disability and Rehabilitation* 2002; **24**:371-7.

Hilari K, .Byng S. Measuring quality of life in people with aphasia: the Stroke-Specific Quality of Life Scale. *International Journal of Language and Communication Disorders* 2001; **36**:86-91.

Hilari K, Byng S, Lamping DL, Smith SC. Stroke and Aphasia Quality of Life Scale-39/SAQOL-39: evaluation of acceptability, reliability, and validity. *Stroke* 2003; **34**:1944-50.

Hobart JC, Williams LS, Moran K, Thompson AJ. Quality of life measurement after stroke: uses and abuses of the SF-36. *Stroke* 2002; **33**:1348-56.

Holbrook M, Skilbeck CE. An activities index for use with stroke patients. Age and Ageing 1983; **12**: 166-70.

Jacob-Lloyd HA, Dunn OM, Brain ND, Lamb SE. Effective measurement of the functional progress of stroke clients. *British-Journal-of-Occupational-Therapy* 2005;**68**:253-9.

Jenkinson C. The SF-36 physical and mental health summary measures: an example of how to interpret scores. Journal of Health Services Research and Policy 1998; **3**: 92-96.

Jenkinson CP, Mant JW, Carter J, Wade DT, Winner S. The London Handicap Scale: a re-evaluation of its validity using standard scoring and simple summation. *Journal of Neurology, Neurosurgery and Psychiatry* 2000; **68**:365-7.

Kersten P, George S, Low J, Ashburn A, McLellan L. The Subjective Index of Physical and Social Outcome: its usefulness in a younger stroke population. *International-Journal-of-Rehabilitation-Research* 2004; **27**:59-63.

King JTJ, Kassam AB, Yonas H, Horowitz MB, Roberts MS. Mental health, anxiety, and depression in patients with cerebral aneurysms. *J Neurosurg.* 2005;**103**:636-41.

Kwon S, Hartzema AG, Duncan PW, Lai SM. Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. *Stroke* 2004; **35**:918-23.

LeVasseur SA, Green S, Talman P. The SEIQoL-DW is a valid method for measuring individual quality of life in stroke survivors attending a secondary prevention clinic. *Quality-of-Life-Research* 2005;**14**:779-88.

Lai, SM, Perera, S, Duncan, PW, Bode, R. Physical and social functioning after stroke: Comparison of the stroke impact scale and SF-36. *Stroke* 2003; **34**:448-493.

Lai SL, Guo XF, Liang WX. Preliminary study on outcome assessment system of treatment of stroke. *Zhongguo Zhong.Xi.Yi Jie.He Za Zhi* 2004; **24**:197-201.

Lim LLY, .Fisher JD. Use of the 12-item Short-Form (SF-12) Health Survey in an Australian heart and stroke population. *Quality of Life Research* 1999; **8**:1-8.

McPherson K, Myers J, Taylor WJ, McNaughton HK, Weatherall M. Self-valuation and societal valuations of health state differ with disease severity in chronic and disabling conditions *Med.Care* 2004;**42**:1143-51.

Nouri FM, Lincoln NB. An extended activities of daily living index for stroke patients. Clinical Rehabilitation 1987; **1**: 301-305.

Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S. Factors influencing stroke survivors quality of life during subacute recovery *Stroke* 2005; **36**: 1480-1484.

Nunnally, J. Psychometric theory (2nd ed.). New York: McGraw-Hill Book Co.

O'Mahony PG, Rodgers H, Thomson RG, Dobson R, James OFW. Is the SF-36 suitable for assessing health status of older stroke patients? *Age and Ageing* 1998; **27**:19-22.

Patrick D, Peach H. Disablement in the Community. Oxford: Oxford University Press, 1989.

Patel MD, Tilling K, Lawrence E, Rudd AG, Wolfe CD, McKevitt C. Relationships between long-term stroke disability, handicap and health-related quality of life. Age Ageing. 2006 May;**35**(3):273-9.

Pickard AS. 'Replicability of SF-36 summary scores by the SF-12 in stroke patients' - erratum. *Stroke* 1999; **30**:1737.

Pickard AS, Johnson JA, Penn A, Lau F, Noseworthy T. Replicability of SF-36 summary scores by the SF-12 in stroke patients. *Stroke* 1999; **30**:1213-7.

Pickard AS, Johnson JA, Feeny DH. Responsiveness of generic health-related quality of life measures in stroke. *Quality-of-Life-Research* 2005;**14**:207-19.

Piercy M, Carter J, Mant JW, Wade DT. Inter-rater reliability of the Frenchay Activities Index in patients with stroke and their carers. *Clinical Rehabilitation* 2000; **14**:433-40.

Poissant, Lise. The development of a Preference-Based Health Index for stroke 1002. Dissertation-Abstracts-International:-Section-B:-The-Sciences-and-Engineering 64(11-B), 5836. 2004.

Polsky D, Willke RJ, Scott K, Schulman KA, Glick HA. A comparison of scoring weights for the EuroQol derived from patients and the general public. *Health Economics* 2001; **10**:27-37.

Rubenach S, Anderson CS, Laubscher S. The Short Form-12 by telephone as a measure of health-related quality of life after stroke. *Age and Ageing* 2000;**29**:553-4.

Sadaria KS, Bohannon RW, Lee N, Maljanian R. Ratings of physical function obtained by interview are legitimate for patients hospitalized after stroke. *Journal of Stroke and Cerebrovascular Diseases* 2001; **10**:79-84.

Secrest JS,.Zeller R. Measuring continuity and discontinuity following stroke. *Journal of Nursing Scholarship* 2003;**35**:243-7

Segal ME, .Schall RR. Determining functional/health status and its relation to disability in stroke survivors. *Stroke* 1994; **25**:2391-7.

Shah S, Vanclay F, Cooper B (1989) Improving the sensitivity of the Barthel Index for stroke rehabilitation. Journal of Clinical Epidemiology 1989; **42** (8): 703-709

Trigg R, .Wood VA. The Subjective Index of Physical and Social Outcome/SIPSO: a new measure for use with stroke patients. *Clinical Rehabilitation* 2000; **14**:288-99.

Trigg R, Wood V, Hewer R. Social integration after stroke: the first stages in the development of the Subjective Index of Physical and Social Outcome (SIPSO). Clinical Rehabilitation 1999; **13**: 341-353.

Trigg R, .Wood VA. The validation of the Subjective Index of Physical and Social Outcome/SIPSO. *Clinical Rehabilitation* 2003; **17**:283-9.

Wade DT, Collin C. The Barthel ADL Index: a standard measure of disability? International Disability Studies 1988; **10**: 64-7.

Wade DT, Legh SJ, Langton HR. Social activities after stroke: measurement and natural history using the Frenchay Activities Index. *International Rehabilitation Medicine* 1985; **7**:176-81.

WHOQOL Group. The World Health Organisation Quality of Life Assessment (WHOQOL). Development and general psychometric properties. Soc Sci Med 1998; **46**: 1569-85

Williams LS, Weinberger M, Harris LE, Clark DO, Biller J. Development of a stroke-specific quality of life scale. *Stroke* 1999a; **30**:1362-9.

Williams LS, Weinberger M, Harris LE, Biller J. Measuring quality of life in a way that is meaningful to stroke patients. *Neurology* 1999b; **53**:1839-43.

Chapter 10: Patient-reported Health Instruments: Carer Impact

The impact of a disease on a patient is an increasingly important outcome measure in medicine and healthcare. Issues such as quality of life are now widely used in clinical trials and in patient management for assessing morbidity and the impact of treatment. For a long time, studies focused almost exclusively on changes in the impact of ill health on patients, but increasing attention is now being paid to the impact on carers of patients with chronic diseases.

Carers (or caregivers) play an important role in the care of chronically ill patients, as the number of people with chronic illnesses is increasing and informal and community care outside of acute services is increasingly encouraged. Carers tend to be family members (often the spouse) or friends, who are called upon to provide significant and continuous support to the person with ill health. It is being increasingly recognized that caring for someone with ill health poses challenges and can represent a stressful and difficult situation to the carer with adverse physical and psychological outcomes for the caregiver.

Two different approaches have been used to study carer burden. The first approach, which is an indirect approach, uses generic instruments as proxy measures such as the SF-36. Generic instruments have usually been extensively tested, although not necessarily in the carer population. They provide a broad perspective of health, but they do not give an insight into carers' specific problems. The second approach, which is direct, investigates the carers' experience, focusing specifically on the content of carers' experiences. It uses either instruments that have been developed for carers generally (hereafter referred to as general carer instrument) or instruments that have been developed for people caring for a person with a specific condition (hereafter referred to as disease-specific carer instrument). An example of a general carer instrument is the Carer Strain Index. These instruments provide a more specific measure of the carer-specific burden. However, these instruments are not specific to a particular disease group, and as such may not capture all the relevant issues for a person caring for a patient with a particular condition. Thus, a number of disease-specific carer instruments have also been developed, for example, the Parkinson's Impact Scale (PIS).

Instruments that have been used to assess carer burden also vary in terms of their dimensions, with some instruments investigating multiple dimensions (including for instance, physical health, psychological health, social roles), and other instruments being dimension-specific (e.g. fatigue, depression). Caregiver well-being has traditionally been considered from a deficit perspective and little attention has been paid to positive aspects of caregiving (Berg-Weger and Tebb 1998) and increasingly some instruments also focus on positive aspects of caregiving.

This review reports the psychometric properties of generic and carer-specific instruments that have been used in people who care for people with ill health. This review does not discuss dimension-specific or disease-specific instruments. Furthermore, the focus of this review is on caregiving for adults with ill health, not for children (either healthy or with ill health) or healthy elderly. The review only includes articles published in English with data from English speaking populations (UK, USA, Canada, Australia and New Zealand).

RESULTS: Patient-reported Health Instruments: Carer impact

Search terms and results: identification of articles

At the time of the review, the PHI database contained 12,000+ records (up to June 2005). Search results are detailed in table 10.1. When assessed against the review inclusion criteria, 44 articles were retrieved and reviewed in full. Of these, 26 articles were included in the review.

Table 10.1

Source	**Results of search**	**No. of articles considered eligible**	**Number of articles included in review**
PHI database: original search (up to June 2005) ***Total number= 12.562***	129	44	26
Supplementary search	-	-	49
TOTAL	-	-	**75**

Supplementary searches included hand-searching of titles from 2004 to 2006 of the following key journals:

-Health and Quality of Life Outcomes
-Medical Care
-Quality of Life Research

Further searches were conducted within the bibliography and using Pub Med per instrument up to September 2006.

Identification of instruments

Five indirect measures in the form of generic health instruments were included in the review, together with 26 general carer instruments. The developmental and evaluative studies relating to the generic health instruments reviewed are listed in Tables 10.2 to 10.6. Those relating to general carer instruments are shown in Tables 10.7 to 10.17. Table 10.18 includes examples of carer disease-specific instruments.

RESULTS: GENERIC INSTRUMENTS (INDIRECT APPROACH)

Five generic instruments were identified which were evaluated for use to assess carer impact. Full details of the development, domains and scoring methods are detailed in Chapter 3.

The following instruments measurement properties are reported:

a) SF-36 and SF-12
b) Health Utilities Index Mark 2
c) Reintegration to Normal Living Index
d) Ferrans and Power Quality of Life Index
e) General Health Questionnaire

a) SF-36 and SF-12

Reliability

Good internal consistency for the SF-36 overall was reported in studies by Jenkinson et al., (2000); Berg-Weger et al., (2003) and for the Physical Component Subscale (PCS) and Mental Component Subscale (MCS) in a study by Clark et al., (2004). One study found adequate internal consistency for the other subscales (Berg-Weger et al., 2003)., whereas another found good internal consistency for the different SF-36 subscales (Cameron et al., 2006b).

The SF-12v2 has been found to have weak internal consistency for a sample of carers of dementia patients (McConaghy and Caltabiano 2005).

Validity

Convergent and discriminant validity

Depression, measured by the Centre for Epidemiologic Studies-Depressed Mood Scale, was found to be significantly related to the Physical Health and Mental Health domains of the SF-36 (Berg-Weger et al., 2003). In the same study, anxiety (assessed by the Self-rating Anxiety Scale) was also significantly negatively related to Physical Health and Mental Health, and physician's visits. Visits to mental health professionals were only significantly and negatively related to The Mental Health summary score. There was also moderate discriminative validity, as alternative mental heath measures correlated more strongly with the Mental Health subscale than with the Physical Health subscale. On the other hand, alternative measures of physical health correlated more strongly with the Physical Health subscale.

Internal validity

Factorial analysis supported the original structure of the SF-36 in a study by Berg-Weger et al., (2003).

Predictive validity

One study found support for the predictive validity of the SF-36, with particularly the Vitality Scale being a predictor of stroke carer stress (Smith et al., 2004).

Socio-demographic variables

The scores on the SF-36 for carers have been found to be below those for the general population in studies by Jenkinson et al., (2000) and Cameron et al., (2006b). The scores on the SF-12 have also been found to be slightly below the general norms (Clark et al., 2004). In a further study, carer scores were reported to be lower than population norms on the Energy and Vitality scales (Smith et al., 2004).

Generic carer instruments

The PCS and MCS has strong correlations with the Caregiver Strain Index (CSI), supporting construct validity in a study by Jenkinson et al., (2000).

Responsiveness

No data available.

Precision

Some floor effects have been found for the SF-36 for carers for the Role Physical (19.3%) and Role Mental (23.4%), as well as some ceiling effects (50.6% and 49.0% respectively) (Jenkinson et al., 2000).

Acceptability

No data available.

Feasibility

The survey was administered by telephone interviews in a study by Berg-Weger et al., (2003), reporting 30 minutes completion time.

Table 10.2: Evaluative studies relating to the SF-36 when completed by carers of people with ill health

Study/ Country	**Population (N) Age (years) Method of administration Setting**	**Measurement and Practical properties**					
SF-36		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Berg-Weger et al., 2003 **USA**	Adult daughters who had been primary caregivers to a parent with Alzheimer Disease or a related disorder, who had died at least 12 months before the study (102) Age 57 years Telephone interview Alzheimer Association members	√	√ Factorial validity Convergent and discriminant validity				
Clark et al., 2004 **USA**	Family caregivers of stroke survivors (132) SF-36v2 mental and psychical scales Age 56.7 Sampled from a sample of a national, multi-site clinical trial Interviewer administered	√					
Smith et al., 2004 **UK**	Carers of stroke patients (90) and stroke patients Age 57.8 Patients identified from 2 hospital stroke registers		√ Predictive				
Jenkinson et al., 2000 **Multi-national including the UK**	Carers of patients with amyotrophic lateral sclerosis (415) Age 55.1 years Carers of patients recruited through 74 clinical sites throughout Europe Self-administered	√	√			√	
Cameron et al., 2006b **Canada**	Caregivers of patients with acute respiratory distress syndrome	√					
SF-12v2							
McConaghy and Caltabiano 2005 **Australia**	Carers of people with dementia (42) Age 62.0 years Self-completion questionnaire Self-administered or face to face interviews	√					

b) HUI 2

Reliability

No data available.

Validity

One study evaluated criterion validity of HUI 2 in carers of patients with Alzheimer Disease (Bell et al., 2001) by comparing HUI 2 to a caregiver time questionnaire, a caregiver burden instrument and the SF-36. It was found that the HUI 2 may not adequately capture differences in the burden of caregivers for patients with Alzheimer Disease.

Responsiveness/ Precision/ Acceptability/Feasibility

No data available.

Table 10.3: Evaluative studies relating to the HUI 2 when completed by carers

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Bell et al., 2001 Canada	679 caregivers of individuals with Alzheimer Disease 63 years Interviewer administered questionnaire		√				

c) RNLI

Reliability

Good internal consistency has been found for the RNLI in a study by Bluvol and Ford-Gilboe (2004).

Validity/ Responsiveness / Precision/ Acceptability

No data available.

Feasibility

The questionnaires (the RNLI, which has 11 items, plus 2 more questionnaires) took 30-40 minutes to complete in a study by Bluvol and Ford-Gilboe (2004).

Table 10.4: Evaluative studies relating to the RLNI when completed by carers

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Bluvol and Ford-Gilboe 2004 **Canada**	Carers (40) and stroke patients (40) Age carers 66.2 Self-completion questionnaire	√					√

d) FPQLI

Reliability

High internal consistency was found for the Total instrument and moderate to high alphas for the Life domains (Weitzner et al., 1997).

Validity

Socio-demographic variables

Caregiver age was significantly correlated with the Health/functioning and Psychological/spiritual domains, as well as the Total score in a study by Weitzner et al., (1997).

Responsiveness/ Precision/ Acceptability/ Feasibility

No data available.

Table 10.5: Evaluative studies relating to the FPQLI when completed by carers

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Weitzner et al., 1997 **USA**	Caregiver of cancer patient (22) Age51.7 Interviewer administered	√					√

e) GHQ

Reliability

No data available

Validity

Discriminative validity

There was a non-significant trend in GHQ Total scores and Depression subscales scores to be higher for carers using Admiral Nurse (AN) teams vs. carers who did not (Woods et al., 2003). On follow-up, a significant difference was found on the Anxiety and Insomnia subscale, where outcome was better for the AN group. Another study showed that carers of dementia patients showed higher levels of distress as measured by GHQ than carers for patients with depression (Rosenvinge et al., 1998). Furthermore, significant differences in GHQ scores have been found between carers of people with anorexia and psychosis (Treasure et al., 2001). GHQ scores have also been found to differ in carers of people with a head injury according to different time intervals post-injury. The GHQ scores were higher for carers of people with a recent head injury, which indicates greater burden in this group (Sander et al., 1997).

Predictive validity

Coping style has been found to contribute significantly to GHQ score variance, with emotion-focused coping being related to GHQ scores in a study by Sander et al., (1997). Furthermore, coping accounted for more of the GHQ variance than disability scores.

Socio-demographic variables

Gender has been found to have a significant effect on GHQ scores, but neither race nor relationship to the injured person had a significant effect (Sander et al., 1997).

Dimension-specific variables

Strong positive correlations were found between the GHQ and the Relatives Stress Scale (Draper et al., 1992).

Responsiveness

The GHQ-28 has been shown to be responsive to change in a study using cognitive behavioural therapy in carers of Parkinson's disease patients. Both the Total score and the scores for 3 of the sub-scales decreased in response to the intervention (Secker and Brown 2005). Both conventional and AN services led to lower GHQ scores overall and 2 of the 4 subscales over an 8-month period (Woods et al., 2003).

Precision/ Acceptability/ Feasibility

No data available.

Table 10.6: Evaluative studies relating to the GHQ when completed by carers

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Draper et al., 1992 **Australia**	Co-resident carers of dementia (51) and stroke patients (48) Age stroke 72.0 and dementia 76.0 Self-completion		√				
Sander et al., 1997 **USA**	Carers of patients with head injury (60) 3 groups corresponding to 3 post-injury intervals: early (26), intermediate (21) and long-term (22) Age 'early' 39.1, 'intermediate' 45.8 and 'late' 45.1 Self-completion		√				
Rosenvinge et al., 1998 **UK**	Carers of patients with dementia (32) or depression (25) Age 67.8 for dementia carers 66.8 for depression carers Interviewer administered		√				
Treasure et al., 2001 **UK**	Carers of patients with anorexia (71) or psychosis (68) Interviewer administered		√				
Woods et al., 2003 **UK**	Carers of people with dementia (128, of which 55 used an admiral nurse (AN) service and 73 did not (comparison group)) Age 62.4 for AN group and 58.8 for comparison group Interviewer administered		√	√			
Secker and Brown 2005 **UK**	Carers of patients with Parkinson's (30) Age 59.1 for treatment group (n=15) and 58.8 for control group (n=15)			√			

RESULTS: DIRECT MEASURES OF CARER IMPACT

Eight instruments that investigate the carers' general burden have been identified. Full details of the development, domains and scoring methods are detailed in Tables 10.7 and 10.8.

The following instruments measurement properties are reported:

a) Appraisal of Caregiving Scale

b) Bakas Caregiver Outcomes Scale

c) Caregiver Burden Inventory

d) Caregiveing Burden Scale

e) Caregiver Impact Scale

f) Caregiver Strain Index

g) Caregiver Well-Being Scale

h) Zarit Burden Interview

a) Appraisal of Caregiving Scale (ACS)

The ACS has been developed in the USA with carers of cancer patients receiving radiotherapy. The ACS is a 53-item instrument that measures the meaning of illness-caregiving situation in terms of the intensity of four dimensions (Harm/loss, Threat, Challenge and Benign).

b) Bakas Caregiver Outcomes Scale (BCOS)

The BCOS was developed to measure changes in caregiving outcomes. The BCOS is a unidimensional scale based on 10 items and addresses changes in caregiving social functioning, subjective well-being and physical health. It was first developed and evaluated in carers of stroke survivors in the USA.

c) Cargiver Burden Inventory (CBI)

The CBI is a 25 item instrument with 5 subscales that was developed in carers of confused or disoriented older people in Canada. The CBI aims to give a reading of caregivers' feelings and a picture of the carers' responses to the demands of caregiving.

d) Caregiver Appraisal Scale (CAS)

The CAS was designed as a 47-item interview questionnaire for caregivers of disabled elderly. The CAS has five domains of caregiving appraisal: Caregiving satisfaction, Perceived caregiving impact, Caregiving mastery, Caregiving ideology and Subjective caregiving burden.

e) Caregiver Impact Scale (CIS)

The CIS was developed in Canada in carers of cancer patients and is based on the Illness Intrusiveness Rating Scale by (Devins et al., 1983). The CIS assesses the extent to which providing care interfered with the caregiver's participation in 14 domains of lifestyle (such as health, employment, recreation).

f) Caregiver Strain Index (CSI)

The CSI was developed in the USA to measure caregiver reactions, including perception and emotional feeling with regards to caregiving. It includes 13 items and was developed in caregivers of patients, aged 65 or over, who had recently been hospitalized for hip surgery or heart problems. At a later stage it was modified, to include a 'sometimes' response category, rather than just the 'yes/no' response options and some items were rephrased.

g) Caregiver Well-Being Scale

The Caregiver Well-Being Scale was developed in the USA in a sample of caregivers of older adults with dementia, caregivers of children with developmental problems and caregivers of healthy children who were younger than 12 years of age. The scale has also been evaluated in a sample of caregivers of chronically ill patients. It includes 45 items, with 2 subscales: Basic Human Needs and Activities of Daily Living. The scale identifies the caregivers' strengths in meeting their basic needs and daily activities.

h) Zarit Burden Interview (ZBI)

The Zarit Burden Interview assess the degree to which a caregiver perceives their caregiving responsibilities to have a negative effect on their health, personal and social life, finances and emotional well-being. Different versions of the ZBI are available, the 22-item full version and the 12-item short version. The ZBI was developed and tested in carers of patients with dementia in the USA.

Table 10.7: General carer instruments

Instrument (no. items)	*Domains (no. items)*	*Response options*	*Score*	*Administration/ Completion (time)*
Appraisal of Caregiving Scale (53)	4 domains harm/loss (15 items) threat (15 items) challenge (15 items) benign (8 items)	5 point Likert Scale 1= very untrue, 5= very true	Higher scores on each subscale represent greater intensity of the appraisal dimension	Self-completion 20 minutes
Bakas Caregiving Outcomes Scale (10)	Unidimensional	7-point Likert scale 1=changed for the worst, 7=changed for the best		Self-completion or interviewer administered
Caregiver Burden Inventory (24)	5 domains: developmental, physical, social, emotional burden and time dependence	5 point Likert scale 0=strongly disagree, 4=strongly agree	Items for each domain are summed. Domain scores range from 0-20, except for physical burden (0-16). For physical burden the summed score is multiplied by 1.25 to give an equivalent score out of 25	Self-completion or interviewer administered
Caregiving Impact Scale	14 domains of caregivers' lifestyle	7-point Likert scale 1=not very much, 6=a lot	Summation of items with higher scores indicating higher interference	Self-completion or interviewer administered
Caregiving Appraisal Scale (original 47, later 35)	5 domains (47 items): caregiving satisfaction, perceived caregiving impact, caregiving mastery, caregiving ideology and subjective caregiving burden. 4 domains (35 items): perceived burden (15), caregiver relationship (11) satisfaction, caregiver ideology (5) and caregiving mastery (4)	Self-completion questionnaire uses 5 point scale 1= strongly disagree and 5=strongly agree	For 35 item instrument: Scores calculated by summing individual domain scores, using reversed scoring for certain items.	Interview or self-completion
Carer Strain Index (CSI) (13 items)	Unidemensional		0-100 (lowest to highest level of strain)	Self-completion or interviewer administered 15-45 minutes
Caregiver Well-Being Scale (45 items)	Basic human needs (4 factors and 22 items) Activities of daily living (5 factors and 23 items)			Self-completion
Zarit Burden Interview (Original 29, full version 22 items, short version 12 items)	For 12 item short version Personal strain (9) Role strain (3)	5-point Likert style 0=never and 4=nearly always	0-88 with higher scores a greater burden	Interviewer administered

Table 10.8 summarizes the domains included in the different instruments. However, some general carer instruments do not include these domains as such, but include items that reflect these domains. For example, the CSI is unidimensional. Another example is the CAS which does not include a 'social well-being' domain, but within the 'perceived burden' domain several items relate to social well-being such as 'my social life has suffered' or 'I feel isolated and alone'. Other domains found in general caregiving instruments are not reflected in the health status domains by Fitzpatrick et al. (1998), such as the domain of 'caregiving mastery' of the CAS, which includes items on how well the carer copes with caring or 'time-dependence burden of the Caregiving Impact Scale, which describes burden due to restrictions of caregivers time. Also, some domains (symptoms, cognitive function and treatment satisfaction) from Fitzpatrick et al. (1998) are of less relevance for carer instruments.

Table 10.8: Summary of carer-specific instruments: health status domains ***(after Fitzpatrick et al., 1998)***

Instrument	***Instrument domains***								
	Physical function	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**	**Treatment satisfaction**
Appraisal of Caregiving Scale	x			x	x		x		
Bakas Caregiving Outcomes Scale (10)			x	x	x				
Caregiving Appraisal Scale (47/35)				x	x			x	
Caregiver Burden Inventory (24)	x			x	x				
Caregiving Impact Scale (14)			x		x		x		
Carer Strain Index (13)	x			x	x		x		
Caregiver wellbeing scale (45)	x			x	x		x	x	
Zarit Burden Interview (29, 22, 12 or 4)			x	x	x		x		

RESULTS: CARER IMPACT

a) Appraisal of Caregiving Scale (ACS)

Reliability

Two studies have found the ACS sub-scales to be internally consistent (Oberst et al., 1989; Carey et al., 1991).

Validity

Strong correlations were found between the Harm/loss and Threat subscales, indicating that they may represent the same construct (Oberst et al., 1989). Also, the high correlation between the challenge and benign subscales represent a problem.

Socio-demographic variables

Each of the four sub-scales was related to at least one other caregiver variable (Oberst et al., 1989). Harm/loss scores were correlated with the carer's level of education, social status and health status. Threat scores were correlated with the carer's level of education and social status. Challenge and benign scores were correlated with caregiver age. Benign scores were related to the carer's perception of the illness (as more or less serious).

Patient variables

Correlations were also found with various patient variables (Oberst et al., 1989). Harm/loss scores for carers were related to the length of time patients received radiation. The carer's relationship to the patient was also related to the benign subscale, with those caring for a parent perceiving the situation as less benign than those caring for a spouse or others.

Table 10.9: Developmental and evaluation studies relating to the Appraisal of Caregiving Scale:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Oberst et al., 1989	Family members of cancer patients receiving radiotherapy.(47) Age 53.3 USA Self-completion	√	√				
Carey et al., 1991	Family caregivers of patients receiving outpatient chemotherapy (49) USA Self-completion	√					

b) Bakas Caregiver Outcomes Scale

Reliability

Good internal consistency was found for both the 10 and 12 item BCOS in a study by Bakas and Champion (1999). High internal consistency for the 10-item BCOS was also found in two further studies (Bakas et al., 2004; Bakas et al., 2006) . One study reported finding good test-retest reliability (0.66 for the 15-item BCOS and 0.68 for the 10-item BCOS) (Bakas et al., 2006). It was not clear whether the findings of test-retest reliability refer to group or individual comparisons.

Validity

Internal validity

Two studies report evidence on construct validity, by using factor analysis that supported unidimensionality of the BCOS (Bakas and Champion 1999; Bakas et al., 2006).

Generic health status

Significant weak to moderate correlations with LIFE-3 and with the SF-36 subscales were found in a study by Bakas and Champion (1999) and a significant weak correlation was found with the SF-36 General Health Subscale by Bakas et al., (2006).

Responsiveness/ Precision/ Acceptability/ Feasibility

No data available.

Table 10.10: Developmental and evaluation studies relating to the Bakas Caregiving Outcomes Scale:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Bakas and Champion, 1999 **USA**	Caregivers of stroke survivors (sample 1= 92, sample 2= 104) Age sample 1 60.5 and sample 2 62.2 Self-completion questionnaire	√	√				
Bakas et al., 2004 **USA**	Caregivers of stroke survivors (114) Age 60.5 Self-completion questionnaire	√					
Bakas et al., 2006 **USA**	Family caregivers of stroke survivors (147) (USA) Age 51.7 years Interviewer administered (face to face or telephone) or self-completed	√ Internal consistency Test-retest	√ Construct				

c) Caregiver Burden Inventory (CBI)

Reliability

The total CBI score (Foster and Chaboyer 2003) and the five subscales of the CBI have been found to be internally consistent in studies by Novak et al., (2001); and Foster and Chaboyer (2003).

Validity

Internal validity

The five factor structure was supported empirically in a study by Novak and Guest (1989).

Dimension-specific variables

The total CBI score, as well as 4 of the 5 subscales (with the exception of emotional burden), have been found to be significantly correlated to filial (family) obligation (Foster and Chaboyer 2003).

Responsiveness

The Total CBI score was responsive to change in carers of patients with Parkinson's disease receiving cognitive behavioural therapy after 3 months of therapy, compared to controls in a study by Secker and Brown (2005).

Precision/ Acceptability/ Feasibility

No data available.

Table 10.11: Developmental and evaluation studies relating to the Caregiving Burden Inventory:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Novak and Guest, 1989 **Canada**	Carers of confused or disoriented older people (107) Age 60.1 Interviewer administered	√	√				
Foster and Chaboyer, 2003 **Australia**	Carers of a family member who had been critically ill & admitted to intensive care (71) Age 50.3 Self-completion	√	√				
Secker and Brown, 2005 **UK**	Carers of patients with Parkinson's (30) Age 59.1 for treatment group (n=15) and 58.8 for control group (n=15)			√			

d) Caregiving Appraisal Scale

Reliability

Internal consistency was good for the three factors of the ACS for 2 different samples of carers (Lawton et al., 1989). However, it has to be noted that the 3 factors were different for the 2 samples. Internal consistency was good for three of the four factors (Struchen et al., 2002). Caregiving mastery showed poor internal consistency, but since the factor emerged from the analysis of 2 different cohorts, Struchen et al. (2002) believed that it was a significant construct of caregiver appraisal. Using two of the original subscales (perceived burden and impact of caregiving role), Dracup et al. (2004) found good internal consistency of these subcscales.

Test-retest reliability was found to be reasonable in one of the samples investigated by Lawton et al (1989).

Validity

Construct Validity

Factor analysis has provided conflicting evidence of the structure of the CAS. Three factors found in a sample of carers of disabled elderly people in respite were not confirmed in a sample of carers of disabled elderly people in institutions (Lawton et al., 1989). Another study found a five-factor solution, but two factors (caregiving mastery and caregiving burden) were found to be less robust (Lawton et al., 1991). Another study concluded that factor analysis showed that the CAS has four sub-scales: perceived burden and caregiver satisfaction with their relationship to the patient, caregivers' ideology and caregiving mastery (Struchen et al., 2002). The four

factor solution was also found with 35 items (rather than the initial 47) and in different samples (carers of the traumatic brain injury model system and carers of the residential treatment programme cohort) (Struchen et al., 2002).

Generic health status

One of the factors (perceived burden) was significantly correlated to the GHQ, Subjective Burden Scale and Objective Burden Scale (Struchen et al., 2002). A negative significant correlation was found between caregiver relationship satisfaction and the Objective Burden Scale. However, this was a weak correlation.

Dimension-specific variables

The caregiving satisfaction subscale of the ACS was found to be significantly related to caregiving burden and caregiving burden was related to depression (Lawton et al., 1991).

Validity/ Responsiveness/ Precision/ / Feasibility

No data available.

Acceptability

Of 241 participants, 11 cases had more than one response missing (Struchen et al., 2002).

Table 10.12: Developmental and evaluation studies relating to the Caregiving Appraisal Scale:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Lawton et al., 1989 USA	Carers of disabled older people in respite care or in institutions (632) Age 59.7 respite care, 76.2 institutionalized care Interviewer administered	√	√				
Lawton et al., 1991 USA	Spouse (285) and adult child (244) carers of elderly people with Alzheimer's Disease Age 76.3 Interviewer administered		√				
Struchen et al., 2002 USA	Carers of person with traumatic brain injury (241) Age47.0 Self-completion	√	√			√	
Dracup et al., 2004 USA	Spouses of patients with heart failure (75) Age 54.0 Self-completion	√					

e) Caregiving Impact Scale

Reliability

The CIS has been found to be internally consistent in two studies (Cameron et al., 2002; Cameron et al., 2006a).

Validity/ Responsiveness/ Precision/ Acceptability/ Feasibility

No data available.

Table 10.13: Developmental and evaluation studies relating to the Caregiving Impact Scale:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Caregiving Impact Scale		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Cameron et al., 2002 **Canada**	Family caregivers of cancer patients (44) Age 55.8 Interviewer administered	√					
Cameron et al., 2006a **Canada**	Informal carers of stroke survivors (94) Age 60.8 Interviewer administered or self-completion	√					

f) Caregiver Strain Index

Reliability

Internal consistency

Four studies found the CSI Total score to be internally consistent (Robinson 1983; Berg-Weger et al., 2000; Jenkinson et al., 2000; Diwan et al., 2004). However, one study found internal consistency for 2 of the 3 subscales below 0.7 (Diwan et al., 2004). The modified CSI has also been found to be internally consistent (Thornton and Travis 2003).

Test-retest reliability

Test-retest reliability had not been investigated in the original CSI studies. However, in the modified CSI study, it was found that test-retest reliability was better for the modified CSI than the parent CSI (Thornton and Travis 2003).

Validity

Construct validity

Principle component analysis identified 3 factors of the CSI, which were comparable but not identical to findings reported by others (Diwan et al., 2004). Exploratory factor analysis and structural equation modelling found three factors that the authors reported to be similar, but not identical, to those proposed originally by Robinson in 1984 (Rubio et al., 1999). However, it is unclear what the authors are referring to, as Robinson (1984) does not report any factors or sub-scales of the CSI in the original development.

Socio-demographic variables

No significant difference was found in the level of strain as measured by the CSI at three or six months after a stroke between men and women in a study by Blake et al., (2003). Adult children were significantly more likely to report role strain compared to spouses and other carers. Higher income of the caregiver was predictive of greater role strain, and perceived lack of support from health care services was associated with greater personal strain (Diwan et al., 2004). For the modified CSI, age was found to be inversely related to carer strain (Thornton and Travis 2003).

Generic health status

The CSI has been found to be significantly moderately to strongly correlated with the General Health Questionnaire-12 (Blake and Lincoln 2000; Blake et al., 2003), patient Extended Activities of Daily Living Scale (EADL), and Negative Affectivity (Blake et al., 2003). In one study, the best predictor of the CSI was the carer's mood and other factors were the perceived patient EADL and negative affectivity (Blake and Lincoln 2000). In a second study, strain was accurately predicted by a model based on the General Health Questionnaire-12, Positive and Negative Affectivity Schedule (Blake et al., 2003). Also, CSI scores correlated moderately with PCS and weakly with MCS scores of the SF-36 in a study by Jenkinson et al., (2000).

Patient variables

Significant correlations were found between the CSI and a variety of patient variables (Robinson 1983). Positive correlations were found for CSI score and the patient's age, re-hospitalization within to months and mental status. Negative correlations were found with the patients' ability to perform activities of daily living and satisfaction with progress during convalescence. Another study however, found no significant correlation was found for CSI score with age of the patient or time since the stroke of the patient (Blake and Lincoln 2000). The modified CSI was found to be significantly correlated to the patient's mental capacity and physical functioning and the patient's age (Thornton and Travis 2003).

Caregiver variables

CSI scores were correlated with a number of variables of caregivers' perceptions (for example carer's perception that they were very involved with caregiving or emotional strain of the caregiver) (Robinson 1983).

Responsiveness

CSI scores have been shown to significantly reduce in a study examining effectiveness of cognitive behavioural therapy for the carer (Secker and Brown 2005).

Feasibility

Interviews took 15-40 minutes in a study by Diwan et al., (2004), but this included the completion of several other questionnaires. For the modified CSI, interview time was between 10 and 20 minutes (Thornton and Travis 2003).

Precision/ Acceptability

No data available.

Table 10.14: Developmental and evaluation studies relating to the Carer Strain Index:

Study/ Country	**Population (N) Age Method of administration Setting**	**Measurement properties**					
Carer Strain Index		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Robinson, 1983 **USA**	Carers of recently hospitalized hip surgery or heart patients (81) Age 22-83 Interviewer administered	√	√				
Rubio et al., 1999 **USA**	Caregivers of adults with dementia (27) and children with developmental problems (8) or 'healthy' children younger than 12 years of age (53); and non-caregivers (77) Self-report questionnaire		√				
Blake and Lincoln, 2000 UK	Co-resident spouses of stroke patients (222) Age 69.0 Self completion questionnaire		√				
Berg-Weger et al., 2000 USA	Caregivers of chronically ill family members (142) Self-report Recruited through six caregiver-support organizations	√					
Jenkinson et al., 2000 **Europe incl. UK**	Carers of patients with amyotrophic lateral sclerosis (415) Age 55.1 years Set in 74 clinical Self-completion questionnaire	√	√				
Blake et al., 2003 UK	Spouses of stroke patients (130 at 3 months and 116 at 6 months) Age 66.4		√				
Thornton and Travis (2003) USA Modified CSI	Caregivers of family members or friends aged 53+ taking medication regularly and receiving formal or informal assistance (158) Interviewer administered	√	√				√
Diwan et al., 2004 USA	Caregivers of community-dwelling dementia patients (150) Age 61.9 Interviewer-administered	√	√				√
Secker and Brown, 2005 UK	Carers of patients with Parkinson's (30) Age 59.1 for treatment group (n=15) and 58.8 for control group (n=15)			√			

g) Caregiver Well-Being Scale

Reliability

Internal consistency was high for the scale overall, for the two subscales, and for the four factors of the first subscale (Basic Human Need) and for three out of the five factors of the second subscale (Activities of Daily Living) (2 factors had low internal consistency) (Tebb 1995). The Caregiver Well-Being Scale had good reliability (Berg-Weger et al., 2000).

Validity

Face validity

Face validity was examined by four people familiar with the caregiving literature (Tebb 1995)

Construct validity

For construct validity, moderate to high correlations were found to the Computerized Stress Inventory (Tebb 1995). Lifestyle satisfaction scores were moderately to highly correlated to the two sub-scales of the Caregiver Well-Being Scale (Tebb 1995).

Factor analysis suggested that some items could be deleted from the questionnaire, as these items did not load highly on any factor (Berg-Weger et al., 2000). Furthermore, structural equation modelling showed that the original models did not fit the data, and consequently the models were revised to fit the data (Rubio et al., 1999). For example, only three of the original four constructs measured the sub-scale of 'basic needs'.

Criterion validity

Criterion validity was assessed by comparing the scores of caregivers to non-caregivers. Not all the expected significant differences were found, but the differences were in the expected direction (Tebb 1995).

Responsiveness/Precision/Acceptability/Feasibility

No data available.

Table 10.15: Developmental and evaluation studies relating to the Caregiver Well-Being Scale:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Caregiver Well-Being Scale		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Tebb, 1995 **USA**	Caregivers of adults with dementia (27) and children with developmental problems (8) or 'healthy' children younger than 12 years of age (53); and non-caregivers (77) Self-report questionnaire	√ Internal consistency	√ Face Criterion and construct				
Rubio et al., 1999 **USA**	Re-analysis of the sample from the study by Tebb (1995)		√				
Berg-Weger et al.,2000 **USA**	Caregivers of chronically ill family members (142) Self-report Recruited through six caregiver-support organizations	√	√				

h) Zarit Burden Interview

Reliability

Good internal consistency was found for the 29-item ZBI (Williams 1993), the 22-item ZBI (Zarit et al., 1987; Whitlatch et al., 1991; Majerovitz 1995; Hebert et al., 2000; Bedard et al., 2001; McConaghy and Caltabiano 2005), for a short (12-item) ZBI (Hebert et al., 2000; Bedard et al., 2001; O'Rourke and Tuokko 2003) and for a 4-item screening ZBI (Bedard et al., 2001). Good internal consistency has also been found for the 2 sub-scales (Personal Strain and Role Strain), as well as individual items of the ZBI (Whitlatch et al., 1991).

Validity

Internal validity

Factor analysis supported a two factor solution for the short ZBI (Hebert et al., 2000; Bedard et al., 2001; O'Rourke and Tuokko 2003), but the short ZBI showed a better adjustment than the original ZBI (Hebert et al., 2000). Strong significant correlations were found between the original ZBI and the short ZBI (Bedard et al., 2001), making the short version of the ZBI comparable to the full version.

Socio-demographic variables

For the 29-item ZBI, no difference was found in the total score between daughters and spouses as caregivers (Zarit et al., 1980). However, in a later study, daughters and wives were found to score significantly higher on the ZBI than other caregivers (Zarit et al., 1987). Significant differences in ZBI score (20-item) were found between husband and wife carers (Zarit et al., 1986). No significant effect was found for education and income (Zarit et al., 1987). Scores for the short ZBI (12-item) were significantly higher for women than for men (Bedard et al., 2001). However, another study did not find any correlation between ZBI score and the gender of the caregiver (Hebert et al., 2000). Furthermore, the same study did not find any correlation of the ZBI score and marital status and employment status. It has also been found that more women than men have a higher ZBI score (Gallicchio et al., 2002). Younger carer age has also been found to be significantly associated with a higher ZBI score (Schneider et al., 1999)

Patient variables

Contrary to expectations, none of the variable on the patient's impaired behaviours were correlated with the carer's burden (Zarit et al., 1980). Significant associations were found between ZBI scores and behavioural disturbance (behavioural deficits) and cognitive impairment of the patient (Schneider et al., 1999). No correlation was found between carer burden and duration of illness (Zarit et al., 1980).

General health status

A lack of association between the GHQ-12 and carer burden, measured by the ZBI, was found in a European study (Schneider et al., 1999).

Dimension-specific outcomes

In the original development of the ZBI, only social support, in terms of frequency of family visits, was significantly (negatively) correlated with carer burden (Zarit et al., 1980). For the 29-item ZBI, the sense of burden was moderately to strongly associated with psychological well-being, but not significantly correlated with most variables of physical well-being (Williams 1993). One study found that high levels of burden were negatively correlated to psychological health (measured by SF-36v2) (McConaghy and Caltabiano 2005). Caregiver burden, as measured by the 22-item ZBI was not correlated with caregiver adaptability, but was correlated moderately to caregiving stress variables and memory and behaviour problems (Majerovitz 1995). Correlations for the 22-item ZBI with other measures were statistically significant, but weak, apart for depression for which a moderate correlation was found and correlations for the short ZBI were also weak to moderate (Hebert et al., 2000). Evidence of predictive validity of the short ZBI for depressive symptoms has also been found (O'Rourke and Tuokko 2003).

Responsiveness

A significant decrease over time in carer burden was found in wives, but not husbands, especially for wives who had placed their spouses into a nursing home (Zarit et al., 1986). ZBI scores have been shown to decrease over time, however the

decrease was greater in the waiting list group than in the intervention groups (support or counseling groups) (Zarit et al., 1987).

Interpretability

According to (Bedard et al., 2001), a score of 17 or above on the short ZBI (representing the top quartiles) may be used as cut off point to identify high burden. However, O'Rourke and Tuokko (2003) found this not optimal, upon comparison of the short ZBI scores to scores of the Centre for Epidemiologic Studies-Depression Scale. Their suggested cut off of 10 was not optimal either, and thus it is too early to propose a definite cut off point.

Acceptability

Only one study reported on missing variables, with data missing only on 10 individual items (3.2%) (Hebert et al., 2000).

Feasibility

Both the long and the short ZBI have a low number of items. The ZBI has mostly been interviewer administered.

Table 10.16: Developmental and evaluation studies relating to the Zarit Burden Interview:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Zarit Burden Interview		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Zarit et al., 1980 **USA** **29 items**	Caregivers or people with senile dementia (29) Age 65.0 Interviewer administered		√				
Zarit et al., 1986 **USA** **20-item**	Husbands (31) and wives (33) who were caring for their spouse with Alzheimer's Disease Age husbands 72.3, wives 63.4 Interviewer administered		√				
Zarit et al., 1987 **USA** **22 items**	Care givers of patients with dementia living in the community (119) Age 62.0 Interviewer administered	√		√			
Whitlatch et al., 1991 **USA** **2-item**	Carers of non-institutionalized dementia patients (113) Age 62.0	√					
Williams, 1993 **USA** **29-item**	Caregivers of stroke patients (29) USA Age 56.4	√	√				
Majerovitz 1995 **USA** **22-item**	Spouses of patients with dementia (54) Age 70.5 Interviewer administered	√	√				

Table 10.16 (contd.): Developmental and evaluation studies relating to the Zarit Burden Interview:

Study/ Country	Population (N) Age Method of administration Setting	Measurement properties					
Zarit Burden Interview		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Schneider et al., 1999 **Europe including the UK** **29-item**	Co-resident spouses of people with probable dementia (20) Age 71.0 Interviewer administered		√				
Hebert et al., 2000 **Canada** **22-item**	Caregivers of people with dementia in the community (327) 61.7 years Face to face interview in the caregiver's home	√	√				
Bedard et al., 2001 **22-, 12- and 4-item versions** **Canada**	Caregivers of cognitively impaired adults referred to a memory clinic (413) Age 61.0 Interviewer administered	√	√				
Gallicchio et al., 2002 **USA** **22-item**	Carers of community-dwelling dementia patients (327) Age 61.6 Interviewer administered		√				
O'Rourke and Tuokko, 2003 **Canada** **12-item version**	Carers of institutionalized and community-dwelling patients with dementia (770) Age 58.6 Interviewer administered	√	√	√			
McConaghy and Caltabiano, 2005 **Australia** **22-item**	Carers of people with dementia (42) Age 62.0 years Self-completion questionnaire	√	√				

Other carer-specific instruments identified from the review

The following table provides an overview of other records of carer-specific instruments identified. They have in common the fact that only one record of a study was found evaluating the instrument; insufficient evidence to justify assessing the instrument in more detail. .

Nineteen single study evaluations of instruments are included. Most of the instruments were evaluated in the USA, only 2 in the UK. The majority of evaluations only gave information on internal consistency and validity. The Caregiver Quality of Life Instrument (CQLI) (Mohide et al., 1988) was tested more extensively, but only in a small sample.

Table 10.17 General carer instruments evaluated in a single study

Instrument/ reference	**Population (N) Age Method of administration Setting**	**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**	**Comments** **No other records identified unless stated**
Brief Assessment Scale for Caregivers (BASC) and Negative Personal Impact (NPI) subscale; Glajchen et al., 2005	Caregivers of patients with chronic illness (102) (USA) Age 49.0 Interviewer administered (face to face or telephone)	√ Internal consistency	√ Construct					Internal consistency was acceptable for the BASC and the NPI. Construct validity was confirmed by significant correlations to other measures of caregiver burden. (such as Objective Caregiver Burden).
Burden Scale; Pruchno, 1990	Carers of spouses with Alzheimer's Disease (315) USA Age 70.2 Interviewer-administered	√	√					The burden scale was found to be internally consistent. The Burden Scale was correlated to the CES-D scale.
Caregiver Activity Survey; Davis et al., 1997	Caregivers of Alzheimer patients (42) (USA) Self-completion questionnaire	√ Test-retest	√					The Caregiver Activity Survey total score had high test-retest reliability. Convergent validity was supported by comparing the Caregiver Activity Survey to other Alzheimer's disease measures and an independent measure of caregiver burden.

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Caregiver Appraisal Scale (CAS) (primary instrument) Subjective Burden Scale, Objective Burden Scale, General Health Questionnaire; Struchen et al., 2002	Caregivers of persons with traumatic brain injury (241) Age 47.0 USA Self-completion questionnaire	√	√					Factor analysis yielded 4 factors for the CAS. Three of the 4 factors showed good internal consistency. The perceived burden sub-scale had moderate correlations with the 3 other instruments, but correlations were weak for the other sub-scales (caregiver relationship satisfaction, caregiving ideology and caregiving mastery).
Caregiver Change Interview; Zarit et al., 1987	Caregivers of dementia patients living in the community (119) Age 62.0 USA Interviewer administered	√						Good internal consistency was found for the 4 sub-scales
Craregiving Burden Scale; Knight et al., 1998	Caregivers of persons with traumatic brain injury (52) Age 47.1 Self-completion	√	√					Five of the seven CBS subscales showed good internal consistency (family impact and physical burden did not have good internal consistency) For validity, Parents scored significantly higher on the pessimism and physical burden subscales of the CBS than spouses when caring for a person with traumatic brain injury. The CBS total score was significantly correlated to symptom distress, coping, social support and depression

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Caregiver Distress Scale (CDS); Cousins et al., 2002	Parkinson's caregivers (80) UK Age 69.3 years Administered by clinician or self-report	√					√	Subscales of CDS were internally consistent. Hierarchical cluster analysis and factor analysis led to a 17 item questionnaire with 5 subscales. Can be answered quickly as only 17 items and is quick to score by adding up responses to the different items.
Caregiver Experience Scale; Lemoine et al., 2005	Caregivers of people with mental health problems (405) Canada Age 41.3 Self-completion questionnaire	√	√					Construct validity was assessed by factor analysis, which led to a reduction of the items in the scale and a grouping of the IEQ into 8 subscales. Each subscale (apart from Stigma) showed good internal consistency.
Caregiver Perceived Burden; Macera et al., 1993	Caregivers of family members with dementia (82) Age 61.0 USA Interviewer administered	√	√					Good internal consistency was found. The burden score was significantly correlated with the Center for Epidemioligc Studies Depression scale.
Caregiver Quality of Life Index (CQLI) Hospice Quality of Life Index (HQLI) to evaluate patients' QOL; McMillan and Mahon 1994	Carers of cancer patients on admission to hospice care (68) USA 57.7 years Self-completion questionnaire	√	√ Content Construct					Reliability for the CQLI was acceptable. Content validity was established through careful review of the literature and by experts evaluating the CQLI. Construct validity testing showed that the instrument and individual items can discriminate differentiate QOL on adults who are caregivers and adults who are not.

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Caregiver Quality of Life Instrument (CQLI); Mohide et al., 1988	Family caregivers of patients with chronic degenerative disorders (30) and carers of well elderly (10) Canada Interviewer administered	√ Test-retest	√ Construct	√		√	√	Good test-retest reliability was established. Construct validity was shown by the instrument discriminating among different degrees of caregiver wellbeing, discriminated among caregivers caring for elderly with different levels of health, and by the CQLI correlating with the general stress measure. For feasibility, one participant felt too tired to finish the questionnaire. The average time to complete the CQLI was 20 minutes (range 7-35 minutes). Responsiveness to with-subject change over time was shown when caregivers received respite.
Caregiver Stress Scale Feldman et al., 2003	Carers of patients with Alzheimer's disease (141 treatment group, 146 in placebo group) Age 65.5 treatment group, 66.8 in placebo group Canada, Australia, France			√				CSS scores at week 24 for the treatment group improved or remained the same from baseline, whereas the CSS scores declined for the placebo group. However it was only the difference for cognitive status that was statistically significant.

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Family Appraisal of Caregiving Questionnaire for Palliative Care (FACQ-PC) (primary instrument) Measures of family functioning, positive and negative affect and subjective burden; Cooper et al., 2006	Family caregivers of a relative with cancer (160) Australia Age 60.0 Self-completion questionnaire	√	√ Content Convergent and discriminant					High internal consistency was found for the 4 sub-scales of the FACQ-PC. Content validity was assessed by a panel of five experts in palliative care. Based on the assessment 26 of 28 items were retained for the questionnaire. Convergent and discriminant validity was assessed and demonstrated by comparing the FACQ-PC subscales to other measures.
Generic Caregiver Instrument; Schofield et al., 1997	Carers of people with a variety of long term conditions (976 at 1st interview, 802 at 2nd interview). Non-carers (200 at 1st interview, 181 at 2nd interview) Australia Telephone interview	√	√					Good internal consistency for the different sub-scales at the 2 times of data collection. The carers' reported levels of disability and dependency were independently validated in a sub-sample of carers through clinician assessment. Factor analysis was used to establish construct validity and to reveal sub-scales.

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Measure on positive aspects of caring (primary instrument) General Health Questionnaire, Memory and Behaviour Problem Checklist, Burden Interview, Past Social Interaction Scale and the Social Support Questionnaire; Cohen et al., 1994	Caregivers of people with dementia (196) Canada Standardized interview at carer's home	√ Test-retest	√					Test-retest reliability was assessed by correlations, which although were statistically significant were of moderate strength at best. Also the sample size reduced considerably at the 4 times of data collection. For validity of the measure, the number of positive aspects correlated with a number of different aspects from the other questionnaires used in the study. Again, although statistically significant, the correlations were at best moderate, and the majority were weak. The authors conclude that there may be need for further development of the measure.
Modified Caregiving Appraisal Scale Modified Katz Index of Independence in Activities of Daily Living; Sevick et al., 1994	Caregivers of home-based ventilator-dependent patients at home (29) 47.2 years USA Self-completion questionnaire	√	√ Face validity			√	√	Good internal consistency was found for the modified Katz Index and most of the sub-scales of the modified Caregiving Appraisal Scale (except ideology sub-scale). Final (modified) instrument was reviewed by home ventilation experts. Responses to the modified Katz Index had a moderate amount of missing data (20 pages of questions), with significant relationships found to 3 items.
Oberst Caregiving Burden Scale (OCBS) (primary instrument) Bakas Caregiving Outcomes Scale; Bakas et al., 2004	Caregivers of stroke survivors (114) Age 60.5 USA Self-completion questionnaire	√	√					High internal consistency was found for the OCBS Factor analysis showed that both OCBS subscales were uni-dimensional, thus providing evidence for construct validity. Female caregivers perceived the management of behavioural problems, provision of emotional support and carrying out household tasks as significantly more difficult than male carers .

Instrument/ reference	Population (N) Age Method of administration Setting	Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility	Comments No other records identified unless stated
Quality of Life Scale (Family Version) (QLS); Sherman et al., 2006	Caregivers for patients with AIDS (43) and cancer (38) USA Age AIDS caregivers Either interviewer administered or self-completion	√	√					QLS found to be reliable. Moderate to strong correlations were found between the different QLS subscales.
Scale for Caregiving Efficacy; Steffen et al., 2002	Two samples of family caregivers of a patient with Alzheimer Disease or another dementing disorder Study 1 n=169, 77.9 years Study 2 n=145, 77.3 years USA Both face to face administration and self-completion	√	√					The 2 subscales show strong internal consistency and adequate test-retest reliability. Support for construct validity was given in both studies, even though different measures were used to assess some of the constructs.

Examples of disease-specific instruments

Additionally, there are disease-specific carer instruments, i.e. instruments that capture the carer's experience in relation to a particular condition. Table 10.18 provides a list of instruments that have been tested for use in the population groups included in this review. This list may not be exhaustive.

The Experience of Caregiving Inventory (ECI), developed for mental health problems, may be of particular interest. The ECI has been developed in the UK and has been used in multiple mental health conditions, for example to compare the experience of caregiving for someone with anorexia or psychosis (Treasure et al., 2001). The ECI could easily be adapted to a general carer instrument.

Table 10.18: Disease-specific carer instruments

Disease	Instrument	Reference (Country)
Neuropsychiatry	Neuropsychiatry Inventory Caregiver Distress Scale (NPI-D)	Kaufer et al., 1998 USA
Cardiac Disease	Quality of Life Questionnaire for Cardiac Spouses (QL-SP)	Ebbesen et al., 1990 USA
Cancer	Caregiver Quality of Life Index- Cancer (CQOLC)	Weitzner et al., 1999; Weitzner and McMillan, 1999 USA
Multiple Sclerosis	Coping with Multiple Sclerosis Caregiving Inventory (CMSCI)	Pakenham 2002 Australia
Mental Health	Involvement Evaluation Questionnaire (IEQ)	Van Wijngaarden et al., 2003 5 European Countries, including the UK
Mental Health	Experience of Caregiving Inventory (ECI)	Szmukler et al., 1996; Joyce et al., 2000 UK

SUMMARY AND RECOMMENDATIONS - GENERIC INSTRUMENTS

Fifteen articles were included in this review of evaluation studies of generic health status instruments used to assess carer impact. Five generic instruments have been used in this way as what we consider indirect assessment of carer burden. The SF-36 and GHQ have had the most evaluations, with 4 studies identified for the SF-36 and 6 studies for the GHQ. The SF-12, a shorter version of the SF-36, has been evaluated once. Furthermore, the GHQ has been used in other studies as a reference measure for construct validity with general carer instruments. Other generic questionnaires that have been evaluated in single studies are the Health Utilities Index Mark 2 (HUI 2), Reintegration to Normal Living Index and the Ferrans and Power Quality of Life Index.

Generally, the sample sizes of the studies were relatively small (approximately 100 carers or less) and ranged from as little as 22 participants to as many as 679. The carers included cared for people with a range of illnesses, with Alzheimer's disease or dementia being the most commonly evaluated (n=6) and carers of stroke patients the second most commonly evaluated (n=3). Both the GHQ and SF-36 have been evaluated in different samples, in terms of the disease of the person that is being cared for. Mostly, the instruments were interviewer administered (n=7). Two studies to evaluate the SF-36 and all but one study to evaluate the GHQ were carried out in the UK.

For the SF-36, the psychometric properties evaluated were reliability (4 studies), validity (3 studies) and acceptability (1 study). For the GHQ, evaluations were restricted to validity (4 studies) and responsiveness (2 studies). The SF-36 is found to be internally consistent in the carer population, although the SF-12 was found to be only weakly internally consistent. No information of internal consistency has been reported for the GHQ.

Validity for the SF-36 was supported by comparing the scores of the SF-36 of carers to population norms, as well as by factor analysis which confirmed the original subscale structure of the SF-36 and by strong correlations with other relevant variables and dimensions such as depression or general carer instruments such as the Carer Strain Index. Validity of the GHQ was shown by comparing GHQ scores of people caring for patients with a different disease or by investigating the relationship of the GHQ with other variables. For the GHQ there is evidence from 2 studies that it is responsive to change, but no studies on the responsiveness of the SF-36 are reported. No or very limited data was available on precision, acceptability and feasibility for either of these instruments.

Disappointingly, there are few evaluations of the instruments reported in this review in the carer population. Thus, a lot of information on the psychometrics on the use of these instruments in the carer population is not available. The range of psychometric properties assessed is very limited, meaning that there is a lack of evidence of the performance of these generic instruments with the carer population.

Given the scarcity of evaluations for generic instruments in this population group, it is not possible to recommend a generic instrument to measure carer burden based on evaluative studies in carers. However, generic instruments such as the SF-36 and

GHQ have been widely evaluated in other population groups. It is therefore likely that these instruments are also useful in the carer population, but further studies need to be carried out to confirm this.

SUMMARY AND RECOMMENDATIONS- GENERAL CARER INSTRUMENTS

A total number of 57 studies were included for general carer instruments. Seven general carer instruments have evidence of measurement properties from multiple evaluations with carers: the Appraisal of Caregiving Scale (ACS) (2 evaluations), Bakas Caregiver Outcomes Scale (BCOS) (3), Caregiver Burden Inventory (CBI) (3), Caregiving Appraisal Scale (CAS) (4), Caregiving Impact Scale (CIS) (2), Caregiver Strain Index CSI (9), Caregiver Well-Being Scale (3) and Zarit Burden Interview (ZBI) (12). A further nineteen general carer instruments were identified which have been evaluated in only a single study each.

The instruments were developed with a range of carers, in terms of the type of patient they were caring for, such as stroke, cancer, Parkinson's disease. The majority of studies have been conducted with carers of patients suffering from Alzheimer's disease or dementia (n=21). The evaluative studies of some instruments remained limited to carers of patients with one type of condition, e.g. the ACS was evaluated in carers of cancer patients or the BCOS in carers of patients with stroke, whereas other instruments, such as the CSI, were evaluated in a range of conditions.

The method of administration was by interview in twenty-one studies, by self-completion in twenty-five studies and by both interview or self-administration in five studies. Five studies did not report the method of administration. Also, the majority of interview studies did not report if the instrument was administered by telephone or face to face (n=17). A large range of sample sizes has been used to evaluate instruments, with samples as small as 20 and as large as 770. The majority of evaluations were conducted in samples with fewer than 100 participants (n=25). Only 11 studies used samples larger than 200 people.

The majority of studies (both for multiple and single evaluations) report only on reliability (internal consistency) and validity. Responsiveness, precision, acceptability and feasibility are neglected issues when evaluating general carer instruments. In terms of the quality of reporting the development and evaluation of general carer instruments, there is often missing information regarding method of administration, population demographic information (e.g. age) or details of scoring of the instrument. Furthermore, the majority of evaluations have been carried out in the USA, with only 7 evaluative studies having been conducted in the UK.

However, given these limitations, the number of instruments identified demonstrates that there is increasing interest in studying the burden of persons caring for someone with ill health. The development of these instruments has helped to identify the important domains in the study of carer burden with the most widely used domains being psychological wellbeing, social well-being and role activities. All of the instruments that having undergone multiple evaluations comprise domains or items reflecting Social well-being, all but one instrument investigate Psychological well-being and 5 of 8 investigate Role activities.

Currently the ZBI and the CSI have been evaluated more extensively. Differences and similarities in the psychometric properties of the two instruments have been found. Both the ZBI (long and short versions) and the CSI have been found to be internally consistent. Test-retest reliability information is only available for the modified CSI.

In terms of validity, the evidence for the ZBI is contradictory. Different studies report conflicting findings on validity, for example one study found that ZBI scores are no different between different types of carers, whereas another study found a difference in ZBI scores between different types of carers. The conflicting evidence about the ZBI validity may be related to the multitude of different versions of the ZBI having been tested. More consistent evidence for the validity of the ZBI has been found by its moderate to strong correlation with dimension-specific variables, such as psychological well-being. There is also controversy for the construct validity of the CSI and it is not clear whether the CSI is uni- or multi-dimensional.

Both the ZBI and CSI have been shown to be responsive to change, but this is based on the information of only one study for each instrument. Disappointingly, the information on precision, feasibility and acceptability is limited.

However, both instruments do have attractive features. They are short, the ZBI is the longest with 29 items, although shorter versions (22, 20, 12 and 4 item versions) exist. The CSI has 13 items. It has been suggested that the short 4-item version of the ZBI may be useful for clinical practice, however no definite cut off points of when a carer experiences a heavy care burden have been established, which limits the usefulness of the instrument. Both have been evaluated in UK carer populations but the ZBI evaluation was part of a multi-national European study. One advantage of the CSI is that it is a self-completion questionnaire, whereas the the ZBI is an interviewer administered instrument (although it was used in one study as a self-completion questionnaire). Although there no evidence was found that the ZBI is disease-specific, the ZBI has been evaluated solely in carers of dementia patients.

A range of carer instruments that are disease-specific have also been developed, examples of which are given in table 10.18. Discussing the measurement properties of these instruments was beyond the scope of this review. However, these instruments may represent an appropriate and valid method to investigate carer burden by making the instrument more specific to the type of care that a carer provides given a specific illness.

Overall, due to the limited information of the psychometric properties of general carer instruments, the majority of instruments cannot be recommended for widespread use., Furthermore, given the scarcity of psychometric information of general carer instruments that have been evaluated in multiple studies, it is not possible to recommend any particular instrument at this point in time. Currently, the ZBI and CSI appear to be the most promising general carer instruments, but further evaluations are necessary before definite recommendation can be made. A small but important advantage of CSI over ZBI is the more substantial evidence of use in the more feasible format of self complete questionnaire.

Because of extensive evidence of their use in a wide range of contexts, two broad measures of health status, SF-36 and GHQ, can be used to provide indirect evidence of carer impact. The CSI and ZBI provide more direct evidence of carer impact, with

the CSI being somewhat more supported for use in the format of self completion questionnaire. Despite it not being possible currently to make definite recommendations for either a generic or general carer instrument to be used to investigate the burden of carers for a person with ill health, the combined use of generic and general carer instruments can be recommended as a strategy. The generic instrument would capture broader health impact and allow comparison of the quality of life of carers with the general population or even with persons with ill health. The use of a general carer instrument would allow capturing information that is more specific to the caregiving experience.

REFERENCES

Bakas T, Austin JK, Jessup SL, Williams LS and Oberst MT. Time and difficulty of tasks provided by family caregivers of stroke survivors. *J Neurosci Nurs* 2004; **36**(2): 95-106.

Bakas T and Champion V. Development and psychometric testing of the Bakas Caregiving Outcomes Scale. *Nursing Research* 1999; **48**(5): 250-259.

Bakas T, Champion V, Perkins SM, Farran CJ and Williams LS. Psychometric Testing of the Revised 15-item Bakas Caregiving Outcomes Scale. *Nurs Res* 2006; **55**(5): 346-355.

Bedard M, Molloy DW, Squire L, Dubois S, Lever JA and O'Donnell M. The Zarit Burden Interview: a new short version and screening version. *Gerontologist* 2001; **41**(5): 652-7.

Bell CM, Araki SS and Neumann PJ. The association between caregiver burden and caregiver health-related quality of life in Alzheimer disease. *Alzheimer Disease and Associated Disorders* 2001; **15**(3): 129-136.

Berg-Weger M, Rauch SM, Rubio DM and Tebb SS. Assessing the health of adult daughter former caregivers for elders with Alzheimer's disease. *American Journal of Alzheimer's Disease and Other Dementias* 2003; **18**(4): 231-239.

Berg-Weger M, Rubio DM and Tebb SS. The Caregiver Well-Being Scale revisited. *Health and Social Work* 2000; **25**(4): 255-263.

Berg-Weger M and Tebb SS. Caregiver well-being: a strengths-based case management approach. *Journal of Case Management* 1998; **7**(2): 67-73.

Blake H and Lincoln NB. Factors associated with strain in co-resident spouses of patients following stroke. *Clin Rehabil* 2000; **14**(3): 307-14.

Blake H, Lincoln NB and Clarke DD. Caregiver strain in spouses of stroke patients. *Clin Rehabil* 2003; **17**(3): 312-7.

Bluvol A and Ford-Gilboe M. Hope, health work and quality of life in families of stroke survivors. *J Adv Nurs* 2004; **48**(4): 322-32.

Cameron JI, Cheung AM, Streiner DL, Coyte PC and Stewart DE. Stroke survivors' behavioral and psychologic symptoms are associated with informal caregivers' experiences of depression. *Arch Phys Med Rehabil* 2006a; **87**(2): 177-83.

Cameron JI, Franche RL, Cheung AM and Stewart DE. Lifestyle interference and emotional distress in family caregivers of advanced cancer patients. *Cancer* 2002; **94**(2): 521-7.

Cameron JI, Herridge MS, Tansey CM, McAndrews MP and Cheung AM. Well-being in informal caregivers of survivors of acute respiratory distress syndrome. *Crit Care Med* 2006b; **34**(1): 81-6.

Carey PJ, Oberst MT, McCubbin MA and Hughes SH. Appraisal and caregiving burden in family members caring for patients receiving chemotherapy. *Oncol Nurs Forum* 1991; **18**(8): 1341-8.

Clark PC, Dunbar SB, Shields CG, Viswanathan B, Aycock DM and Wolf SL. Influence of stroke survivor characteristics and family conflict surrounding recovery on caregivers' mental and physical health. *Nurs Res* 2004; **53**(6): 406-13.

Cohen CA, Gold DP, Shulman KI and Zucchero CA. Positive aspects in caregiving: an overlooked variable in research. *Canadian Journal on Aging/La Revue Canadienne du Vieillissement* 1994; **13**(3): 378-391.

Cooper B, Kinsella GJ and Picton C. Development and initial validation of a family appraisal of caregiving questionnaire for palliative care. *Psychooncology* 2006; **15**(7): 613-22.

Cousins R, Davies ADM, Turnbull CJ and Playfer JR. Assessing caregiving distress: a conceptual analysis and a brief scale. *British Journal of Clinical Psychology* 2002; **41**(Pt 4): 387-403.

Davis KL, Marin DB, Kane R, Patrick D, Peskind ER, Raskind MA and Puder KL. The Caregiver Activity Survey (CAS): development and validation of a new measure for caregivers of persons with Alzheimer's disease. *Int J Geriatr Psychiatry* 1997; **12**(10): 978-88.

Devins GM, Binik YM, Hutchinson TA, Hollomby DJ, Barre PE and Guttmann RD. The emotional impact of end-stage renal disease: importance of patients' perception of intrusiveness and control. *Int J Psychiatry Med* 1983; **13**(4): 327-43.

Diwan S, Hougham GW and Sachs GA. Strain experienced by caregivers of dementia patients receiving palliative care: findings from the Palliative Excellence in Alzheimer Care Efforts (PEACE) Program. *J Palliat Med* 2004; **7**(6): 797-807.

Dracup K, Evangelista LS, Doering L, Tullman D, Moser DK and Hamilton M. Emotional well-being in spouses of patients with advanced heart failure. *Heart Lung* 2004; **33**(6): 354-61.

Draper BM, Poulos CJ, Cole AM, Poulos RG and Ehrlich F. A comparison of caregivers for elderly stroke and dementia victims. *J Am Geriatr Soc* 1992; **40**(9): 896-901.

Ebbesen LS, Guyatt GH, McCartney N and Oldridge NB. Measuring quality of life in cardiac spouses. *Journal of Clinical Epidemiology* 1990; **43**(5): 481-487.

Feldman H, Gauthier S, Hecker J, Vellas B, Emir B, Mastey V and Subbiah P. Efficacy of donepezil on maintenance of activities of daily living in patients with moderate to severe Alzheimer's disease and the effect on caregiver burden. *J Am Geriatr Soc* 2003; **51**(6): 737-44.

Fitzpatrick R, Davey C, Buxton MJ, and Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment* 1998;**2**(14).

Foster M and Chaboyer W. Family carers of ICU survivors: a survey of the burden they experience. *Scand J Caring Sci* 2003; **17**(3): 205-14.

Gallicchio L, Siddiqi N, Langenberg P and Baumgarten M. Gender differences in burden and depression among informal caregivers of demented elders in the community. *Int J Geriatr Psychiatry* 2002; **17**(2): 154-63.

Glajchen M, Kornblith A, Homel P, Fraidin L, Mauskop A and Portenoy RK. Development of a brief assessment scale for caregivers of the medically ill. *J Pain Symptom Manage* 2005; **29**(3): 245-54.

Hebert R, Bravo G and Preville M. Reliability, validity and reference values of the Zarit Burden Interview for assessing informal caregivers of community-dwelling older persons with dementia. *Canadian Journal on Aging/La Revue Canadienne du Vieillissement* 2000; **19**(4): 494-507.

Jenkinson CP, Fitzpatrick R, Swash M and Peto V. The ALS Health Profile Study: quality of life of amyotrophic lateral sclerosis patients and carers in Europe. *Journal of Neurology* 2000; **247**(11): 835-840.

Joyce J, Leese M and Szmukler G. The Experience of Caregiving Inventory: further evidence. *Soc Psychiatry Psychiatr Epidemiol* 2000; **35**(4): 185-9.

Kaufer DI, Cummings JL, Christine D, Bray T, Castellon S, Masterman D, MacMillan A, Ketchel P and DeKosky ST. Assessing the impact of neuropsychiatric symptoms in Alzheimer's disease: the neuropsychiatric inventory caregiver distress scale. *Journal of the American Geriatrics Society* 1998; **46**(2): 210-215.

Knight RG, Devereux R and Godfrey HP. Caring for a family member with a traumatic brain injury. *Brain Inj* 1998; **12**(6): 467-81.

Lawton MP, Kleban MH, Moss M, Rovine M and Glicksman A. Measuring caregiving appraisal. *J Gerontol* 1989; **44**(3): P61-71.

Lawton MP, Moss M, Kleban MH, Glicksman A and Rovine M. A two-factor model of caregiving appraisal and psychological well-being. *J Gerontol* 1991; **46**(4): P181-9.

Lemoine O, Lavoie S, Poulin C, Poirier LR and Fournier L. Being the caregiver of a person with a mental health problem. *Can J Commun Ment Health* 2005; **24**(2): 127-43.

Macera CA, Eaker ED, Jannarone RJ, Davis DR and Stoskopf CH. A measure of perceived burden among caregivers. *Eval Health Prof* 1993; **16**(2): 205-11.

Majerovitz SD. Role of family adaptability in the psychological adjustment of spouse caregivers to patients with dementia. *Psychol Aging* 1995; **10**(3): 447-57.

McConaghy R and Caltabiano ML. Caring for a person with dementia: exploring relationships between perceived burden, depression, coping and well-being. *Nurs Health Sci* 2005; **7**(2): 81-91.

McMillan SC and Mahon M. The impact of hospice services on the quality of life of primary caregivers. *Oncology Nursing Forum* 1994; **21**(7): 1189-1195.

Mohide EA, Torrance GW, Streiner DL, Pringle DM and Gilbert R. Measuring the wellbeing of family caregivers using the time trade-off technique. *Journal of Clinical Epidemiology* 1988; **41**(5): 475-482.

Novak B, Kolcaba K, Steiner R and Dowd T. Measuring comfort in caregivers and patients during late end-of-life care. *American Journal of Hospice and Palliative Care* 2001; **18**(3): 170, 216-180, 216.

Novak M and Guest C. Application of a multidimensional caregiver burden inventory. *Gerontologist* 1989; **29**(6): 798-803.

O'Rourke N and Tuokko HA. Psychometric properties of an abridged version of The Zarit Burden Interview within a representative Canadian caregiver sample. *Gerontologist* 2003; **43**(1): 121-7.

Oberst MT, Thomas SE, Gass KA and Ward SE. Caregiving demands and appraisal of stress among family caregivers. *Cancer Nurs* 1989; **12**(4): 209-15.

Pakenham KI. Development of a measure of coping with multiple sclerosis caregiving. *Psychology and Health* 2002; **17**(1): 97-118.

Pruchno RA. The effects of help patterns on the mental health of spouse caregivers. *Res Aging* 1990; **12**(1): 57-71.

Robinson BC. Validation of a Caregiver Strain Index. *Journals of Gerontology* 1983; **38**(3): 344-348.

Rosenvinge H, Jones D, Judge E and Martin A. Demented and chronic depressed patients attending a day hospital: stress experienced by carers. *Int J Geriatr Psychiatry* 1998; **13**(1): 8-11.

Rubio DM, Berg-Weger M and Tebb SS. Assessing the validity and reliability of well-being and stress in family caregivers. *Social Work Research* 1999; **23**(1): 54-64.

Sander AM, High WM, Jr., Hannay HJ and Sherer M. Predictors of psychological health in caregivers of patients with closed head injury. *Brain Inj* 1997; **11**(4): 235-49.

Schneider J, Murray J, Banerjee S and Mann A. EUROCARE: a cross-national study of co-resident spouse carers for people with Alzheimer's disease: I--Factors associated with carer burden. *Int J Geriatr Psychiatry* 1999; **14**(8): 651-61.

Schofield HL, Murphy B, Herrman HE, Bloch S and Singh B. Family caregiving: measurement of emotional wellbeing and various aspects of the caregiving role. *Psychological Medicine* 1997; **27**(3): 647-657.

Secker DL and Brown RG. Cognitive behavioural therapy (CBT) for carers of patients with Parkinson's disease: a preliminary randomised controlled trial. *J Neurol Neurosurg Psychiatry* 2005; **76**(4): 491-7.

Sevick MA, Sereika S, Matthews JT, Zucconi S, Wielobob C, Puczynski S, Ahmad SM and Barsh LF. Home-based ventilator-dependent patients: measurement of the emotional aspects of home caregiving. *Heart and Lung* 1994; **23**(4): 269-278.

Sherman DW, Ye XY, McSherry C, Parkas V, Calabrese M and Gatto M. Quality of life of patients with advanced cancer and acquired immune deficiency syndrome and their family caregivers. *J Palliat Med* 2006; **9**(4): 948-63.

Smith LN, Norrie J, Kerr SM, Lawrence IM, Langhorne P and Lees KR. Impact and influences on caregiver outcomes at one year post-stroke. *Cerebrovascular Diseases* 2004; **18**(2): 145-153.

Steffen AM, McKibbin C, Zeiss AM, Gallagher-Thompson D and Bandura A. The revised scale for caregiving self-efficacy: reliability and validity studies. *Journals of Gerontology - Series B: Psychological Sciences and Social Sciences* 2002; **57**(1): 74-86.

Struchen MA, Atchison TB, Roebuck TM, Caroselli JS and Sander AM. A multidimensional measure of caregiving appraisal: validation of the Caregiver Appraisal Scale in traumatic brain injury. *Journal of Head Trauma Rehabilitation* 2002; **17**(2): 132-154.

Szmukler GI, Burgess P, Herrman H, Benson A, Colusa S and Bloch S. Caring for relatives with serious mental illness: the development of the Experience of Caregiving Inventory. *Soc Psychiatry Psychiatr Epidemiol* 1996; **31**(3-4): 137-48.

Tebb SS. An aid to empowerment: a caregiver well-being scale. *Health and Social Work* 1995; **20**(2): 87-92.

Thornton M and Travis SS. Analysis of the reliability of the modified caregiver strain index. *J Gerontol B Psychol Sci Soc Sci* 2003; **58**(2): S127-32.

Treasure J, Murphy T, Szmukler G, Todd G, Gavan K and Joyce J. The experience of caregiving for severe mental illness: a comparison between anorexia nervosa and psychosis. *Soc Psychiatry Psychiatr Epidemiol* 2001; **36**(7): 343-7.

Van Wijngaarden B, Schene A, Koeter M, Becker T, Knapp M, Knudsen HC, Tansella M, Thornicroft G, Vazquez-Barquero JL, Lasalvia A and Leese M. People with schizophrenia in five countries: conceptual similarities and intercultural differences in family caregiving. *Schizophrenia Bulletin* 2003; **29**(3): 573-586.

Weitzner MA, Jacobsen PB, Wagner H, Friedland J and Cox C. The Caregiver Quality of Life Index-Cancer/CQOLC scale: development and validation of an

instrument to measure quality of life of the family caregiver of patients with cancer. *Quality of Life Research* 1999; **8**(1-2): 55-63.

Weitzner MA and McMillan SC. The Caregiver Quality of Life Index-Cancer/CQOLC scale: revalidation in a home hospice setting. *Journal of Palliative Care* 1999; **15**(2): 13-20.

Weitzner MA, Meyers CA, Steinbruecker S, Saleeba AK and Sandifer SD. Developing a caregiver quality of life instrument: preliminary steps. *Cancer Practice* 1997; **5**(1): 25-31.

Whitlatch CJ, Zarit SH and von Eye A. Efficacy of interventions with caregivers: a reanalysis. *Gerontologist* 1991; **31**(1): 9-14.

Williams AM. Caregivers of persons with stroke: their physical and emotional wellbeing. *Quality of Life Research* 1993; **2**(3): 213-220.

Woods RT, Wills W, Higginson IJ, Hobbins J and Whitby M. Support in the community for people with dementia and their carers: a comparative outcome study of specialist mental health service interventions. *Int J Geriatr Psychiatry* 2003; **18**(4): 298-307.

Zarit SH, Anthony CR and Boutselis M. Interventions with care givers of dementia patients: comparison of two approaches. *Psychol Aging* 1987; **2**(3): 225-32.

Zarit SH, Reever KE and Bach-Peterson J. Relatives of the impaired elderly: correlates of feelings of burden. *Gerontologist* 1980; **20**(6): 649-55.

Zarit SH, Todd PA and Zarit JM. Subjective burden of husbands and wives as caregivers: a longitudinal study. *Gerontologist* 1986; **26**(3): 260-6.

Chapter 11: Measuring patient perceptions of quality in health care: a structured review to inform service delivery for chronic disease

Summary

This review provides a structured synthesis of published evidence for the measurement and practical properties of patient-reported measures that communicate patients' experience of health care quality of relevance to long-term physical conditions or chronic disease management. The review aims to inform the future selection of multi-dimensional measures of patient-perceived health care quality.

BACKGROUND

a) Chronic disease

Chronic disease, defined as long-term conditions that can be controlled but not cured (DoH, 2004), represents the major cause of health problems in the United Kingdom (UK) (DoH, 2004). The growing demands to provide care appropriate to the needs of people with chronic disease are significant, representing a shift from the demands for acute health care: in the UK, 60% of all adults are diagnosed with one or more chronic condition, 60% of all hospitalisations are due to chronic disease or exacerbations, and 80% of GP consultations are related to chronic ill-health (DoH, 2004). It is evident that a high proportion of the UK's health service is currently devoted to health care provision for people with chronic disease; ensuring that the provision of health care is appropriate and of the highest quality is a major challenge, and an essential component of quality improvement efforts.

Many health care systems were designed for acute, episodic health care, and are, at best inappropriate for the management of chronic disease (Bodenheimer et al., 2002). Health care pathways for people with long-term chronic disease are often complex, and numerous shortcomings in the provision of quality health care have been described, including access to care (Davis and Wagner, 2000; Haggulund et al., 2005), continuity of care (Davis and Wagner, 2000; Thapar and Roland, 2005), integration of care between service providers (DoH, 2004), and respect for patients values, preferences and expressed needs (Hibbard, 2003; Davis and Wagner, 2000; Groves and Wagner, 2005).

Various strategies for improving care provision for people with chronic disease have been suggested, including patient involvement in decision making, care planning, and the regular monitoring of care provided (Groves and Wagner, 2005). The Chronic Care, or Chronic Disease, Model (CCM/CDM) was proposed to inform the provision of health care that embraced the needs of patients with long-term conditions (Bodenheimer et al., 2002). The model emphasizes a patient-centred approach to long-term health care, within which well-informed and self-motivated patients are supported proactive members of a multidisciplinary team. Moreover, a patient-centred approach should be responsive to the values, needs and preferences of patients (Hibbard, 2003); respect for and incorporation of patient values may be used to inform both the provision of care and evaluation of patient experience, providing a basis for improving service delivery. The multi-disciplinary nature of health care

suggests that care may often be provided across a range of different settings, by different members of the health care team, and often by more than one member from different disciplines over prolonged periods of time (Campbell et al., 2000). Hence, issues such as access to care, continuity of care, and co-ordination of care between team members become important requirements to ensuring care of the highest quality.

b) Quality health care

Modernisation of health care systems and associated advances in evidence-based healthcare has raised expectations of improvements in the quality of care (Powell et al., 2003; Sheldon, 2005). Moreover, the growing demand for health care, combined with rising costs and limited resources, has increased the emphasis on the efficient use of health care resources (Campbell et al., 2000). It is predicted that chronic disease will be the leading cause of disability by 2020; unless accompanied by good management, it will also become the most expensive health care problem (DoH, 2004). The drive for accountability and associated growth in quality improvement initiatives and performance measurement has ensued. Chronic disease management is now an essential component of quality improvement efforts within health care (Davis and Wagner, 2000). The provision of clinically effective, evidence-based health care, which is both acceptable and beneficial to patients, are important elements in understanding quality in health care (Jenkinson et al., 2002; Fitzpatrick, 1997).

Measuring and communicating health care quality requires rigorous and appropriate measurement of key and consensual variables that reflect the breadth and complexity of health care. Health care quality is, by necessity, a multi-factorial and broad ranging concept; the identification and appropriate measurement of key dimensions of health care service delivery of relevance to chronic disease management is a prerequisite to improving quality in health care (Hibbard et al., 2005). However, discrepant views between different stakeholders within the health care system, ranging from patients to providers, exist with regards to the definition and prioritization of quality issues (Campbell et al., 2002; Leatherman and Sutherland, 2003).

Numerous attempts to describe dimensions of relevance to a patient's experience of health care have been described. A patient's perspective of quality may include their desired health outcome (Mitchell and Lang, 2004; Swan and Boruch, 2004), their relationship with healthcare providers, the qualifications and performance of healthcare providers, and access to and choice of healthcare (Campbell et al., 2002; Hibbard, 2003). Exploring the concept of chronic disease management and patient-centred health care, the Institute of Medicine (IoM) (Committee on Quality of Health Care in America, 2001) engaged with health professionals and patients to describe nine core dimensions reflective of patient-centred quality health care: these include respect for patient values; attention to patient preferences and expressed needs; co-ordination and integration of care; information, communication and education; physical comfort; emotional support; involvement of family and friends; transition and continuity; and access to care. Intermediate outcomes, considered important intermediary steps in the achievement of improved health status, and reflective of key elements within the chronic care model, such as patient knowledge, self-efficacy and self-management skills (Hibbard et al., 2004), have also been described.

Other authors have described similar dimensions to those proposed by the IoM, reflective of central aspects of patient care (Gerteis et al., 1993): access; physical comfort / pain management; hospital environment; patient involvement; information and communication; co-ordination of care; and discharge planning. With the context of primary care, similar patient generated dimensions have also been described as important to the provision of good quality care (Coulter, 2005): fast access; trust in professional providing care; respect for patient preferences; patient involvement; information, education and support for self-care; attention to physical and environmental needs; emotional support; involvement of family and carers; continuity of care and smooth transition and coordination of care.

c) Patient reported quality in health care

Traditionally, health care quality has been assessed in terms of measures of structure, process and outcome (Donabedian, 1966; Campbell et al., 2000; Parchman et al., 2002): structure considers the accessibility and relative quality of the many components of health care, for example, how accessible was care for an individual with chronic disease?; process considers the appropriateness of care, location and timing, for example, did an individual with chronic disease receive care that was appropriate to their needs, at the right time, and in a suitable location? Measures of outcome assess the outcomes of health care, and may include functional and clinical outcomes, or clinical targets.

However, assessment has often focused on the perspectives of the care-provider or health care organization, such as cost, length of stay and patient mortality; within a chronic disease context few assessments have included the patient's experience of care (Groves and Wagner, 2005). Failure to sufficiently involve the patient perspective may reduce the credibility and relevance of assessment, particularly if used to support patient involvement and inform patient choice. Moreover, rigorous evidence of measurement reliability and validity is often lacking for more traditional measures or 'indicators' of care quality (Kendrick, 2001), and limited empirical evidence supports their contribution towards actually improving health care quality (Appleby and Devlin, 2004; Mitchell and Lang, 2004).

Patient-reported measures of health care quality aim to include the patient's perspective across a range of quality concerns in the assessment process. Well developed measures, particularly those that have involved patients in development and embrace the complex and multi-dimensional nature of health care, provide an important resource for assessing and communicating the quality of health care (Campbell et al., 2000). Although there may be occasions where a focus on specific elements of health care quality is important (Bredart et al., 2005), for example, a focus on the continuity of care, approaches that embrace individual dimensions may provide only a partial illustration of health care. Measurement that embraces the multi-dimensional nature of health care quality may be more meaningful to informing quality improvement initiatives.

d) Assessment of quality health care for chronic disease

Although evidence for measurement properties is important to ensuring scientific rigor in quality assessment, the appropriateness and relevance to the clinical setting and policy context, feasibility of incorporating such measures into routine practice settings, and relevance and interpretation of data to inform quality improvement initiatives, are also important issues in recommending measures for practice: 'the true utility in quality measurement lies in its ability to inspire quality improvement' (Kerr et al., 2001).

REVIEW AIM AND OBJECTIVES

Review aim

To provide guidance to policy-makers, clinicians and researchers on the most appropriate, valid and acceptable patient-reported measures of health care quality, of relevance to long-term physical conditions or chronic disease management, for use in routine practice, clinical audit and research settings.

Objectives

Structured review of published international evidence:

a) to identify patient-reported measures of health service quality of relevance to long-term physical conditions or chronic disease management; measures that are broadly applicable across conditions, include key elements of health care quality, and have been applied in the settings in which care may be received / delivered will be reviewed.
b) to extract and assess evidence relating to the development and evaluation of these measures in relation to pre-defined measurement and practical properties.
c) to make recommendations for the application of patient-reported measures of relevance to health care and service delivery for people with long-term physical conditions or chronic disease. These recommendations will consider evidence for the practicability and viability of patient-reported measures as mechanisms for incorporating the patient voice in routine practice settings.
d) to make recommendations for the further evaluation of measurement performance.
e) to make recommendations for future development of measures where appropriate

METHODS

Search strategy

A structured, but pragmatic approach to identifying and retrieving references for the review was adopted.

The primary search strategy ('main search') was designed to retrieve studies exploring the evaluation of health care quality from the patient perspective, and of relevance to long-term physical conditions or chronic disease, including the development and testing of measures, and evaluation of both measurement and practical properties of particular relevance to 'real-world' application. All searches were restricted to English language publications.

Medline, accessed through Ovid software, was searched for the years 1980-2006 (August). For the main search, terms related to 'health care quality', 'measurement (from the patient perspective)', and 'chronic disease' were employed as illustrated in Table 11.1 below.

The reference lists of all included articles were reviewed for additional articles. The reference lists of existing reviews of patient reported measures or patient completed 'surveys' of relevance to the assessment of health care quality were also reviewed.

Table 11.1 Main search strategy (Medline via OVID software (020806); limits Humans, English)

Health Care Quality / Elements of Health Care Quality (all joined by 'OR')	Measurement (all joined by 'OR')	Chronic Disease (all joined by 'OR')
*"Quality of Health Care"/	exp "Outcome and Process Assessment (Health Care)"/	(chronic adj2 disease$).kf,tw,ti,kw.
*Quality Assurance, Health Care/	(patient adj report$).kf,tw,ti,kw.	(chronic adj2 illness$).kf,tw,ti,kw.
*Health Services Accessibility/	(user adj report$).kf,tw,ti,kw.	(chronic adj2 condition$).kf,tw,ti,kw.
exp Patient-Centred Care/	(client adj report$).kf,tw,ti,kw.	(chronic adj2 ill-health).kf,tw,ti,kw.
exp Patient Satisfaction/	(self adj report$).kf,tw,ti,kw.	(chronic adj2 care).kf,tw,ti,kw.
*Patient Satisfaction/	(consumer adj report$).kf,tw,ti,kw.	(long$term adj2 disease$).kf,tw,ti,kw.
*"Continuity of Patient Care"/	(patient adj2 evaluat$).kf,tw,ti,kw.	(long$term adj2 disorder$).kf,tw,ti,kw.
exp Professional-Patient Relations/	(patient adj2 assess$).kf,tw,ti,kw.	(long$term adj2 illness$).kf,tw,ti,kw.
*Health Facility Environment/	(self adj2 assess$).kf,tw,ti,kw.	(long$term adj2 condition$).kf,tw,ti,kw.
(quality adj4 measure$).kf,tw,ti,kw.	(consumer adj2 assess$).kf,tw,ti,kw.	(musculoskeletal$ adj2 disorder$) .kf,tw,ti,kw.
(quality adj4 assess$).kf,tw,ti,kw.	(patient adj2 question$).kf,tw,ti,kw.	(musculoskeletal$ adj2 disease$) .kf,tw,ti,kw.
(quality adj4 health care).kf,tw,ti,kw.	(consumer adj2 question$).kf,tw,ti,kw.	(musculoskeletal$ adj2 condition$) .kf,tw,ti,kw.
(quality adj4 service).kf,tw,ti,kw.	(patient adj complet$).kf,tw,ti,kw.	(rheum$ adj2 disorder$).kf,tw,ti,kw.
(satisfaction$ adj4 health care).kf,tw,ti,kw.	(consumer adj2 survey$).kf,tw,ti,kw.	(rheum$ adj2 disease$).kf,tw,ti,kw.
(satisfaction$ adj2 care).kf,tw,ti,kw.	(patient adj2 survey$).kf,tw,ti,kw.	(rheum$ adj2 condition$).kf,tw,ti,kw.
(quality adj2 care).kf,tw,ti,kw.	(patient adj2 perspective$).kf,tw,ti,kw.	
(satisfaction$ adj4 service).kf,tw,ti,kw.	(patient adj2 appraisal$).kf,tw,ti,kw.	
(patient$ adj satisfaction$).kf,tw,ti,kw.	(patient adj2 preference$).kf,tw,ti,kw.	
(patient$ adj experience$).kf,tw,ti,kw.	(patient adj2 perception$).kf,tw,ti,kw.	
(experience$ adj4 health care).kf,tw,ti,kw.	(patient adj2 view$).kf,tw,ti,kw.	
(expectation$ adj4 health care).kf,tw,ti,kw.	(patient adj experience$).kf,tw,ti,kw.	
(expectation$ adj2 care).kf,tw,ti,kw.	(consumer adj2 perspective$).kf,tw,ti,kw.	
(experience$ adj2 care).kf,tw,ti,kw.	(consumer adj2 appraisal$).kf,tw,ti,kw.	
(continuity adj2 health care).kf,tw,ti,kw.	(consumer adj2 preference$).kf,tw,ti,kw.	
(continuity adj2 care).kf,tw,ti,kw.	(consumer adj2 perception$).kf,tw,ti,kw.	
(access adj4 health care).kf,tw,ti,kw.	(consumer adj2 view$).kf,tw,ti,kw.	
(access adj2 care).kf,tw,ti,kw.	(consumer adj2 experience$).kf,tw,ti,kw.	
	AND	**AND** **Total: 679**

Footnote: kf – XX; tw – text word; ti – title; kw – key word

These searches were further supported by personal knowledge of the field contributed by members of the review team, consultation with experts in the field, and reviews of web-sites for reviewed measures.

Inclusion criteria

Titles and abstracts of all articles were assessed for inclusion/exclusion. Included articles were retrieved in full. All articles and patient-reported measures were required to satisfy certain criteria of relevance to the study question, patient population, elements of health care quality, type of outcome and language. Moreover, the appropriateness of measures to the UK context was an important consideration. Article and PROM inclusion/exclusion criteria are summarized in Tables 11.2 and 11.3 respectively.

Table 11.2 Article inclusion and exclusion criteria

Articles	
Inclusion	**Exclusion**
1. Published articles providing evidence in support of the development / evaluation / application of patient-reported measures of health care service quality of relevance to the receipt of care for long-term physical conditions or chronic disease management in an adult population 2. Evaluation has relevance to current UK policy context for chronic disease management	1. Evidence of measurement and/or practical properties not reported 2. Assessed outcomes focus on patient experience of disease and not on experience of health care. 3. Focus on the evaluation of care for non-physical chronic conditions – e.g., mental or behavioural health problems. 4. Focus on the evaluation of care for non-adult populations – i.e., paediatrics or adolescents. 5. Non-English language 6. Do not describe evaluative measures in sufficient detail to allow identification 7. Non-published data 8. Narrative reviews

Table 11.3 PROM inclusion and exclusion criteria

PROM	
Inclusion	**Exclusion**
1. Identifiable (and reproducible) multi-item patient-reported measures specific to the evaluation of health care service quality of relevance to long-term physical conditions or chronic disease management (1980-2006). 2. Item content has relevance to the current UK policy context* 2. Data synthesis will focus on PROMs with evidence of at least reliability or validity in the UK setting*	1. Not specific to the evaluation of health service quality 2. Evaluations are specific to health care investigations or interventions – e.g., mammography service 3. Single dimension measures of health care quality – e.g., interpersonal skills or care. 4. Single item measures of health care quality 5. Measure is specific to the experience of health care of relevance to non-physical long-term conditions – e.g.,

	mental or behavioural health. 6. Clinician or proxy (or other) completed measures of health care quality. 7. No evidence of reliability or validity (in UK population)* 8. Measure not clearly identifiable

Footnotes: * The appropriateness of measures to the UK context is an important consideration for inclusion in the review. Where item content has relevance to the UK policy setting, but measures lack evidence of measurement properties in the UK population, relevance to the current policy context will override formal exclusion criteria.

Flexibility in inclusion criteria was considered an essential requirement of the review. For example, where promising measures were identified that, although specific to condition or setting, addressed a wide range of dimensions of relevance to other conditions and appeared to address issues of relevance to current UK health policy, such measures were included in the review.

One reviewer (KH) assessed all returned titles and excluded clearly irrelevant or duplicate items. Borderline studies or measures were discussed with another member of the review team (RF).

Data extraction

Data extraction was informed by a form designed for the purposes of the review, and included both study-specific issues such as study design, and respondent characteristics such as type of chronic illness and age, and measurement specific issues, for example, type and description of measure including the dimension of health care quality covered, response format, extent of patient involvement in development, length, and evidence of measurement and practical properties, such as time to complete and ease of administration and scoring (Fitzpatrick et al., 1998; Haywood et al., 2004).

Evidence for the appropriateness of content to the UK policy context for people with chronic disease was extracted.

Format of the reviews

The summary of the evidence follows that of previous reviews (McDowell and Newell, 1996; Haywood et al., 2004). The following information is provided for each measure:

Title

The measurement title as given by the original developer. Instrument developers, year of original publication, and subsequent revision.

Description
The purpose and proposed application of each measure as defined by the developers.

Development, including item derivation, is summarized where available. Item content, the dimensions of health care quality covered, for example, patient involvement and continuity of care, the number of items, response options, and method of scoring are reported. Measurement modifications are described.

Measurement and practical properties
For all included measures published evidence of measurement properties (reliability, validity, and responsiveness, precision) and practical properties (acceptability, feasibility and interpretation) is summarized.

Review summaries (Discussion)
Reviewed evidence is summarized for each included measure. The nine core dimensions of patient-centred quality health care described by the IoM (Committee on Quality of Health Care in America, 2001) were used to inform a tabulated summary of core dimensions included in the reviewed measures, as shown in Table 11.4. To support comparison between measures, dimension coverage was reviewed against this general classification.

The number of studies in which the measures have been evaluated is provided.

Discussion and Conclusion
The discussion and conclusion to this chapter summarises the current state of health care quality assessment for chronic disease, and suggests areas for future evaluative work.

RESULTS

Search results

The main search returned 679 references. All abstracts were reviewed. When assessed against the inclusion criteria, 86 articles were retrieved and reviewed in full. Checking the reference-lists of included articles and websites generated a significant number of additional articles, and associated measures, that were read and considered for the review.

However, a relatively small final total of 22 articles contributed required evidence of development, measurement and/ or practical properties for the included measures.

a) Identification of patient-reported measures of health care quality

Eleven patient-reported measures of health care quality, of relevance to chronic disease were included in the review, as listed in section 11.5 and Tables 11.4 to 11.8. An additional oncology-specific measure was also included due to its relevance to the review. In addition, although not specific to chronic disease, the General Practice Assessment Questionnaire (GPAQ) was also included for its relevance to the UK policy context (not included in count).

Three organizations are significant within the field of health care evaluation for their development of a range of patient reported measures or surveys; these websites were reviewed for current (and future) developments:

- Consumer Assessment of Health Plans (CAHPS) (USA) https://www.cahps.ahrq.gov/default.asp),
- Picker Institute (USA and Europe) http://www.pickereurope.org/,
- Netherlands Institute for Health Service Research http://nivel.nl. (QUOTE measures)

The work of these groups, and relevance to the review, is summarized in the following sections.

b) Existing reviews of patient-reported measures of health care quality

Three structured reviews of patient-reported measures of health care quality and service delivery were identified (General practice – Wensing et al., 1994; Hospital surveys - Castle et al., 2005; Disease management industry - Sen et al., 2005); these reviews do not refer specifically to the evaluation of patient-reported measures of health care quality of relevance to chronic disease. Two further reviews of measures for the evaluation of quality of care and patient satisfaction were also reviewed assessed (van Campen et al., 1995; Weaver et al., 1997). A literature review of patient reported measures of general practitioner care was reviewed (Sixma and Spreeuwenberg, 2006).

Table 11.4 Dimensions of health care quality (informed by IOM: Committee on Quality of Health Care in America, 2001)

	Core dimensions of Health Care Quality									
Measure (items)	**Respect - patient values, needs / preference**	**Co-ordination / Integration**	**Information, Communication, Education**	**Physical comfort**	**Emotional support**	**Involvement of family / friends**	**Continuity /transition**	**Access to Care (include waiting)**	**Environment**	**Overall impression**
General application across condition and setting (Table 11.5)										
ICICE (>50) ♣	√		√	√				(√)		(√)
PACIC (20) ♣	√	√	√				√	√		
QUOTE-generic	√	√	√		√		√	√	√*	
General application across condition, but specific to setting (Table 11.6)										
Primary Care										
CEP-Q (18) **Dr	√	√	√		√	√	√	√	√	
GPAQ (25) ** DrNR	√	√	√				√	√		
HSHQ (16)	√	√	√				√	√		
SOSQ (21)	??		√					√		
Out-patients										
OPEQ (26)	√	√	√		√		√	√	√	
In-patient										
I-PEQ (40)	√	√	√	√	√	√	√			√
PPE-15 (15)	√	√	√	√	√	√	√			

Core dimensions of Health Care Quality										
Measure (items)	**Respect - patient values, needs / preference**	**Co-ordination / Integration**	**Information, Communication, Education**	**Physical comfort**	**Emotional support**	**Involvement of family / friends**	**Continuity /transition**	**Access to Care (include waiting)**	**Environment**	**Overall impression**
Specific to chronic condition and specific to setting										
Picker MSD (16)	√	√	√		√		√			√
I-PEQ (CHD) (38)	√	√	√	√			√	√	√	
Cancer-specific										
EORTC ** DrNR IN-PATSAT32 (32)	√	√	√	√	√	√	√	√	√	√

Footnotes: Core IoM domains of quality care: respect for patient values; attention to patient preferences and expressed needs (first two domains combined for purpose of review); co-ordination and integration of care; information, communication and education; physical comfort; emotional support; involvement of family and friends; transition and continuity; and access to care.

- ♣ Chronic Care Model informs item content
- ** includes sections specific to evaluation of doctor-related care^{Dr}, nurse-related care^{NR}, services and care organisation, and overall assessment.

INSTRUMENT REVIEWS

The following sub-headings were used to categorise the reviewed measures:

General application across condition and setting (Table 11.5)

Improving Chronic Illness Care Evaluation (ICICE)
Patient Assessment of Chronic Illness Care (PACIC)
(Assessment of Chronic Illness Care (ACIC))
Quality of Care through the Patients Eyes (QUOTE)

General application across condition, but specific to setting (Table 11.6)

Primary care
Clients Evaluate Practice locations Questionnaire (CEP-Q)
General Practice Assessment Questionnaire (GPAQ)*
Health care System Hassles Questionnaire (HSHQ)
Seattle Out-patient Satisfaction Questionnaire (SOSQ)

Out-patients
Out-Patient Experience Questionnaire (OPEQ)

In-patients (and ambulatory care)
Picker Institute
Adult In-Patient Experiences Questionnaire (I-PEQ)
Picker Patient Experiences Questionnaire (PPE-15)
Consumer Assessment of Health Plans (CAHPS)*

Specific application to condition and setting (Table 11.7)

Out-patients
Picker Musculoskeletal Disorder (MSD) Questionnaire
Picker I-PEQ Coronary Heart Disease (I-PEQ (CHD))

Cancer-specific (Table 11.8)

European Organisation for Research and Treatment in Cancer –
In-patient Satisfaction Questionnaire EORTC IN-PATSAT32

** Measures not included in total number of reviewed measures*

GENERAL APPLICATION ACROSS CONDITION AND SETTING

a) Improving Chronic Illness Care Evaluation (ICICE) (Baker et al., 2005 a,b)

The Improving Chronic Illness Care Evaluation (ICICE) was developed in the USA as a comprehensive, patient-reported measure to evaluate the Chronic Care Model (CCM) of chronic illness care; the developers suggest that the model enables one – 'to look inside the black box' and see what elements of the CCM work (Baker et al., 2005a). The multidimensional measure was developed for the evaluation of quality of care across larger population groups to determine the effectiveness of quality improvement activities targeted specifically at groups with chronic illness; it may have particular relevance to research studies.

The ICICE was developed for the ICICE study, which sought to measure the impact of the CCM for several chronic conditions as part of a quality improvement initiative. However, the ICICE model reported by Baker et al., (2005a, b) is applied to the evaluation of a CCM for patients with heart failure only, and hence several items are specific to this condition. The CCM provides the conceptual basis to the measure, including core dimensions such as communication, patient education and information, support for self-management and patient goal setting, and links to community services. Items were generated from already existing measures, experts in the field, and reviews of the literature. Additional items are specific to chronic care management for people with heart failure. Specific involvement of patients and health care professionals is not reported.

The ICICE dimensions are: communication (4 items with 5-point scale; the mean of 4 items is calculated; 3 additional items have yes/no response options); satisfaction (4 modified items from the Consumer Assessment of Health Plans Study (CAHPS) instrument; each item has a 5–point agreement scale; the mean value is calculated); patient education (13 items across 3 condition-specific factors – pathophysiology and treatment; medication adherence; lifestyle modification and weight monitoring; yes/no response); patient knowledge (15 items across three condition-specific factors); and patient behaviours (items related to self-management for heart failure); self-efficacy (3 items relating to self-management relevant to heart failure; 5-point agreement scale; mean value calculated); and health status (generic health status assessed using the SF-12; condition-specific health assessed with the ICICE Heart Failure Symptom Scale; 7 items modified from several other hearth failure-specific measures; 5 or fewer response options per item; score 0-100, where 100 indicates no symptoms), as shown in Tables 11.4 and 11.5. Access to care and overall quality of care (information accessed from patients' medical notes) are also assessed; however, the methodology is not reported in the published literature. Although the total number of items is not clear, there are more than 50 items for the described dimensions.

Administration is via telephone, with an average completion time of 34 minutes. Relevance and comprehension to patients and / health professionals has not been reported. The complete version of the telephone survey is available on-line (http://www.rand.org/health/ICICE/pdfs/chf.pdf.)

Measurement and practical properties (Table 11.9)
There is acceptable evidence of internal consistency reliability, and some evidence of validity for several dimensions following completion by a large group of patients with heart failure, identified from hospital clinics and health plans in the USA. Evidence suggests that the communication dimension may support the detection of differences between groups and improvements in communication over time (although evidence for responsiveness to quality improvement initiatives are limited (Baker et al., 2005b)). The ceiling effects reported for the satisfaction dimension limits the ability of this dimension to detect group differences or temporal trends. The ICICE is a relatively long questionnaire, requiring a significant time period for completion; self-completion has not been assessed.

Discussion
The ICICE represents a generic model for the evaluation of chronic care for long-term conditions, congruent with the Chronic Care Model. Key dimensions are informed by the CCM model and, where appropriate to the model, made specific to the target condition. Hence, although the model is clearly generic across chronic conditions and the provision of care, measurement is specific to conditions.

Although the developers suggest that the number of dimensions included in the ICICE allow for the evaluation of specific elements of the CCM, the full questionnaire is long and resource intensive in completion. It may be more appropriate for health service research settings as opposed to routine practice settings. The ICICE is a relatively new model for evaluation, and evidence of measurement and practical properties are limited, and only assessed in a US population; feasibility in a routine setting has not been explored.

b) Patient Assessment of Chronic Illness Care (PACIC) (Glasgow et al., 2005)

The Patient Assessment of Chronic Illness Care (PACIC) was developed to evaluate the extent to which patients with chronic illness receive care that aligns with the Chronic Care Model (CCM) (http://www.improvingchroniccare.org) (Glasgow et al., 2005a). The developers suggest that there are no comparable patient-reported measures that evaluate the quality of patient-centred care, congruent with the CCM, for people with chronic illness. The measure was developed for application in a variety of health care settings, by individuals with one or more of a range of chronic illnesses.

The CCM emphasises an evidence-based approach to health care that is population-based, patient-centred, proactive and planned. Moreover, care includes key elements of self-management support such as collaborative goal setting, problem-solving and follow-up support (Glasgow et al., 2005a). In developing the PACIC, the CCM framework was evaluated by experts in the field of chronic disease management, and used to inform qualitative interviews with patients. The initial item pool was informed by interviews with national experts in chronic disease management and the CCM from the USA. Items, and earlier versions of the measure, were subsequently piloted and re-tested with patients with one or more chronic disease and further experts to ensure that items were both acceptable and representative of the underlying constructs in the CCM.

The PACIC includes 20 items across 5 dimensions: patient activation/involvement (3 items), delivery system design/decision support (3), goal setting (5 items), problem solving/contextual counselling (4 items), and follow-up/coordination (5 items), as shown in Tables 11.4 and 11.5. Although the CCM defines 6 dimensions of health care quality, issues such as organisation of health care and clinical information systems were omitted from the PACIC due to lack of specific visibility to patients. For each item, patients rate the frequency with which they experienced a particular event / action over the previous six-months, on a five-point scale ranging from 1 (no or never) to 5 (yes or always). Patients evaluate care delivered from their primary health care team for the chronic disease they perceive as impacting most on their life. Items scores are summed and a mean score for each dimension and a total mean score is produced (range 0-20, where 20 is best quality care).

Measurement and practical properties (Table 11.9)

Although a relatively new measure, the PACIC has been completed by large numbers of patients, aged 50 years and over, in the USA with one or more chronic diseases; most commonly hypertension, arthritis, depression, diabetes, asthma, and chronic pain (Glasgow et al., 2005a, b). Early evidence supports high levels of internal consistency reliability (greater than 0.77), but moderate levels of test-retest reliability (three-month retest: range 0.47 to 0.68; overall 0.58). Strong evidence of construct validity, supporting a priori hypotheses, was reported when assessed against other patient-reported measures of health care quality (subscales from the revised Primary Care Experiences Questionnaire; Safran, 2003), and a measure of self-activation which assesses the extent to which patients feel able to take responsibility for their care – an important consideration in chronic disease management (Patient Self-Activation Scale - Hibbard et al., 2004). Data quality was good across all patient groups, with evidence to support the proposed factor structure and no evidence of ceiling effects. The responsiveness of the measure to a quality improvement initiative has not been reported.

Self-completion reportedly required between 2 and 5 minutes; slightly longer for telephone administration (between 7 and 8 minutes). A copy of the PACIC is available from the 'Improving Chronic Illness Care' website: http://www.improvingchroniccare.org

An earlier measure proposed by the development team is the ***Assessment of Chronic Illness Care (ACIC)***, completed by clinicians and health care team members to evaluate the extent to which the 'team' employs elements of the CCM in the routine care of patients (Bonomi et al., 2001). Evidence supports the reliability and validity of the ACIC, and indicates that it is responsive to improvements in the quality of care following CCM-based quality improvements (Bonomi et al., 2001; Wagner et al., 2001). It is recommended that the ACIC is applied alongside the PACIC, providing complementary provider and patient ('consumer') assessments of health care quality for chronic illness. The feasibility of completing and reporting on both measures has not been reported.

Discussion

Unlike other patient-reported measures of health care quality which report on the overall receipt of health care, or the experience of health care which may have relevance to chronic care, for example, including issues such as access and continuity

of care, the PACIC is the only multi-item measure that is specifically aligned to the provision of health care defined by the Chronic Care Model. As such, the measure has good face and content validity for the evaluation of quality in chronic disease management.

Development involved experts in chronic disease management, patients with one or more chronic diseases, and reference to detailed literature reviews, further enhancing content validity. Moreover, evidence of acceptability to patients, and measurement reliability and validity across these patient groups is good. There is limited evidence detailing the feasibility of application; however, it is a relatively brief measure with a simple scoring process.

There is no evidence of measurement responsiveness to change following quality improvement initiatives, and evidence of application in a UK setting is lacking. The PACIC, and the ACIC, are promising measures and warrant further consideration for application in the UK policy context.

c) Quality of Care Through the Patient's Eyes (QUOTE) (van Campen et al, 1998)

A team from the Netherlands Institute for Health Service Research (Nivel: http://nivel.nl) has developed a suite of patient-reported measures designed to understand the patients experience of health care 'through the patients' eyes' (van Campen et al., 1998). The original development of questionnaires took place during the late 1980's, and development continues to date. Both the structure (continuity of care, costs, accommodation, accessibility) and process (courtesy, information, autonomy and competence) of health care service delivery were considered important elements to understanding the patient experience of health care and are included as key dimensions in the multidimensional questionnaires developed.

The QUOTE questionnaires have two sections: the first evaluates patient expectations from health care (how important are specific aspects of care?); the second evaluates an individual's actual experience (perceived experience and problems?). First, patients are asked to rate the importance of several key indicators of health care quality ('Important or not?'). For example, *'Doctors … should be conversant with my health problem'*. The four response options inform the 'importance score': Not important, Fairly important, Important, Extremely important. Second, patients are asked to score their actual experience. For example, *'Doctors … were conversant with my health problem'*. The four response options inform the 'performance score': No, Not really, On the whole yes, Yes.

The measurement of importance acknowledges that patients do not value all aspects of quality similarly. For the purpose of statistical analysis, the quality judgment is equal to the importance score multiplied by the (perceived) performance score.

Each QUOTE questionnaire contains a core generic set of items applicable to a range of users of health care services; the four original QUOTE questionnaires (Rheumatology (Rheum), chronic non-specific lung disease (CNSLD), disabled and elderly) share the same generic set of items (van Campen et al., 1998). Core dimensions include access to care, coordination and integration of care, information

and communication, respect for patient values, preferences and expressed needs, continuity and transition of care, as shown in Tables 11.4 and 11.5.

Additional specific items support the evaluation of health care experience of relevance to specific conditions or patient groups, as listed below. Core and specific items for all questionnaires were informed by detailed qualitative interviews and focus groups with representative patients exploring patients concerns in relation to health care (for example, van Campen et al., 1998); health care professionals were also consulted, and detailed literature searches performed. Involvement of patients ('clients') ensures that the questionnaires are written in plain language that is understandable, supporting face content and acceptability.

QUOTE questionnaires are self-completed and are currently available for the following conditions or patient groups (www.nivel.nl/oc2/page.asp?PageID=5386) (Accessed August 2006):

- Breast Cancer
- Cancer (generic)
- Cataract
- Chronic non-specific lung disease (QUOTE-CNSLD) – asthma and chronic obstructive lung disease
- Diabetes (QUOTE-DM)
- HIV
- Inflammatory Bowel Disease (QUOTE-IBD)*
- Rheumatic Patients (QUOTE-Rheum)
- Elderly people
- Patients undergoing fertility treatment
- Disabled persons (QUOTE-disabled)
- Occupational Therapy Users (QUOTE-OT)*

All questionnaires were developed in Dutch; only the QUOTE-IBD and QUOTE-OT have English translations. Items referring to cost of health care have been removed from the English translations to improve relevance to the UK context. Published evidence of measurement and practical properties for these two measures has not been identified.

Several questionnaires have specific relevance to the current review: QUOTE-Rheum, CNSLD, DM and disabled. Published evidence reporting measurement and/or practical properties for the QUOTE-DM and QUOTE-disabled have not been identified. There is limited published evidence describing the development and initial evaluation (van Campen et al., 1997) or application (Temmink et al., 1999) of the QUOTE-CNSLD (evidence summarized in Table 11.9): high levels of internal consistency reliability and promising evidence in support of measurement validity has been reported (van Campen et al., 1997). The majority of evidence is available for the QUOTE-Rheum.

QUOTE-Rheum

Although several publications describe application of the QUOTE-Rheum in the Dutch population (van Campen et al., 1998; Temmink et al., 2000; Jacobi et al., 2004), few provide evidence of measurement and/or practical properties (van Campen

et al., 1998). The QUOTE-Rheum has been evaluated following completion by patients across a range of inflammatory and non-inflammatory conditions including rheumatoid arthritis (RA), osteoarthritis, low back pain, ankylosing spondylitis and osteoporosis (van Campen et al., 1998). High levels of internal consistency reliability and good evidence in support of measurement validity have been reported.
The questionnaire is suitable for application across the range of health care services accessed by non-institutionalised patients with rheumatic conditions, including care provided by general practitioners, physiotherapy and nursing.

There is little evidence of acceptability to patients (completion rates are not reported) or feasibility of application within a routine practice setting.

Discussion
The concept of a core set of generic items of relevance to the patient experience of health care has intuitive appeal: the needs of different patient groups are similar across a number of core aspects of care. For example, most patients want the opportunity to express their health concerns and to have these taken seriously. Moreover, a core set supports the comparison of health care quality across conditions, settings and hospitals.

However, the model also acknowledges the importance of recognising issues of relevance to specific conditions or population groups in the evaluation of health care quality. For example, access to care and continuity of care are important issues for people with long-term chronic conditions.

The QUOTE programme of work provides an important contribution to the evaluation of health care quality – and there are several interesting developments in the evaluation of heath care quality of relevance to chronic disease. However, there is limited evidence of application in the UK setting. Modification of item content would be required to improve relevance to the UK policy context, particularly with reference to pay for service items. Further evidence for the performance of modules relating to chronic conditions, such as the QUOTE-CNSLD and QUOTE-disabled is also required.

Table 11.5. Patient-reported measures of health service quality: General application across condition and setting

Measure (Developer)	Aim / Focus	Domains (no. items)	Response options	Score	Completion	Setting	other
ICICE Improving Chronic Illness Care Evaluations Baker et al., 2005a,b	To evaluate the impact of Chronic Care Models of care Conceptual model – CCM 'model for a comprehensive instrument to measure all elements of the CCM' – to 'look inside the black box' and see what elements worked Measuring quality of care for larger population groups; determine effectiveness of QI activities	1.a Communication (4) - participatory decision-making; encouragement; perceived interest in pt questions; regular review of mgt plan. 1.b 3 yes/no items re goal setting 2. Satisfaction (4) – modified from CAHPs – satisfaction with Dr / nurse 3. Patient education (13), 4. Knowledge (15) 5. Self-management, 6. Self-efficacy (3) 7..a Health status: generic (SF-12) and 7.b HF specific (7) 8. Access to Care – not reported NR 9. Quality of care – from notes** NR	Items 1a, 2, 6 use 5 point Likert agreement scale (1 strongly agree to 5 strongly disagree) Items 1b, 3,4 5 - Yes/No	1.a Mean of 4 items. 1.b Yes/No items 2. Mean of 4 items 3, 4 5 – % Yes/No 6 Mean of 3 items	Patient interview – telephone survey	Evaluation of CCM model of care – across all settings (tested in primary care) Initial development / evaluation in patients with Heart Failure (HF)	http://www.rand.org/health/projects/icice/pdfs/chf.pdf Single domain scores for domains 1and 2 (but Sat scores very skewed); separate scores for 3 – 6 = considered critical elements of quality care for HF and other chronic diseases. USA only
PACIC Patient Assessment of Chronic Illness Care Glasgow et al., 2005a, b	To assess the extent to which patients with chronic illness receive care that aligns with the CCM To complement the ACIC – providing the patient perspective on receipt of CCM-related chronic illness care	Index: Overall PACIC (20 items) 1. Pt Activation/ Involvement* (3) 2. Delivery System Design / Decision Support (2) 3. Goal setting* / Tailoring items (5) 4. Problem-solving / Contextual* (4) 5. Follow-up / Coordination (5) Scales emphasise pt-HCT interactions – esp aspects of self-mgt support*	5 point categorical / Likert-type (almost never) to (almost always) extent to which actions / care received over past 6mths congruent with CCM	Mean for dimensions ; mean index 0-20	Patient self	All settings in which care can be received; tested in primary care setting	Copy; web-site (www.improvingchroniccare.org) From Improving Chronic Illness Care programme USA only

Measure (Developer)	Aim / Focus	Domains (no. items)	Response options	Score	Completion	Setting	other
PACIC with 5'As Glasgow et al., 2005b	Original PACIC plus 6 items to reflect 5 A's model of behavioural counselling: Ask, Advise, Agree, Assist, Arrange		Addition of 5 A's			Supports evaluation of the receipt of patient-centred care in accordance to CCM framework	
ACIC Assessment of Chronic Illness Care *Compliments PACIC Bonomi et al., 2001	Quality improvement tool to evaluate strengths and weaknesses of care delivery for chronic illness in 6 areas. Focus – organization of care for chronic illness. Aligned with CCM	1.Community linkages 2.Self-management support 3.Decision support 4.Delivery system design 5.Information systems 6.Organization of care		Sub-scales scores 0-11 (11 optimum care)	*Completed by 'Team' NOT patient*	All settings in which care can be provided	Copy in appendix. web-site (www.improvingchroniccare.org/ACIC%20docs/ACICV3.5.pdf)
QUOTE 1.Generic core 2.Specific Rheumatic Patients Chronic non-specific lung disease (CNSPD) Diabetes (DM) Disabled population Netherlands Van Campen et al., 1998	Measure quality of care from the perspective of the non-institutionalised patient. Consider needs, expectations and experiences Multi-dimensional taxonomy: focus structure and process	Total 32 indicators 1.General QUOTE** (16) as: Patient indicators (items 1-16) Structure quality (items 7-14) Process quality (items 1-6, 15, 16) • includes items relating to cost of care • 2.Specific items: Rheum-specific (16) (items 17-32) CNSPD DM Disabled	1.Relative importance (I) rated 4-point scale (unimportant to extremely important) 2.Performance (P) rated 4-point scale (no; not really; on the whole yes; yes)	Q = I x P	Self (postal)	Focus on range of care services used by non-institutionalised patients with rheumatic conditions (RA, AS, OP, OA, LBP) (range home care to specialist care)	Contains common generic and condition-specific items (core of 4 measures developed**) Developed in collaboration with patients Non-English translation

Footnote: ** QUOTE original four population-specific questionnaires have same generic items: Rheumatic patients (Rheum), chronic non-specific lung disease (CNSPD), Disabled (disabled) elderly (elderly).

Instrument Reviews
General application across condition, but specific to setting (Table 11.6)

Primary Care
The review process identified a large number of measures for the evaluation of health care within a primary care setting, for example, the Components of Primary Care Instrument (Flocke, 1997) and the Primary Care Assessment Survey (Safran et al., 1998). However, only three measures specifically considered health care quality from the perspective of an individual with long-term, chronic disease:

Clients Evaluate Practice locations Questionnaire (CEP-Q)
Health System Hassles Questionnaire (HSSQ)
Seattle Outpatient Satisfaction Questionnaire (SOSQ)

a) Clients Evaluate Practice locations (CEP) Questionnaire (CEP-Q) (Wensing et al., 1996a, 1998)

The Clients Evaluate Practice locations Questionnaire (CEP-Q) was developed to enable patients with one or more chronic condition to assess several aspects of health care provided by their general practitioner (GP) within a primary care setting (Wensing et al., 1996a,b). It is suggested that the measure could be used to support a patients' positive or negative assessment of health care, and to stimulate physicians towards improving the quality and outcomes of their care.

The initial item pool was informed by a consensus study with 19 General Practitioners and 34 patients with chronic illness; a further evaluation of items and questionnaire performance involved 249 patients from six general practices within the Netherlands (Wensing et al., 1996a).

The CEP-Q includes 51 items across 9 core dimensions of health care (Vingerhoets et al., 2001): organization of care (9 items), availability for emergencies (3 items), premises/environment (3 items), continuity (4 items), cooperation (4 items), medical Care (6 items), relation / communication (10 items), information and advise (6 items), and support (6), as shown in Tables 11.4 and 11.6. Two broader dimensions are described: first, organisation of care includes items relating to access to care, the environment, continuity of care, and co-ordination between care providers. The second dimension includes items to evaluate the humaneness of the GP, patient involvement, the provision of information, emotional support and support for the patients' network. Patients respond to all items using 6-point descriptive scale ranging from 'poor' to 'very good'. Two items evaluating patients' needs for more care are evaluated on 3-point descriptive scales: 'yes, I like to'; 'it doesn't matter' and 'no, I'd rather not'. For each dimension item scores are summed and a mean score for each dimension produced (range 0-'x', where 0 is poor quality care). A revised 12-item questionnaire has been referred to (Sixma and Spreeuwenberg, 2006), but at the time of writing limited evidence of development and performance was identified.

Measurement and practical properties (Table 11.10)
The CEP-Q was developed in Dutch and has not been evaluated in the UK health care system. Extensive focus groups with patients with chronic illness and general practitioners were described to inform item content (Wensing et al., 1996a). However, limited published evidence of measurement and practical properties was identified.

Discussion
The CEP-Q was first published in 1996, and although early evidence for its measurement and practical properties was promising, there is limited evidence of its application in more recent published studies.

The General Practice Assessment Questionnaire (GPAQ) is recommended in the UK for the evaluation of health care quality of relevance to the primary care setting (http://www.gpaq.info/about%20GPAQ.htm). The GPAQ is summarized in the next section 11.5.2 (b); although not specific to the evaluation of health care for people with chronic disease, reference to the GPAQ in the current review is appropriate for completeness and to reflect the current status of health care quality assessment in the UK primary care setting.

b) General Practice Assessment Questionnaire (GPAQ) (Roland et al., 2006: www.gpaq.info/)

The General Practice Assessment Questionnaire (GPAQ) was developed in the UK and has recently been proposed as the successor to the original General Practice Assessment Survey (GPAS) questionnaire, on which it was based. The GPAQ is a patient-reported questionnaire to inform general practices of what patients think about care provided. It includes multiple aspects of general practice - specifically, including access, interpersonal aspects of care and continuity of care, and is the UK recommended assessment tool for general practice for the new GP contract to inform the 'Quality and Outcomes Framework' (http://www.ic.nhs.uk/services/qof).

Items for the GPAQ were informed by the GPAS, which in turn was developed from the Primary Care Assessment Survey (PCAS) (Safran et al., 1998); the PCAS has been widely used in the United States. The development team from the National Primary Care Research and Development Centre, Manchester, has led the development and validation of the GPAS and GPAQ. The GPAS was widely used in UK-based research to evaluate to quality of primary health care, as perceived by members of the general population attending primary care practices, during the late 1990's and early 2000's (for example, Campbell et al., 2003, 2005; Bower et al., 2003): the GPAQ is a shorter, more easily completed version of the GPAS.

Item content focuses on access, interpersonal aspects of care and continuity of care – items of health care quality not covered by other elements of the GP contract. There are two versions of the GPAQ: the postal version contains 25 items – receptionist (2 items), access (11 items), continuity (2 items), communication (8 items), and practice nurse (3 items). Items relating to the practice nurse are replaced in the 'consultation version' by items relating to enablement (3 items). The majority of items require patients to rate the level of 'excellence' of care on a 6-point adjectival scale ('very poor' to 'excellent'). Several item have 'time' related responses.

Computer software programmes are available to support scoring and data analysis. Mean scores are calculated for each question (each question has more than one item), and expressed as a percentage of the maximum possible score for each 'question' (where 100% is best possible score).

Measurement and practical properties
Evidence for the GPAQ is currently limited; however, the developers anticipate that the measurement properties will be similar to those established for the GPAS (www.gpaw.info/validation.htm). Measurement properties for the GPAS have been reported following completion by random samples of adults attending primary care practices in the UK, and evidence suggests acceptable levels of reliability and validity; there is limited evidence of performance specific to patients with long-term or chronic disease.

Discussion
The GPAQ is designed to be widely applicable to the general population accessing care from their general practice, and across a wide range of health states presenting in a primary care setting. Although not specific to the needs of individuals with chronic ill-health, the wide range of dimensions included in the questionnaire have clear relevance to this patient population group. Increasingly patients with chronic disease receive care within the primary care setting; the appropriateness of questionnaires to the evaluation of health care of relevance to patients with long-term conditions is important to future quality improvement initiatives. Further evidence of measurement and practical properties for the revised GPAQ is required.

c) Healthcare System Hassles Questionnaire (HSHQ) (Parchman et al., 2005)

The Healthcare System Hassles Questionnaire (HSHQ) was developed in the USA to enable patients with chronic illness to report difficulties experienced in accessing care within a primary health care setting; more specifically, to report how the receipt of care for chronic illness is facilitated by the primary care provider (Parchman et al, 2005).

The concept of 'hassles' is further defined as 'troubles' or 'bothers' that patients may experience during their numerous encounters with the health care service; qualitative focus groups with patients with chronic illness raised major concerns with health service interactions. Knowledge of the requirements and experience of people with chronic disease and the increasing role of primary care physicians in providing care highlighted key dimensions for consideration in the provision of health care: access to care, continuity of care, knowledge of the patient by the physician, coordination of care, and communication between patient and physician.

Items for the HSHQ were developed to reflect problems encountered by patients when visiting general health care providers; more idiosyncratic variables such satisfaction with one specific clinic, provider or specific visit, or factors not amenable to change (e.g., cost or facilities) were avoided. 'Hassles' was operationalised as 'the reporting of events', as distinguished from the concept of 'dissatisfaction' which provides an evaluation of these events. An initial pool of 126 items (generation not clear) was piloted with a convenience sample of primary care patients (n= 132). Additional focus groups with 60 patients with two or more chronic illnesses further informed item development and comprehension. The final questionnaire contains 16 items coded on a 4-point descriptive scale, where 0 is 'not a problem' and 4 is 'a very big problem', as shown in Table 11.6; dimensions of care are listed in Table 11.4. Respondents are

asked to rate problems that may make it difficult for them to care for their chronic illness. Items are summed, 0-64, where 64 is the greatest level of hassle experienced.

Measurement and practical properties (Table 11.10)
Initial evaluation of the HSHQ involved participants from a larger study of US primary care (Parchman et al, 2005). Patients had one or more chronic illness. Initial evidence supports high levels of internal consistency reliability (greater than 0.90). Acceptable evidence of construct validity, supporting a priori hypotheses, was reported when assessed against the Components of Primary Care Instrument (CPCI), a patient-reported measure of primary health care quality. The CPCI includes dimensions to assess communication, preference for physician, knowledge of patient, and coordination of care (Flocke, 1997). The types of hassles experienced by patients with single or multiple morbidities was also assessed; patients with multiple chronic illnesses reported more problems with accessing information, problems with medication, and lack of time with clinicians. Evidence suggests that health care system hassles, as measured by the HSHQ, are inversely related to the level of communication and the extent to which care is coordinated within the primary care setting: as coordination and communication improved, the level of hassles also improved.

Data quality was acceptable, although a tendency towards end effects was reported; additional evidence supports the proposed factor structure. The responsiveness of the measure to quality improvement initiatives has not been reported. The questionnaire is brief and would appear to be simple to complete and score; the involvement of patients in item development supports the face validity and acceptability of the measure. However, response rates to the overall survey were low. Evidence of feasibility within a routine setting is not reported.

Discussion
The HSHQ provides a brief report of 'hassle' or difficulties encountered by patients with chronic illness when accessing care in a primary care setting. Development involved a large number of patients with a range of chronic conditions, and items were designed to be generic across conditions and across primary care providers.

Evidence suggests that key components of the HSHQ relate to communication and coordination of care, both essential elements of quality health care in a primary care setting, and for people with chronic disease. The measure may provide a useful resource to inform the reduction of patient 'hassles' during their interaction with the health care delivery system; in particular difficulties with communication and coordination of care in a primary care setting would be highlighted.

d) Seattle Outpatient Satisfaction Questionnaire (SOSQ) (Fihn et al., 2004)

The Seattle Outpatient Satisfaction Questionnaire (SOSQ) was developed to evaluate patient-reported satisfaction with health care provision from primary care providers and outpatient clinics for people with chronic disease (Fihn et al., 2004). The original development involved patients with one or more chronic conditions, including ischaemic heart disease, chronic obstructive pulmonary disease and diabetes. Patients were participants in the Ambulatory Care Quality Improvement Project; a project designed to evaluate if a comprehensive programme of sustained feedback to health

care providers about their patients' general and specific health, and their satisfaction with care, would result in improved health outcomes over time. It appears that the SOSQ was constructed specifically for this trial; all three publications refer to application of the SOSQ in this trial (Fihn et al., 2004; Reiber et al., 2004; Fan et al., 2005b).

The SOSQ has a total of 21 items and consists of two scales taken from available measures of patient satisfaction: 1) the Humanistic Scale (12 items) addresses the personal attributes of the primary care physician, taken from the 23-item American Board of Internal Medicine Patient Satisfaction Questionnaire (Webster, 1989); 2) the Organisational Scale (9 items) addresses issues related to the delivery and organisation of care, for example, access, waiting time, and choice of physician (Tables 11.4 and 11.6). This scale was modified from the RAND Patient Satisfaction Questionnaire. Clarity of item content across the two dimensions is not provided. The involvement of patients and health care professionals in item selection and initial questionnaire development is not reported; hence, evidence for the acceptability, face validity, and potential appropriateness of the questionnaire is unclear. Further detail pertaining to development and initial testing is limited (Fan et al, 2005).

Responses to each item are on a 5-point descriptive scale (from poor to excellent). Items scores are summed into two scales, and transformed to scores ranging from 0 (least satisfied) to 100 (most satisfied).

Measurement and practical properties (Table 11.10)

There is limited evidence for the internal consistency reliability of the SOSQ (Fan et al., 2005a), and only limited evidence of validity following completion by patients with one or more chronic conditions in the USA (Fan et al, 2005a,b). Evidence suggests that for patients with IHD, COPD and DM, patient education and ability to cope with their disease was more strongly associated with patient satisfaction with health care (as measured by the SOSQ), than disease severity. However, evidence for the acceptability of the questionnaire to patients and feasibility for completion in a clinic setting is limited.

Discussion

The availability of a measure to evaluate patient satisfaction with the provision of health care in a primary care setting for people with chronic disease has value and relevance to this review. Although the SOSQ is relatively brief and simple to complete, there is limited evidence supporting the involvement of patients or health care professionals in the initial development and testing of the questionnaire; evidence for the acceptability, content and face validity of the questionnaire is therefore limited. Minimal evidence for measurement and / or practical properties exists.

The SOSQ attempts to provide a multi-dimensional assessment of patient satisfaction with the provision of health care, in a primary care setting, for people with one or more chronic, long-term conditions. However, confidence in the performance of this measure is limited due to poor evidence of development and subsequent evaluation.

Out-patients

e) OutPatient Experiences Questionnaire (OPEQ) (Garratt et al., 2005)

The OutPatient Experiences Questionnaire (OPEQ) was developed in a Norwegian population to provide a patient-report of health care experience suitable for application across a range of out-patient clinic settings (Garratt et al., 2005). The measure was developed for application in a variety of health care settings, by individuals with one or more of a range of somatic conditions. The questionnaire was developed for application across a range of outpatient clinic settings, and was designed to be brief, acceptable to patients and easily completed.

Development involved extensive literature reviews of Anglo-American and Scandinavian literature of relevance to patients' experience of outpatient settings, and was further informed by earlier experience with a Norwegian measure of inpatient experience of health care quality (Patient Experience Questionnaire; Pettersen et al., 2004). Focus groups with health care professionals (doctors and nurses) across a range of outpatient clinic settings – cardiology, gynaecology, neurology, oncology, respiratory medicine, surgery – assessed items and dimensions of health care quality for their relevance to patient experience and the Norwegian setting. Relevance and comprehension was further assessed during qualitative patient interviews.

The OPEQ includes 26 items across 6 dimensions (Tables 11.4 and 11.6): clinic access (2 items), communication (6 items), organisation (4 items), hospital standards/ environment (3 items), information (6 items), pre-visit communication (3 items). Each item has a 10-point scale with descriptive anchors (not detailed); patients report on their experience at the outpatient clinic. Items scores are summed and a mean score for each dimension ('scale') is produced (range 0-100, where 100 is the best experience).

Measurement and practical properties (Table 11.10)

The OPEQ has good evidence for internal consistency reliability, test-retest reliability and validity following completion by a large patient population sample (aged 16 years and above) identified from a range of outpatient clinics, and is recommended as an appropriate measure of patient experiences of outpatient clinics in Norway. Although only moderate survey response rates were reported (53.9%), low levels of questionnaire missing data were reported, suggesting high levels of acceptability.

Discussion

The OPEQ provides a multi-dimensional measure of outpatient experience across a range of clinic settings, and hence across a range of conditions. The views of a range of health care professionals and patients were central to questionnaire development. The OPEQ has been completed by a large number of patients representing a wide range conditions and age range; it appears that both acute and chronic disease populations were included in development and testing. The questionnaire is brief, simple to complete and evidence suggests high levels of patient acceptability. Good evidence of measurement reliability and validity is provided. Evidence of responsiveness to change in health care provision is not available.

Published evidence of a similar measure, generic across conditions and suitable for the evaluation of outpatient experience has not been identified in the UK setting.

Replication of the results for the OPEQ or similar measures specific to the evaluation of outpatient care, of relevance to chronic disease, in the UK health care system is required.

In-patients

f) The Picker Institute

The Picker Institute (USA and Europe) has developed a suite of patient-reported measures designed to seek detailed information relating to a patients experience of health care. The original development of questionnaires took place in the USA during the late 1980's, funded by the Picker/Commonwealth Programme for Patient-Centred Care (www.pickereurope.org). The questionnaires have been used in postal surveys in the UK since 1994.

The questionnaires all address multiple dimensions of health care informed by detailed qualitative interviews and focus groups with patients exploring patients concerns in relation to health care, consultation with experts, and systematic literature reviews (Gerteis et al., 1993; cited by Jenkinson et al., 2002a, b). These dimensions include access to care, coordination and integration of care, information and communication, respect of patient values, preferences and expressed needs, involvement of family and friends, continuity and transition of care. Items are phrased to explore if certain processes or events occurred during an episode of care, and hence reflect patient experience of health care; they do not assess patient satisfaction. Response options are generally 'Yes, completely; Yes, to some extent; No'. For the purpose of statistical analysis, all items are coded as dichotomous 'problem scores', indicating the presence or absence of a problem; a problem is defined as an aspect of health care that, in the eyes of a patient could be improved upon (Jenkinson et al., 2002a).

The Picker questionnaires are self-completed and are currently available for a range of hospital settings (website lists the following settings: outpatients, accident and emergency, maternity, day surgery, primary care, rehabilitation and home care – accessed 21/10/06) and conditions (website lists cancer, heart disease, hip replacement and back pain – accessed 21/10/06). Published evidence for the following questionnaires has been identified and used to inform the current review:

Picker Adult In-patient Experience Questionnaire (I-PEQ)
I-PEQ Coronary Heart Disease (I-PEQ CHD)
(detailed in section 11.5.3)
Picker Patient Experience questionnaire (PPE-15)
Picker Musculoskeletal Disease Questionnaire (Picker MSD Questionnaire)
(detailed in section 11.5.3).

Picker adult In-patient Experience Questionnaire (I-PEQ)

The original Picker adult in-patient experience questionnaire (I-PEQ) contains 40 items across seven dimensions of health care (Tables 11.4 and 11.7), and is suitable for the evaluation of in-patient hospital care across a range of surgical and medical conditions. The original I-PEQ was developed in the USA and has undergone multiple translations. The I-PEQ was modified for a UK population by the removal of items referring to payment for health care; the semantic equivalence of items for the UK audience was also evaluated (Jenkinson et al., 2002a). Good evidence in support

of measurement validity has been reported following completion by a range of UK, including those receiving in-patient care for medical (not detailed), orthopaedic, and surgical 27.9% conditions, and care for older people (Jenkinson et al., 2002a). However, published evidence in support of measurement reliability has not been identified. Acceptable survey response rates have been reported (range 46% USA to 74% Germany; UK 65%) (Jenkinson et al., 2002a) (Table 11.10).

Picker Patient Experience questionnaire (PPE-15)

A shortened version of the I-PEQ, the Picker Patient Experience questionnaire (PPE-15), was published in 2002 (Jenkinson et al., 2002b). The questionnaire contains 15 items and provides a core set of generic items suitable for the evaluation of in-patient health care across different settings (including both planned and emergency admissions); the developers suggest that optional 'specific' modules could be added to the core set of items. However, at the time of writing, published evidence of the development or evaluation of these 'add-on' components has not been identified. Evidence following completion by patients in the UK evaluating their experience of acute in-patient hospital care (conditions not detailed) supports high levels of internal consistency reliability, with promising evidence of both face and criterion validity. Survey response rates were acceptable (65%) (Table 11.10).

Discussion

The core dimensions of health care captured within the Picker questionnaires reflect the key elements of health care quality reported by a range of studies exploring health care quality from the perspective of patients (for example, Coulter, 2005). There is also a strong overlap with the nine core dimensions of care recommended by the Institute of Medicine, as shown in Table 11.4. The extensive involvement of patients and health care professionals in item development and subsequent testing of the range of questionnaires provides good evidence of face and content validity; there is acceptable evidence in support of patient acceptability. Limited evidence suggests acceptable levels of reliability. The extensive experience of Picker Institute Europe in running large scale surveys of patients experience in the UK is of note (www.pickereurope.org).

g) The Consumer Assessment of Healthcare Providers and Systems (CAHPS)

The Consumer Assessment of Healthcare Providers and Systems (CAHPS) programme represents a public-private finance initiative focused towards the development of a suite of patient-completed surveys for the assessment of patient and consumer experience of health care within the USA (https://www.cahps.ahrq.gov/default.asp). The original development of the CAHPS surveys commenced during the mid to late 1990's, and were originally focused towards the standardized evaluation of health plans (CAHPS I). However, development has now extended to cover a wide range of conditions and settings (CAHPS II). The CAHPS programme is funded by the US Agency for Healthcare Research and Quality (AHRQ). Evidence of application of CAHPS surveys in the UK setting has not been identified. However, the extensive development of a range of patient completed surveys of experience within the health care setting is of relevance to this review.

The dimensions of the original CAHPS hospital survey were informed by the Institute of Medicines recommendations for key dimensions of health care: the included dimensions were nurse communication, nursing services, doctor communication, physical environment, pain control, communication about medicines and discharge information. The development of CAHPS surveys involves interviews and focus groups with patients, health professionals and a wide range of stakeholders, further supplemented by extensive literature reviews and information from web-chats and stakeholder meetings. Rigorous and scientific methods are applied to support the development of credible and relevant measures of health care quality (Darby et al, 2005). Patients are invited to report on their experience of health care, rather than the evaluation of satisfaction; response options generally encourage a response indicating whether or not an event / action occurred (most items have 4 or more response options).

Several CAHPS surveys are currently available (as listed on the website); several have specific relevance to the evaluation of health care for people with long-term or chronic disease:

CAHPS People with Mobility Impairments Survey: for the evaluation of health care experience by adults with mobility impairment. This survey may be used as a 'stand-alone' survey or as an additional module within the CAHPS Hospital and Ambulatory questionnaire. However, this is a relatively new survey and evidence of measurement properties is not yet published (website: accessed August 2006). These surveys are designed to inform health care providers of the needs and experiences of patients receiving health care within an ambulatory care setting.

CAHPS Hospital Survey: has been developed to provide a patient reported evaluation of health care experience within medical, surgical or obstetric hospital departments (Darby et al., 2005; Goldstein et al., 2005).

Table 11.6. Patient-reported measures of health service quality: general application across condition, but specific to setting

Measure (Developer)	**Aim / Focus**	**Domains (no. items)**	**Response options**	**Score**	**Completion**	**Setting**	**other**
Primary Care							
CEP-Q Client Evaluates Practice location Questionnaire Wensing et al., 1996a,b	To measure patients opinion about aspects of treatment provided by their GP – of relevance to chronic disease Extensive involvement of health care professionals and patients	51 items 1. Organization of care (9) 2. Availability for emergencies (3) 3. Premises (3) 4. Continuity (4) 5. Cooperation (4) 6. Medical Care (6) 7. Relation / Communication (10) 8. Information and Advise (6) 9. Support	6-point descriptive scale (poor to very good) 2 items use 3 point scale: 'yes, I like to' to 'no, rather not'	Summation across dimensions (0 poor quality)	Patient self	Primary care	Completed by patients with range of chronic conditions to evaluate primary care service Netherlands only
GPAQ General Practice Assessment Questionnaire Roland et al., 2006	Patient evaluation of 9 key areas of primary care activity – supersedes the GPAS Recommended for evaluation of UK primary health care	9 domains, 25 items (2 versions: postal or clinic) 1. Access (11) 2. Continuity of care (2) 3. Communication (8) 4. Practice nurse (3) (postal only) 5. Enablement (3) (clinic only)	Range of response options. Descriptive scales: 'very poor' to 'excellent' Waiting / time for access etc.	Domain scores: summary expressed as percentage (100 best care)	Patient self	Primary care	Specific to evaluation of health care in UK primary care setting Developed under aupices of National Primary Care R&D Centre (www.gpaq.co.uk)
HSHQ Health Care System Hassles Parchman et al., 2005	To measure the level of 'hassles' experienced by patients with regards to care for chronic illness in primary care – *how is receipt of care for CI facilitated etc by PC*	16 items Include items relating to: information; access and waiting; communication; continuity; respect for patient values. Extensive involvement of health care professionals and patients	4-point scale: 0 'no problem' to 4 'very big problem'	Summation 0-64	Patient self	Primary care	Patient report experience; not evaluate experience (not a measure of dissatisfaction) USA only

Measure (Developer)	Aim / Focus	Domains (no. items)	Response options	Score	Completion	Setting	other
SOSQ Seattle Outpatient Satisfaction Survey Fihn et al., 2004	To assess satisfaction with primary care provider in relation to healthcare provision for chronic illness (IHD, COPD, DM)	21 items 1.Organiziational Scale – satisfaction with HC services in internal medicine (x) 2. Humanistic Scale – satisfaction with communication and humanistic qualities of physician (x)	5-point descriptive scale (poor to excellent)	2 summary scales. Transforme d 0 to 100 (100 most satisfied)	Patient self	Primary care	USA only
Out-patients							
OPEQ OutPatient Experiences Questionnaire Garratt et al., 2005	To assess patient experience of hospital out-patient care; completed by somatic patients (not specifically chronic disease)	26 items Clinic access (2) Communication (6) Organisation (4) Hospital standards (3) Information (6) Pre-visit communication (3)	10-point scale; descriptive anchors	Item scores and mean value across domains ('scales')	Patient self	Range of clinics – cardio, gynae, neuro, oncology, respiratory, surgery. Not specify acute / chronic	Developed and completed in Norwegian population only
Picker Questionnaires							
I-PEQ UK Jenkinson et al., 2002a,b	To measure adult experience of in-patient health care Surgical / medical	40 items Information and Education (5) Respect for Patient Preferences (4) Emotional Support (5) Coordination of Care (6) Continuity and transition (4) Physical comfort (5) Involvement of family / friends (3) Overall impression (8)	3 options – Yes (completely / always / to large extent); Yes (to some extent / somewhat); No 'Problem scores' – no problem / problem	40 problem scores (index or item scores)	Postal self-completion	Evaluation of in-patient acute care	Evidence suggests sensitive to change over time, useful for setting priorities for quality improvement and measuring change in care delivery (www.pickereuropse.org/)

Measure (Developer)	Aim / Focus	Domains (no. items)	Response options	Score	Completion	Setting	other
PPE-15 Picker Patient Experience Questionnaire UK Jenkinson et al., 2002b	Core set of items to measure patients' experience of in-patient care Surgical / medical	15 items Information and Education (2) Respect for Patient Preferences (3) Emotional Support (3) Coordination of Care (1) Continuity and transition (3) Physical comfort (1) Involvement of family / friends (2)	3 options – Yes (completely / always / to large extent); Yes (to some extent / somewhat); No 'Problem scores' – no problem / problem	15 problem scores: score 0-16, where 0 is best quality care (index or item scores)	Postal self-completion	Evaluation of in-patient acute care	Short form version of I-PEQ

Instrument Reviews
Specific application to condition and setting (Table 11.7)

a) Picker Musculoskeletal Disease Questionnaire (Picker MSD Questionnaire)
A questionnaire specific to the evaluation of patients' experience of out-patient health care for musculoskeletal disease (MSD) (mainly non-inflammatory neck and back pain), the Picker MSD Questionnaire, was developed in 2002 (Jenkinson et al, 2002d). As typifies development of Picker questionnaires, there was extensive patient involvement in item generation and development. Patients (n=13 patients with back or neck pain) and health care professionals (two physicians, two physiotherapists, a chiropractor and osteopath) contributed to initial item development. Reference was also made to an existing Picker out-patient questionnaire to inform item development; however, at the time of writing further evidence or detail pertaining to this questionnaire has not been identified.

A further 13 patients participated in two focus groups; cognitive interviews were subsequently run with 11 additional patients to explore the item content of the proposed questionnaire. The Picker MSD Questionnaire contains 16 items relating to the patient health care experience of relevance to musculoskeletal out-patient care, as summarized in Tables 11.4 and 11.7. The initial long-form version of the questionnaire included items across nine dimensions: access to care, information and education, respect for patient preferences, emotional support, coordination of care, continuity and transition, overall impression; the final version does not include items in the access to care (Table 11.4). An index score may be calculated, or individual scores across the 16 items.

Measurement and practical properties (Table 11.11)
The initial development and testing of the questionnaire was in a Swedish population (Jenkinson et al, 2002d); there is no published evidence of application and evaluation in a UK population. Initial evidence suggests high levels of internal consistency reliability (Kuder-Richardson 0.86), and promising evidence for measurement face and construct validity. However, survey response rates were relatively low (51%) (mean age 54 years (SD 13.84); range 16 – 88 years). Evidence for the feasibility of application is not reported. The low response rate to the survey is surprising; the small number of items, high level of patient involvement and associated evidence of face and content validity would have suggested a higher rate of completion, comparable to that of other Picker questionnaires. The authors suggest that the low response could have been influenced by the large number of questionnaires included in the survey package, or inaccuracies in the sample frame.

b) Picker In-patient Experience Questionnaire – Coronary Heart Disease (I-PEQ (CHD))
A version of the I-PEQ appropriate to the evaluation of in-patient care for patients with coronary heart disease (I-PEQ (CHD)) (Jenkinson et al, 2002c) has also been developed (Table 11.7). The questionnaire contains 38 items across seven dimensions of care, as shown in Table 11.4. Evidence following completion by UK patients who had received hospital in-patient care for coronary heart disease suggests high levels of internal consistency reliability, with good evidence to support the proposed seven measurement dimensions. There is evidence in support of construct validity. Good

acceptability as evidenced by high completion rates was reported (74%) (Table 11.11).

Discussion – Picker MSD and I-PEQ (CHD)

All but one of the reviewed Picker questionnaires, the Picker MSD, are specific to the evaluation of in-patient hospital health care. Although specific to the evaluation of health care experienced by patients with musculoskeletal conditions, several items in the Picker MSD are specific to neck and/ or back pain, and hence the questionnaire is not suitable for completion by patients with more general chronic or long-term physical conditions. Moreover, the questionnaire does not include items reflecting several core dimensions that, evidence suggests, may be important to people with long-term physical conditions or chronic disease, including access to care (Parchman et al., 2005; Haggulund et al., 2005), physical comfort and involvement of family and friends. At the time of writing, published evidence for the availability and performance of a musculoskeletal out-patient questionnaire referred to in the development of the MSD had not been identified. Although further reference to the availability of questionnaires to assess out-patient health care is made on the Picker Institute website (www.pickereurope.org), further contact with Picker Institute Europe is required to explore the availability and evidence for these measures.

Table 11.7. Patient-reported measures of health service quality: specific application to condition and setting

Measure (Developer)	**Aim / Focus**	**Domains (no. items)**	**Response options**	**Score**	**Completion**	**Setting**	**other**
Picker MSD Q Picker Musculoskeletal Disease Questionnaire Sweden Jenkinson et al., 2002d	To measure experience of health care for patients with musculoskeletal disorders (neck or back pain) Monitor performance of providers; provide direction for improvements in health care delivery Provide info that can be acted upon by providers	16 items Information and Education (2) Respect for Patient Preferences (3) Emotional Support (4) Coordination of Care (3) Continuity and transition (2) Overall impression (2) *Access to care not included in final version Validation only: Overall Satisfaction: 5-point scale (poor to excellent) Recommend clinic to others?: Yes / yes probably / No	3 options – Yes (completely / always / to large extent); Yes (to some extent / somewhat); No Problem scores' – no problem / problem	16 problem scores: score 0-16, where 0 is best quality care (index or item scores)	Postal self-completion	For patients attending OP clinic	Interviews / FGs with HCPs and patients – important aspects of patient experience of HC for MSD Based on existing Picker OP questionnaire
I-PEQ (CHD) UK Jenkinson et al., 2002c	To measure adult experience of in-patient health care specific to Coronary Heart Disease	Total 38 items, 7 domains Information and communication (6) Patient involvement / respect for patient preferences (6) Hospital environment (8) Coordination of Care (7) Discharge and transition (6) Pain / Physical comfort (2) Access (1)	3 options – Yes (completely / always / to large extent); Yes (to some extent / somewhat); No 'Problem scores' – no problem / problem	38 problem scores (index or item scores)	Postal self-completion	Evaluation of in-patient acute care	Evidence suggests sensitive to change over time, useful for setting priorities for quality improvement and measuring change in care delivery (www.pickereuropse.org/)

Instrument Reviews
Cancer-specific (Table 11.8)

Several condition-specific measures of health care quality were identified in the review process, for example, the Diabetes Measurement and Evaluation Tool (Paddock et al., 2000) and the Osteoarthritis Treatment Satisfaction Questionnaire (ARTS) (Pouchot et al., 2005). However, most of these measures included an evaluation of the patient experience of disease as well as condition-specific health care, and were therefore excluded from further review. The EORTC IN-PATSAT32 questionnaire was included in the review due to the extensive involvement of health care professionals and patients with a wide range of cancer diagnoses, and the wide range of health care domains included in the final measure.

a) European Organisation for Research and Treatment of Cancer Quality of Care Patient Satisfaction Questionnaire (EORTC IN-PATSAT32) (Bredart et al., 2004, 2005)

The European Organisation for Research and Treatment of Cancer (EORTC) In-Patient Satisfaction with Care Questionnaire (EORTC IN-PAT32) was developed under the auspices of the EORTC quality of life group to provide a patient-reported evaluation of in-patient hospital based care for people with cancer, including perceived care from hospital doctors and nurses, and aspects of health care organisation and service delivery (Bredart et al., 1998, 2004, 2005). Development took place during the late 1990's and early 2000 and involved international collaboration with health care professionals and patients with cancer from across Europe.

The initial item pool was informed by interviews with oncology experts and patients with cancer from several North European countries including the UK. An existing patient satisfaction questionnaire was also utilised. Items, and earlier versions of the measure, were subsequently piloted and re-tested with patients with cancer and additional experts to ensure that items were acceptable and comprehensive.

The EORTC IN-PATSAT 32 includes 32 items across 11 multi-item and three single item dimensions or 'scales'. Patients are asked to rate doctors in terms of their technical skills (3 items), interpersonal skills (3 items), information provision (3 items) and availability (2 items). Nurses are similarly rated across the same dimensions, with items phrased to represent the nursing role (total 11 items). Additional items ask patients to rate other services and the care organisation, and include items relating to interpersonal skills and information provision (3 items), waiting time (2 items) and hospital access (2 items). The three single items refer to the exchange of information between carers, the level of comfort specific to the environment, and an overall rating of care received. All items use five-point categorical response options – 'poor' through to 'excellent'. All scores are linearly transformed to a 0-100 scale, where a higher score indicates a higher level of satisfaction. Patients are invited to complete the questionnaire at home, within six-weeks of discharge from hospital.

Measurement and practical properties
Although a relatively new measure, the EORTC IN-PATSAT32 has been completed by large numbers of patients from diverse cultural groups across Europe, including the UK; there are multiple translations available. Patients involved in measurement testing, aged 18 years and over, have included a wide range of oncology diagnoses, and experienced a range of medical and/ or surgical interventions (Bredart et al., 2005). Evidence supports high levels of internal consistency reliability (greater than 0.88 for all scales except Hospital Access (0.67)), and test-retest reliability (two-week retest >0.70; single item for general satisfaction 0.66). Strong evidence of construct validity, supporting a priori hypotheses, was reported when assessed against other patient-reported measures of health care quality and quality of life (EORTC Quality of Life Questionnaire (QLQ30)). Data quality was good across all patient groups, with evidence to support the proposed factor structure. There was some evidence of potential ceiling effects for several 'scales' (> 20% of respondents scoring the highest rating). The responsiveness of the measure to a quality improvement initiative has not been reported.

For the majority of patients, self-completion of both the EORTC IN-PATSAT32 and the EORTC QLQ30 required less than 15 minutes (Bredart et al., 2005); completion time for the EORTC IN-PATSAT32 alone has not been reported. Although the majority of patients did not require assistance with questionnaire completion, older patients were more likely to request assistance.

Discussion
Unlike other reviewed measures, the EORTC IN-PATSAT32 is specific to the evaluation of health care quality for patients receiving in-patient care for cancer. However, it supports the evaluation of surgical and / or medical in-patient care, and is appropriate for completion by patients with a range of cancer diagnoses.

Development involved a range of health care professionals in oncology and patients with a range of cancer diagnoses who had received medical and/or surgical treatment. The extensive involvement of patients and health professionals, from a wide range of cultural settings, in addition to reference to existing literature and an existing measure of satisfaction, contributes to the evidence for good levels of content validity. A wide range of dimensions important to the overall concept of health care quality are included in the measure. Moreover, evidence of acceptability to patients following self-completion, and measurement reliability and validity across these patient groups is very good. Although there is limited evidence detailing the feasibility of application, it is a relatively brief measure with a simple scoring process and has been completed across a large number of oncology settings. There is no evidence of measurement responsiveness to change following quality improvement initiatives.

The EORTC IN-PATSAT32 is a well developed measure of in-patient 'satisfaction', or experience, of health care specific to oncology in-patient care. The development and subsequent testing of the measure provides acceptable evidence of both measurement and practical properties. However, evidence of ceiling effects for some items may be a function of the request for respondents to rate levels or satisfaction with or excellence of care provided, as opposed to their experience of care. Although oncology-specific, items are not uniquely tied to oncology. The measure addresses a wide (the widest of all reviewed measures) range of dimensions considered important

in health care evaluation, and clearly of relevance to other long-term conditions; it also includes items of relevance to care provided by different members of the health care team. The measure also appears to address issues of relevance to the current health policy context in the UK.

Table 11.8. Patient-reported measures of health service quality: Cancer-specific

Measure (Developer)	**Aim / Focus**	**Domains (no. items)**	**Response options**	**Score**	**Completion**	**Setting**	**other**
EORTC IN-PATSAT32 Bredart et al., 1998; 2004 ; 2005 Cross-cultural development	Evaluate cancer in-patient perceptions of the quality of medical and nursing care, and the organisation of care and services received during admission to oncology unit	32 aspects of care 1.Doctors: technical skills (3); information (3); interpersonal qualities (3); availability (2) 2. Nurses: : technical skills (3); information (3); interpersonal qualities (3); availability (2) 3. Services: interpersonal quality/information (3); exchange of information (1); waiting time (2); accessibility (2); comfort (1) 4. General satisfaction: global evaluation of care (1) Extensive involvement of health care professionals and patients	5 point scale: poor, fair, good, very good, excellent	Linear transformation 0-100 scale: 100 is most satisfied	Self or interview	Relevance to hospital inpatient experience; completion once discharged to home	Interviews with patient, oncologists; review of satisfaction literature and available questionnaires. Cross cultural (European) development and evaluation. Multiple translations – include English Copy – Bredart 2005

CHAPTER 11: SUMMARY OF EVIDENCE

Table 11.9. General application across condition and setting - summary of evidence.

Study/ Country	***Population (N) Age (years) Method of administration Setting***	***Measurement and Practical properties***					
Improving Chronic Illness Care Evaluations (ICICE)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Baker at al., 2005a	Chronic heart disease (781) Age 62% older than 65yrs Telephone interview	Internal consistency ✓ Test re-test	Construct ✓		✓	✓	✓
Baker et al., 2005b	Chronic heart disease (828) Age Telephone interview	Internal consistency Test re-test	Construct ✓	✓		✓	✓
PACIC							
Glasgow et al., 2005a USA	Chronic illness/disease (266) All with more than 1 chronic condition (Hypertension, Arthritis, Depression, DM, Asthma, Pain) Age: mean 64.2 (10.5). 56% female Self-completed Primary Care	Internal consistency ✓ PACIC 0.93 Domains range 0.77 (Decision) to 0.90 (Contextual) Test re-test ✓ 3-months PACIC 0.58 Domains range 0.52 (Pt Activation) to 0.68 (Coordination)	Construct ✓ Hypothesised relations stated Health service use Number of conditions Other measures				
Glasgow et al., 2005b Diabetes Care USA	Type 2 DM (363) Age: mean 64.1 (11.5). 47% female Self-completed Primary Care	Internal consistency ✓ PACIC 0.96 Test re-test	Construct ✓ Sociodemographic variables - no difference Number of conditions – no difference			✓	

Table 11.9 continued **QUOTE measures**		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
QUOTE – Chronic Non-Specific Lung Disease (CNSLD)							
Van Campen et al., 1997 Netherlands	Total 357 with range of chronic non-specific lung conditions Mean 57yrs (sd 18.6) ; range 15 to 95 54% female Self (postal)	Internal consistency ✓ Index alpha 0.93 General QUOTE: Patient indicators 0.88 Structure quality 0.84 Process quality 0.80 QUOTE CNSLD-specific 0.90 Test re-test	Face and Content ✓ PCA supported factor structure ✓				**Not clear**
QUOTE – Rheumatic Patients (Rheum)							
Van Campen et al., 1998 Netherlands	Total 425 with range of rheumatic conditions (70% RA; 44% OA; 29% LBP; 25% other) Mean 62yrs (sd 14.5) ; range 15 to 95 Self (postal)	Internal consistency ✓ Index alpha 0.92 General QUOTE: Patient indicators 0.84 Structure quality 0.81 Process quality 0.74 QUOTE Rheum-specific 0.88 Test re-test	Face and Content ✓ PCA supported factor structure ✓				**Not clear**

Table 11.10 General application across condition, but specific to setting

Study/ Country	Population (N) Age (years) Method of administration Setting	Measurement and Practical properties					
		Reliability	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
CEP-Questionnaire Clients Evaluate Practice locations Questionnaire							
Wensing et al., 1996a Netherlands	Patients with 1+ chronic conditions (34); GPs (19) Age Focus Groups Primary Care		Item development Face and Content ✓			Extensive involvement of patients and GPs	
Wensing et al., 1996b Netherlands	Patients with 1+ chronic conditions (not severe illness or psychiatric disease) (n 345) Mean 59.5 (sd NR) rge 18->70 Mail (202) or by hand in clinic (143): all self completed Primary Care	Internal consistency ✓ alpha range 0.54 to 0.94				✓ Mail RR 63% Hand RR 72% No significant difference in completion of Q or evaluations of care	
Thoonen et al., 2002 Netherlands	Asthmatics (n= 193) Mean 39.5 (sd 11.5) Self Participants in RCT – tailored education vs. usual care Primary Care Only applied 3 domains (20): 1.Medical Care 2.Relation and Communication 3.Information and Advise			✓ Stat. significant change in score over 6-mths for intervention group; and between groups			
Health care System Hassles Questionnaire (HSHQ)							
Parchman 2005* USA	Chronic illness/disease (422) Age: mean 64.2 (10.5). 56% female Self-completed Primary Care	Internal consistency ✓ >0.90 Test re-test	Face ✓ Construct ✓ Health service use Number of conditions Other measures – CPCI		✓	✓	

		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Seattle Outpatient Satisfaction Questionnaire (SOSQ)							
Fihn et al., 2004 and Reiber et al., 2004 USA	DM-specific population (completed DM-specific, SF-36, SOSQ) n= 1,593 (baseline and 2-yr data) Age mean 65 (SD 10) Self Primary Care - ACQIP			✓ No statistically significant difference in score on either domain: between groups (at baseline or 2yrs) or over time (2 years)			
Fan et al., 2005a USA	Chronic illness (IHD, COPD, DM) (28,689 – returned SOSQ) Age mean 65 (SD 10) Self Primary Care - Ambulatory Care Quality Improvement Project (ACQIP)*	Internal consistency Test re-test	Construct ✓			✓	
Fan et al., 2005b USA	Chronic illness (IHD, COPD, DM) (28,689 – returned SOSQ) Age mean 65 (SD 10) Self Primary Care - Ambulatory Care Quality Improvement Project (ACQIP)*	Internal consistency Test re-test	Construct ✓			✓ 61% response rate	
OutPatient Experiences Questionnaire (OPEQ)							
Garratt et al ., 2005 Norway	Wide range of conditions. Age 55.5 (sd 17.4) Self-completed Hospital out-patient – range of departments	Internal consistency ✓ Test re-test ✓ > 0.70	Face ✓ Construct ✓			✓	
Picker In-Patient Experiences Questionnaire (I-PEQ)							
Jenkinson et al., 2002a UK	In-patients – acute medical / surgical Age range 60.9 (sd18) Postal Acute in-patient care	Internal consistency Test re-test	Face and Content ✓ Participation of patients and HCPs in development Construct ✓ Correlation with overall satisfaction and willingness to recommend			✓	✓

		Reliability	Validity	Responsiveness	Precision	Acceptability	Feasibility
Picker Patient Experiences Questionnaire (PPE-15)							
Jenkinson et al., 2002b UK, Europe, USA	In-patients – acute medical / surgical Age range 60.9 (sd18) Postal Acute in-patient care	Internal consistency ✓ Test re-test	Face and Content ✓ Construct ✓			✓	✓

Footnote:

* Ambulatory Care Quality Improvement Project (ACQIP): randomized controlled trial of feedback/no feedback to provider (Aim: could outcomes be improved by comprehensive feedback?)

Table 11.11 Specific application to condition and setting

Study/ Country	***Population (N) Age (years) Method of administration Setting***	***Measurement and Practical properties***					
Picker Musculoskeletal Disease Questionnaire (Picker MSD Questionnaire)		**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**	**Feasibility**
Jenkinson et al., 2002d Sweden	Patients with Musculoskeletal problems (mainly back/neck) Age 54 yrs (sd 13.84) Postal Patients attending MS clinics	Internal consistency ✓ KR-20 0.86 Test re-test	Face and Content ✓ Participation of patients and HCPs in development Construct ✓ Correlation with overall satisfaction and willingness to recommend			✓	✓
Picker In-patient Questionnaire - Coronary Heart Disease I-PEQ (CHD)							
Jenkinson et al., 2002c UK	Patients with Coronary Heart Disease Age range 25->75 Postal Following in-patient care for CHD	Internal consistency ✓ KR-20 range 0.60 to 0.74 Test re-test	Face and Content ✓ Participation of patients and HCPs in development Construct ✓			✓	

DISCUSSION

Long-term, chronic diseases can have a substantial impact on the well-being of patients. Ensuring the provision of good quality and timely care that is responsive to the needs of patients is essential to patient centred care, and an important challenge to quality improvement initiatives. Recent years have seen an increasing acceptance of the role of patients in evaluating health care; consequently an increasing number of surveys and questionnaires exploring patient satisfaction or experience of health care are now available.

A large number and wide range of patient reported measures were initially identified in the review process, including numerous condition and profession-specific measures, those specific to different health care settings, for example, primary care and hospital in-patient care, and those specific to single dimensions of health care quality, such as continuity of care and access. However, few multi-dimensional measures were identified that were specific to the evaluation of health care quality and of relevance to multiple chronic conditions. Furthermore, there was limited published evidence of application and measurement or practical properties for reviewed measures.

A range of measurement and practical properties were stated *a priori* to inform data synthesis and subsequent recommendations for patient reported measures of health care quality of relevance to a chronic disease setting. Four components were considered key to this review:

> 1) evidence of scientific rigour informed by published evidence of measurement properties;
> 2) the diversity and range of dimensions of relevance to chronic disease and health care evaluation;
> 3) evidence of, or perceived, feasibility of application within a real world setting;
> 4) evidence of, or relevance to, application within a UK health care setting.

Although no single measure was considered outstanding across these four components, several of the reviewed measures have evidence to commend them:

- The conceptual base of the Patient Assessment of Chronic Illness Care (**PACIC**) (Glasgow et al., 2005a, b) was informed by the Chronic Disease Model (CDM); hence, evaluation of health care is judged against a clearly defined conceptual framework. There was extensive involvement of patients with a range of chronic diseases and health care professionals in item generation and subsequent assessment of dimensions. The measure includes several dimensions reflective of key elements of the CDM. The PACIC is a relatively new measure: early evidence suggests acceptable levels of reliability and good evidence of validity as a measure of health care quality, where health care provision aligns with the CDM, following completion by patients with a range of chronic conditions. To date, the measure has only been applied in the US population. The extensive involvement of patients and health professionals is to be commended, resulting in a relatively short (20 item) and focused questionnaire, with good evidence of face and content validity, and acceptability. Although not reported,

these factors are likely to result in good completion rates. However, the strong theoretical background of the PACIC may influence the relevance of the measure to any future application in a UK population; the provision of health care would need to be closely aligned with the CDM. Further evidence of performance with a UK setting is required.

- The **Picker Institute** (Europe) have developed a range of patient reported questionnaires applicable to the evaluation of patients experience of health care within different health care settings, in particular patient experience of hospital in-patient care. Several measures are condition-specific. The range of Picker measures have been widely applied and evaluated within the UK health setting, with good evidence of patient and professional involvement in item development, and promising evidence of measurement properties, acceptability and feasibility. However, at the time of writing, there is little clear evidence of the availability and performance of measures applicable to the evaluation of health care for people with chronic disease, particularly for care received outside of a hospital in-patient setting.

- The Out-Patient Experience Questionnaire (**OPEQ)** provides a multi-dimensional measure of outpatient experience across a range of clinic settings, and hence across a range of conditions. Development and subsequent evaluation has involved large numbers of health professionals and patients representing a wide range acute and chronic conditions and age groups. The questionnaire is brief, simple to complete and with good patient acceptability. Published evidence of a similar measure, generic across conditions and suitable for the evaluation of outpatient experience has not been identified in the UK setting. Replication of the results for the OPEQ, or similar measure specific to the evaluation of outpatient care, of relevance to chronic disease, in the UK health care system is required.

- The **QUOTE** measures include both a patient's expectation from health care ('importance') and actual experience in the generation of a final score. A range of QUOTE questionnaires are available; each measure has a 'generic' set of common dimensions of relevance to a wide range of health care users. Additional condition-specific 'add-on' items are available, including those applicable to rheumatology, chronic non-specific lung disease and 'disabled' patient groups. However, limited evidence of measurement reliability and validity has been identified for the QUOTE-Rheum only; evidence of acceptability and feasibility is not reported. There is no published evidence of completion by a UK population; evidence suggests that modification of item content would be required to improve relevance to the UK policy context. Clarity is lacking with regards to the format of these measures.

- The **Healthcare System Hassles Questionnaire (HSHQ)** involved patients with a wide range of chronic conditions and health professionals in item generation. Although the broad concept of 'hassles' with the receipt of care within a primary care setting was proposed, evidence suggests that the measure more specifically addresses concerns related to communication and co-ordination of care. The HSHQ is a relatively new measure with limited evidence of measurement and practical properties. It has not been applied in the UK setting.

- Although specific to the evaluation of health care for in-patients (receiving medical and/or surgical care) with cancer, the **EORTC IN-PATSAT32** is commended for the extensive involvement of patients and health professionals, across a wide range of cultural settings, in item development and subsequent evaluation. The result is a relatively short measure (32 items) that includes the widest range of dimensions of relevance to health care quality of all reviewed measures; it also includes items specific to care provided by specific members of the health care team: doctor and nurse-specific items. There is promising evidence of measurement properties and feasibility of application, including evaluation within a UK setting. However, patients are asked to rate their level of satisfaction with health care; evidence of potential ceiling effects has been reported.

Although at this time it is impossible to recommend an 'off the shelf' patient reported measure of health care quality that could be recommended for a chronic disease setting, there are several promising developments in the field.

Dimensions of health care quality

The multi-dimensional evaluation of health care quality is recommended to support data interpretation (Weaver et al., 1997; Coulter, 2005) and to inform quality improvement activities (Cleary, 1999). There is clearly a convergence towards key dimensions of relevance to health care quality both generally, and specific to the evaluation of health care of relevance to chronic disease (as summarised in Table 11.4). All reviewed measures include a wide range of dimensions, embracing a broader understanding of health care quality, than was observed in some of the earlier reviews of measures of health care quality (for example, Wensing et al., 1994). Four dimensions are common to the majority of reviewed measures – 1) respect for patient values; 2) co-ordination/ integration of care; 3) information, communication and education, and 4) continuity/ transition of care.

The developers of all reviewed measures have made some attempt to include patients and health care professionals in item generation; several developers also make reference to theoretical and conceptual frameworks for chronic disease management and/or health care quality. The appropriate involvement of patients should enhance the comprehensiveness and relevance of questionnaire content, and is increasingly recognised as an essential component of questionnaire development (Fitzpatrick et al., 1998; Burke et al., 2006).

Types of measure – patient experience

There appears to be general consensus that measures which aim to extract evidence, or reports, of a patient's experience within the health care setting are more reflective of health care quality than measures exploring levels of satisfaction or relative excellence. Evidence suggests that measures addressing satisfaction alone are generally unhelpful and lack discrimination (Wensing and Elwyn, 2003; Street, 2006). Several of the reviewed measures that require patients to indicate their level of satisfaction with elements of health care have evidence of potential ceiling effects; for example, the ICICE and EORTC IN-PATSAT32. Moreover, it is suggested that data interpretation for well developed measures of patient experience is easier and hence more actionable for quality improvement initiatives (Sixma and Spreeuwenberg, 2006).

Reviewed measures that aim to explore patient experience of health care include: the PACIC, the HSHQ, the OPEQ, the Picker suite of measures, the QUOTE measures, and the CAHPS survey questionnaires.

Questionnaire format – ease of completion
Evidence of response, or completion, rates, as a measure of acceptability was not readily available from the majority of publications. However, there is a convergence towards the type of questionnaire likely to be both acceptable and feasible: that is, simple to complete, of acceptable length, and with relevant and meaningful item content. Reviewed measures with better evidence of response rates had extensive patient and health professional involvement at all stages of development.

Measurement properties
The relevance of item content is important to interpretation, and hence to using information to inform quality improvement initiatives. However, few studies report evidence of measurement responsiveness to such initiatives. Further evidence of responsiveness is an important requirement if these measures are to be used to inform quality improvement activities.

CONCLUSION
Fundamentally, the use and interpretation of health outcomes, and associated health outcomes research, is concerned with the evaluation of health care quality (O'Connor 2004). Enabling patients to effectively communicate personal values, priorities and expectations for health care, within the context of long-term chronic disease, to health care providers and to evaluate the relative success of health care are important elements of patient-centred care.

People with long-term chronic conditions experience the receipt of health care across a range of settings. Well developed multi-dimensional measures that capture the range of health care dimensions of relevance to patient-centred care and the needs of people with long-term conditions are essential to informing quality improvement activities. However, the relative benefits of measures that are generalisable across conditions and / or health care settings, versus those that are more specific to condition and / or setting are not clear. The appropriateness of a measure should consider the underlying objectives of any quality improvement initiative, an overriding feature of which should be to facilitate quality improvement efforts (Cleary, 1999).

Overall, there is limited supporting evidence for the patient reported evaluation of health care quality of relevance to chronic disease; and where evidence is available this is generally not available within a UK setting. No single measure fulfilled all requirements of scientific rigour, content, feasibility and relevance to the UK policy context. However, there is growing convergence towards key dimensions of relevance to the provision of good quality health care for individuals with long-term chronic conditions. Moreover, evidence suggests that those measures that aim to evaluate a patients experience of health care provide a more rigorous and interpretable assessment of health care quality than those measures where patients are asked to rate their level of satisfaction with a service. Several measures have promising evidence of measurement and practical properties; the review should inform future development, or where appropriate modification, of patient reported measures. The current review

clearly highlights the need for a well-developed, multi-dimensional, patient-reported measure of health care quality of relevance to chronic disease and the UK policy setting.

REFERENCES

Appleby J, Devlin N. (2004) Measuring Success in the NHS. Using patient-assessed health outcomes to manage the performance of healthcare providers. London, King's Fund.

Baker DW, Asch SM, Keesey JW, Brown JA, Chan KS, Joyce G, Keeler EB. (2005a) Differences in education, knowledge, self-management activities, and health outcomes for patients with heart failure cared for under the chronic disease model: the improving chronic illness care evaluation. *Journal of Cardiac Failure*. **11**(6):405-13.

Baker DW, Brown J, Chan KS, Dracup KA, Keeler EB. (2005b) A telephone survey to measure communication, education, self-management, and health status for patients with heart failure: the Improving Chronic Illness Care Evaluation (ICICE). *Journal of Cardiac Failure*. **11**(1):36-42.

Bodenheimer, T, Lorig K, Holman H, Grumbach K (2002) patient self-management of chronic disease in primary care. *JAMA*. **288**:2469-75

Bonomi AE, Wagner EH, Glasgow RE, Von Korff M (2001) Assessment of chronic illness care (ACIC): a practical tool to measure quality improvement. *Health Service Research*. **37**(3):791-820.

Bower P, Roland M. (2003) Bias in patient assessments of general practice: general practice assessment survey scores in surgery and postal responders. *British Journal of General Practice.* Feb; **53**(487):126-8.

Bredart A, Bottomley A, Blazeby J, et al. (2005) An international prospective study of the EORTC cancer in-patient satisfaction with care measure (EORTC IN-PATSAT32). *European Journal of Cancer*. **41**(14):2120-31.

Bredart A, Mignot V, Rousseau A, et al. (2004) Validation of the EORTC QLQ-SAT32 cancer inpatient satisfaction questionnaire by self- versus interview-assessment comparison. *Patient Education and Counseling*. **54**(2):207-12.

Bredart A, Razavi D, Delvaux N et al. (1998) A comprehensive assessment of satisfaction with care for cancer patients. *Support Care Cancer*. **6**(6):518-23.

Burke L, Stifano T, Dawisha S. (2006) Guidance For Industry - Patient-Reported Outcome Measures: Use In Medical Product Development To Support Labelling Claims. Rockville, MD, U.S Department Of Health And Human Sciences, Food And Drug Administration

Campbell SM, Roland MO, Buetow SA. (2000) Defining quality of care. *Social Science and Medicine*. Dec; **51**(11):1611-25.

Campbell SM, Braspenning J, Hutchinson A, Marshall M. (2002) Research methods used in developing and applying quality indicators in primary care. *Quality and Safety in Health Care.* **11**(4), 358-64.

Campbell SM, Ronison J, Steiner A, (2003) Is the quality of care in general medical practice improving? Results of a longitudinal observational study. *British Journal of General Practice.* Apr; **53**(489):298-304.

Campbell S. Steiner A. Robison J. Webb D. Raven A. Richards S. Roland M. (2005) Do Personal Medical Services contracts improve quality of care? A multi-method evaluation. *Journal of* Health Services *& Research Policy*. **10**(1):31-9.

Castle NG, Brown J, Hepner KA, Hays RD. (2005) Review of the literature on survey instruments used to collect data on hospital patients' perceptions of care. *Health Service Research*. Dec; **40**(6 Pt 2):1996-2017.

Cleary PD. (1999) The increasing importance of patient surveys. *British Medical Journal.* **319**:720-1

Committee on Quality of Health Care in America (2001) Crossing the Quality Chasm: a new health system for the 21st Century. Washington, DC: National Academy Press.

Coulter A. (2005) What do patients and the public want from primary care? *British Medical Journal*. Nov 19; **331**(7526):1199-201.

Darby C, Hays RD, Kletke P. (2005) Development and evaluation of the CAHPS hospital survey. *Health Service Research*. Dec; **40**(6 Pt 2):1973-6.

Davis RM, Wagner EG, Groves T (2000) Advances in managing chronic disease. *British Medical Journal*. Feb; **320**:525 – 526

Department of Health (2004) Improving Chronic Disease Management. Published 03/03/2004. Electronic access (25/10/2006) http://www.dh.gov.uk/PublicationsAndStatistics/Publications/PublicationsPolicyAndGuidance/PublicationsPolicyAndGuidanceArticle/fs/en?CONTENT_ID=4075214&chk=YxS1Yj.

Donabedian A. (1966) Evaluating the quality of medical care. Milbank Quarterly, 44 (Suppl.):166-206

Fan VS, Burman M, McDonell MB, Fihn SD. (2005a) Continuity of Care and Other Determinants of Patient Satisfaction with Primary Care. *Journal of General Internal Medicine*. **20**(5):226-233.

Fan VS, Reiber GE, Diehr P, Burman M, McDonell MB, Fihn SD. (2005b) Functional status and patient satisfaction: a comparison of ischemic heart disease, obstructive lung disease, and diabetes mellitus. *Journal of General Internal Medicine.* May; **20**(5):452-9.

Fihn SD. McDonell MB. Diehr P. et al. (2004). Effects of sustained audit/feedback on self-reported health status of primary care patients. *American Journal of Medicine.* **116**(4):241-8.

Fitzpatrick R (1997) The assessment of patient satisfaction. In: Jenkinson C (Ed) Assessment and Evaluation of Health and Medical Care. Buckingham, UK: Open University Press.

Fitzpatrick R, Davey C, Buxton M, Jones D. (1998) Evaluating Patient-Based Outcome Measures For Use In Clinical Trials. Health Technology Assessment, 2, I-Iv, 1-74.

Flocke SA. (1997) Measuring attributes of primary care: development of a new instrument. *Journal* of *Family Practice*. **45**(1):64-74.

Garratt AM AM, Bjaertnes OA, Krogstad U, Gulbrandsen P. (2005) The OutPatient Experiences Questionnaire (OPEQ): data quality, reliability, and validity in patients attending 52 Norwegian hospitals. *Quality and Safety in Health Care.* **14**(6) 433-7

Gerteis M, Edgman-Levitan S, Daley J, Delbanco T. (1993) Through the Patient's Eyes. San Francisco, CAL Jossey-Bass.

Glasgow RE, Wagner EH, Schaefer J, Mahoney LD, Reid RJ, Greene SM. (2005a) Development and validation of the Patient Assessment of Chronic Illness Care (PACIC). *Medical Care*. May; **43**(5):436-44.

Glasgow RE, Whitesides H, Nelson CC, King DK. (2005b) Use of the Patient Assessment of Chronic Illness Care (PACIC) with diabetic patients: relationship to patient characteristics, receipt of care, and self-management. *Diabetes Care*. **28**(11):2655-61.

Goldstein E, Farquhar M, Crofton C, Darby C, Garfinkel S. (2005) Measuring hospital care from the patients' perspective: an overview of the CAHPS Hospital Survey development process. *Health Service Research*. **40**(6 Pt 2):1977-95.

Groves T, Wagner EH (2005) High quality care for people with chronic diseases. *British Medical Journal.* **330**: 609 - 610

Hagglund KJ. Clark MJ. Hilton SA. Hewett JE. (2005) Access to healthcare services among persons with osteoarthritis and rheumatoid arthritis. *American Journal of Physical Medicine & Rehabilitation*. **84**(9):702-11, Sep

Haywood KL, Garratt AM, Schmidt LJ, Mackintosh AE, Fitzpatrick R (2004) Health status and quality of life in Older People: a structured review of patient-reported health instruments. Report from the Patient-reported Health Instruments Programme to the UK Department of Health, April.

Hibbard, J. (2003) Engaging Health Care Consumers To Improve The Quality Of Care. *Medical Care*. **41**: I61-70

Hibbard JH, Stockard J, Mahoney ER, Tusler M. (2004) Development of the Patient Activation Measure (PAM): conceptualizing and measuring activation in patients and consumers. *Health Service Research*. Aug; **39**(4 Pt 1):1005-26.

Hibbard JH, Mahoney ER, Stockard J, Tusler M. (2005) Development and testing of a short form of the patient activation measure. *Health Service Research*. Dec; **40**(6 Pt 1):1918-30.

Jacobi CE, Boshuizen HC, Rupp I, Dinant HJ, van den Bos GA (2004).Quality of rheumatoid arthritis care: the patient's perspective. *International Journal of Quality in Health Care*. Feb; **16**(1):73-81

Jenkinson C, Coulter A, Bruster S, Richards N, Chandola T. (2002a) Patient experiences and satisfaction with health care: results of a questionnaire study of specific aspects of care. *Quality and Safety in Health Care*. **11**:335-339

Jenkinson C, Coulter A, Bruster S. (2002b) The Picker Patient Experience Questionnaire: development and validation using data from in-patient surveys in five countries. *International Journal for Quality in Health Care*. Oct; **14**(5):353-8.

Jenkinson C, Coulter A, Bruster S, Richards N. (2002c) The coronary heart disease in-patient experience questionnaire (I-PEQ (CHD)): results from the survey of National Health Service patients. *Quality of Life Research* Dec; **11**(8):721-7

Jenkinson C, Coulter A, Gyll R, Lindstrom P, Avner L, Hoglund E. (2002d) Measuring the experiences of health care for patients with musculoskeletal disorders (MSD): development of the Picker MSD questionnaire. *Scandinavian Journal of Caring Sciences.* Sep; **16**(3):329-33

Kerr EA, Krein SL, Vijan S, Hofer TP, Hayward RA. (2001) Avoiding pitfalls in chronic disease quality measurement: a case for the next generation of technical quality measures. *American Journal of Managed Care*. Nov; **7**(11):1033-43.

Kendrick, S. (2001) Using all the evidence: towards a truly intelligent National Health Service. Clinical Indicators Support Team (Chief Medical Officer, Scotland), http://www.show.scot.nhs.uk/indicators/Publications/Evid.htm (Accessed: August, 2006).

Leatherman, S. Sutherland, K. (2003) The Quest For Quality. A Mid-Term Evaluation Of The Ten-Year Quality Agenda. The Nuffield Trust

Mitchell, P, Lang, N. (2004) Framing the Problem Of Measuring and Improving Healthcare Quality: Has The Quality Health Outcomes Model Been Useful? *Medical Car*e, **42**, Ii4-11.

McDowell I, Newell C (1996) Measuring Health – a guide to rating scales and questionnaires. Oxford University Press, Oxford. Second Edition.

O'Connor, R. (2004) Measuring Quality of Life in Health, Churchill Livingstone, London.

Paddock LE, Veloski J, Chatterton ML, Gevirtz FO, Nash DB. (2000) Development and validation of a questionnaire to evaluate patient satisfaction with diabetes disease management. *Diabetes Care*. Jul; **23**(7):951-6.

Parchman ML, Burge SK; Residency Research Network of South Texas Investigators. (2002) Continuity and quality of care in type 2 diabetes: a Residency Research Network of South Texas study. *Journal of Family Practice*. Jul; **51**(7):619-24.

Parchman ML, Noel PH, Lee S. (2005) Primary care attributes, health care system hassles, and chronic illness. *Medical Care*. Nov; **43**(11):1123-9

Pettersen KI, Veenstra M, Guldvog B, Kolstad A. (2004) The Patient Experiences Questionnaire: development, validity and reliability. *International Journal for Quality in Health Care*. Dec; **16**(6):453-63.

Pouchot J, Trudeau E, Hellot SC, Meric G, Waeckel A, Goguel J. (2005) Development and psychometric validation of a new patient satisfaction instrument: the osteoARthritis Treatment Satisfaction (ARTS) questionnaire. *Quality of Life Research*. Jun; **14**(5):1387-99

Powell AE, Davies HT, Thomson RG. (2003) Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls. *Quality and Safety in Health Care*. **12**(2), 122-8.

Reiber GE, Au D, McDonell M, Fihn SD. (2004) Diabetes quality improvement in Department of Veterans Affairs Ambulatory Care Clinics: a group-randomized clinical trial. *Diabetes Care*. May; 27 Suppl 2:B61-8.

Roland, M. (2006) General practice assessment questionnaire (GPAQ). http://www.gpaq.info/team.htm.

Safran DG Kosinski M, Tarlov AR, Rogers WH, Taira DH, Lieberman N, Ware JE. (1998) The Primary Care Assessment Survey: tests of data quality and measurement performance. *Medical Care*. May; **36**(5):728-39.

Safran DG (2003) Defining the future of primary care: what can we learn from patients? *Annals of Internal Medicine*. Fe 4; **138**(3):248-55

Sen S. Fawson P. Cherrington G. Douglas K. Friedman N. Maljanian R. Fitzner K. Tang P. Soper S. Wood S (2005). Patient satisfaction measurement in the disease management industry. *Disease Management*. Oct; **8**(5):288-300.

Sixma H, Spreeuwenberg P (2006) Chapter 24: Quality of general practitioner care from the patients' perspective: facts, trends and differences. IN: Westert G, Schellevis FG (Eds) Morbidity, Performance and Quality in Health Care: Dutch General Practice on Stage. Radcliffe Publishing Ltd., Oxford.

Sheldon T. (2005) The healthcare quality measurement industry: time to slow the juggernaut? *Quality and Safety in Health Care*. Feb; **14**(1):3-4.

Street A. (2006) Future of quality measurement in the National Health Service. Expert *Reviews in Pharmacoeconomics and Outcomes Research*. **6**(3):245-248

Swan BA, Boruch RF. (2004) Quality of evidence: usefulness in measuring the quality of health care*. Medical Care*. Feb; **42**(2 Suppl):II12-20.

Temmink D, Hutten JB, van der Zee J, Huyer Abu-Saad H. (1999) Dutch nurse clinics for children with asthma: views of professionals and parents. *Journal of Nursing Care and Quality.* Oct; **14**(1):63-71.

Temmink D, Hutten JB, Francke AL, Abu-Saad HH, van der Zee J. (2000) Quality and continuity of care in Dutch nurse clinics for people with rheumatic diseases. *International Journal for Quality in Health Care*. Apr; **12**(2):89-95.

Thapar AK, Roland MO (2005) General practitioner attitudes to the care of people with epilepsy: an examination of clustering within practices and prediction of patient-rated quality of care. *BMC Family Practice.* Mar 1; **6**(1):9.

van Campen C, Sixma H, Friele RD, Kerssens JJ, Peters L. (1995) Quality of care and patient satisfaction: a review of measuring instruments. *Medical Care Research and Review.* **52(**1):109-133

van Campen C, Sixma HJ, Kerssens JJ, Peters L. (1997) Assessing non-institutionalized asthma and COPD patients' priorities and perceptions of quality of health care: the development of the QUOTE-CNSLD instrument. *Journal of Asthma.* **34**(6):531-8

van Campen C, Sixma HJ, Kerssens JJ, Peters L, Rasker JJ. (1998) Assessing patients' priorities and perceptions of the quality of health care: the development of the QUOTE-Rheumatic-Patients instrument. *British Journal of Rheumatology*. Apr; **37**(4):362-8.

Vingerhoets E, Wensing M, Grol R (2001) Feedback of patients' evaluations of general practice care: a randomised trial. *Quality in Health Care*. **10**:224-228

Webster G. (1989) Final report on the Patient Satisfaction Questionnaire Project. Philadelphia, PA> American Board of Internal Medicine Committee on Evaluation of Clinical Competence.

Wagner EH. Glasgow RE. Davis C. Bonomi AE. Provost L. McCulloch D. Carver P. Sixta C. (2001) Quality improvement in chronic illness care: a collaborative approach. *Joint Commission Journal on* Quality Improvement. **27**(2):63-80.

Weaver M. Patrick DL. Markson LE. Martin D. Frederic I. Berger M. (1997) Issues in the measurement of satisfaction with treatment. *American Journal of Managed Care.* **3**(4):579-94.

Wensing M, Grol R, Smits A. (1994) Quality judgments by patients on general practice care: A literature analysis *Social Science & Medicine* **38**(1); January: 45-53.

Wensing M. Grol R. van Montfort P. Smits A. (1996a) Indicators of the quality of general practice care of patients with chronic illness: a step towards the real involvement of patients in the assessment of the quality of care. *Quality* in *Health Care*. **5**(2):73-80.

Wensing M, Grol R, Smits A, Van Montfort P. (1996b) Evaluation of general practice care by chronically ill patients: effect of the method of administration. Family *Practice*. Aug; **13**(4):386-90.

Wensing M, Grol R, van Weel C, et al. (1998) Quality assessment using patients' evaluations of care. *European Journal of General Practice*. **4**: 155-8. *Article not received (cited by Vingerhoets et al, 2001 as development article for CEP-Q).*

Wensing M, Elwyn G. (2003) Methods for incorporating patients' views in health care. *British Medical Journal*. Apr 19; **326**(7394):877-9.

Chapter 12 DISCUSSION

An enormous array of patient-reported health instruments have been reported in the literature. At least 1275 different instruments were identified a few years ago and the number will have grown substantially since then (Garratt et al., 2002). This makes the task of selecting instruments for any given purpose challenging. The reviews reported here have used standard and fairly widely accepted criteria to assess instruments, basically focused on measurement properties and evidence of practical feasibility to reach, wherever possible, specific recommendations. As emphasised in the Introduction, whilst such criteria are widely accepted and to a large extent can be expressed in operational terms, how a review should weigh up the sum of evidence available has been less extensively discussed and agreed. To take an example, how should a review assess the overall merits of an instrument if there is extensive positive evidence supporting the reliability and responsiveness of an instrument but a small amount of weak evidence indicating poor response rates? There are problems of weighing up contrasting positive and negative features of performance of instruments on different criteria, weighing the volume versus quality of evidence, possible biases against reporting negative evidence, the likelihood of more long-established instruments accumulating more favourable evidence and so on. In the end considerable judgement is unavoidably involved in assessing overall performance of one instrument with another.

Despite such unresolved difficulties in the assessment of health instruments, for the most part it was possible to reach reasonably confident recommendations, with some caveats, for each of the conditions. In some reviews, for example, diabetes, there did not seem to be sufficient evidence to highlight one diabetes-specific instrument over others. It is not surprising that for each of the six specific conditions, the largest amount and best quality of supportive evidence was found for SF-36 as a generic health instrument. It is generally the most widely used and most extensively examined of instruments, to capture broad aspects of health in general populations as well as in studies of specific health problems.

It is frequently recommended that a generic instrument should be used in combination with a disease- or condition-specific instrument when assessing the health problems of individuals with a particular condition, so that both broad general features of health and rather more specific problems are equally captured in assessment. It is assumed here that that is the optimal strategy for group-based uses in the NHS whenever it is feasible and especially when change over time is an important issue. Generally the evidence is that disease-specific instruments are more sensitive to changes within patient groups with specifically identified health problems (Wiebe et al., 2003). Most of the recommendations made in this report are compatible with such a strategy. No very long disease-specific instruments have been recommended, so that being combined in a battery with, for example, SF-36 is feasible. However in a small number of instances it does not make sense to use some specific combinations of generic and disease-specific instruments, for example where the content of the disease-specific instrument has been partly or largely derived from a generic instrument and there would be resulting repetition if used in a battery. An example is discussed in the chapter on epilepsy.

It may be an important goal to assess utilities or values regarding health, for example in the context of an economic evaluation. The reviews of long term conditions presented here have found encouraging evidence for instruments such as EQ-5D and HUI, designed and developed to assess utilities. Instruments such as EQ-5D are also quite short and hence appropriate to be used in combination with disease-specific instruments. Most recently, it has been argued that an alternative strategy is to use disease-specific instruments but to derive appropriate utilities for the various health states described in them. There is nothing inherently difficult in such a 'hybrid' solution and examples of utility-based disease-specific instruments have begun to appear (Wasserman et al., 2005; Casey et al., 2006). However there was little evidence of such work in the conditions specifically reviewed here.

The reviews have tried to give as much attention as possible to whether instruments are feasible to use on a routine or regular basis, for example taking account of evidence response rates, respondent burden and acceptability. Unfortunately these aspects of instruments are not remotely as commonly assessed and reported as are traditional measurement properties. Typically the only direct evidence available are matters such as the number of questionnaire items, any apparent complexity of format, and whether or not a trained interviewer is required for instrument administration. The instruments recommended in this report are for the most part suitable for respondents to self-complete without supervision, for example in a postal survey. This is usually the most cost effective form of administration whilst potentially sacrificing some of the more favourable response rate that can be obtained if a questionnaire is either administered by an interviewer or under supervision. This is supported in a study by Duncan et al., (2005) where it is reported that telephone administration was twice as expensive as postal administration. However, no actual costs were reported.

The simplest and most quantitative expression of an instrument's overall performance in terms of feasibility and acceptability, as stated in the individual reviews, is particularly poorly reported and operational definitions of response rates in any case vary. Reports of response rates in research studies do not necessarily generalise to the routine context. It was not possible to assess relative performance of instruments of instruments in terms of response rates. Nevertheless the general literature provides some useful general guidance on factors that may impact on response rates for patient reported health instruments.

A systematic review of randomised controlled trials to improve response rates to questionnaires across a broad spectrum of contexts including health, shorter length of questionnaire improved response rate (Edwards et al., 2002); A subsequent report from this group showed that the greatest improvement in response rates were achieved by very short questionnaires (Edwards et al., 2004). A second, independent systematic review of trials to improve response rates to questionnaires in health research confirms the importance of shorter questionnaires (Nakesh et al., 2006). A review of response rates to patient satisfaction surveys showed modest improvements to responses by shorter questionnaires (Sitzia and Wood (1998).

Two trials illustrate potential implications for patient-reported health instruments. A randomised trial compared response rates in patients with stroke to (the shorter) EQ-

5D compared with SF-36 and obtained a 5% better response rate with EQ-5D (Dorman et al., 1997). Similarly Iglesias and Torgerson (2000), in a survey of women aged 70 years and older conducted through general practices, randomised respondents to receive a questionnaire that included the EQ-5D and (a longer) SF-12. A difference of 9% was observed in favour of the EQ-5D.

Other aspects of the overall design and conduct of a survey may make a contribution to the response rate, over and above specific instruments selected (McColl et al., 2001). Personally conducted interviews may not necessarily result in higher response rates but do tend to have less missing data (Smeeth et al., 2001). Personalised letters, reminder letters, stamped return envelopes and telephone contact may also positively impact on reponse rates (Edwards et al. 2002; Nakesh et al., 2006.).

It should also be noted that characteristics of respondents have an important effect on response rates. For example, older respondents do appear to have difficulties with questionnaires such as SF-36, especially if they have additional physical or psychological morbidity and impairments (Malison 1998; Parker et al., 2006). Overall it is difficult, from the available evidence to separate out evidence of instrument-specific effects on response rates from the many broader determinants.

Even in contexts where it is possible to automate collection, processing and analysis of information gained from patient-reported health instruments, for example by use of computer interface, the costs of analysis and interpretation of information to relevant audiences should not be neglected. There is some evidence supporting electronic administration of the SF-36 in terms of patient preference and shorter completion times but no details were provided of costs, acceptability to staff and training needed (Caro et al., 2001). In a study by Bendsten et al., (2003 (Sweden)), the feasibility of implementing a computerised system for collecting and analysing patient responses to the SF-36 with patients with COPD reported the thoughts and attitudes among physicians of the utility of the results. Patients completed the SF-36 prior to consultation and the physician reviewed the results following the interview. The physicians rated the patient's health status and then compared the patient's assessment with their own. While there was correlation between physician and patient's responses, the physicians reported that the information from the SF36 did not provide any new information or lead to further clinican decisions. The physicians in this study embraced the concept of incorporating patient's perspectives of their health at an individual level, but there was more interest in incorporating such measurement for the purposes of group evaluation and quality improvement initiatives.

Unlike traditional measurement performance which is reported in quantitative data, feasibility and patient acceptability has been explored qualitatively with focus groups with physicians and patients. Several barriers have been reported and McHorney et al., (2002) illustrates potential and actual barriers to the completion of the SF-36 and a disease-specific instrument at home prior to an appointment at an asthma clinic. Patients reported a lack of feedback from physicians about their responses but they did prefer completing the instruments at home. Interestingly, patients in this study expressed preference for instruments tailored to their specific problems and there was resistance to the inclusion of mental health questions. Patients though could see the benefit of collecting such information from a physician and healthcare provider perspective and also for their own benefit increasing self-awareness of their health in

general. The physicians in this study could see the benefits of using such measurement particularly with people with chronic illness promoting a more holistic view of the patient and the information provided extra protection for documented consultation. Further benefits perceived by the physicians related to the potential marketing value in terms of patient satisfaction with services and for evaluation of services. Again there was preference for disease-specific instruments. Also, the. physicians in this study were concerned about several issues echoed in other studies regarding the economical impact of implementing such assessment as well as the organisational aspects in terms of staff training, real time analysis of data and score interpretation.

It has been suggested that the use of patient-reported health instruments is a useful strategy for improving collaboration between patient and providers and as a screening tool. Espallargues et al., (2000) suggests though that using patient-reported instruments as a screening tool is minimally collaborative. The use of instruments as a strategy to improve communication suggests that it increases topics discussed and has a positive effect on provider behaviours. However, there is little evidence to suggest this method results in better patient reported-health overtime (Espallargues et al., 2000).

Greenhalgh (1998) describes criteria for assessment of instruments for use in clinical practice. Whilst there is still emphasis on measurement performance, it is suggested that there may be a trade-off between psychometric measurement standards and practical aspects of feasibility and clinical utility. For example, some instruments are lengthy and take 15-20 minutes to complete. Practically this may not be feasible in a clinical setting. Patients may require assistance or find completing questionnaires too burdensome. Shorted instruments which have been developed from longer parent versions may be more acceptable to both patients and clinicans with the added benefit of less time for completion and analysis. The SF-12 is an example of a generic instrument and the MiniAQLQ asthma-specific. Greenhalgh (1998) points out that feasibility and clinical utility should have higher priority than traditional quantitative measurement properties when selecting an instrument for use in clinical practice.

The evaluation of services is an important feature of quality management in the NHS. A whole systems approach to care is being advocated including not only the measurement of patient-reported health outcomes, but the use of services, satisfaction with care and user involvement in defining quality. In a study by Steinwachs et al., (1994), as part of a quality improvement programme, evaluation of the feasibility of implementing an outcomes management system for patients undergoing cardiac angiography and patients with asthma across 13 different types of organisations was examined. Outcomes management in this study was defined as a systematic approach to collecting information on the impact of medical care on patients' health outcomes. Feasibility was defined as successful collection of outcomes data; is the information collected reliable, valid and discriminative; and will the information provide useful predictors of outcomes to improve the overall quality of care. Response rates were higher for physicians than patients in this study and no differences were observed between different organisations. There was a positive relationship between response rates and data collection and staffing levels especially where there was committed personnel allocated to collecte data with a specified protocol to increase responses

from patients. This study also found that those patients who were satisfied with their care also had a positive change in their health status.

Whilst many of the instruments included in this review for people with chronic conditions are suitable for self-completion individualised components/domains within instruments may prove difficult for some patients (Garratt 2000).

Several barriers have been asserted with reference to successful implementation of patient-reported health measurement. Complex scoring systems advocated by the developers may be an important barrier and specific training is required. Jenkinson (2000) argues that for the London Handicap Scale simple summation yields almost identical information to the complex weighed scheme suggested by the developers.

The greater challenge is to ensure that information gathered by this methodology is made meaningful to those audiences. Although increasing effort is put into interpretability of health status scores, it remains one major barrier to wide-spread up-take. To date evidence demonstrating that information from patient-reported outcomes can make a difference to individual clinical practice and health outcomes is not persuasive (Gilbody et al., 2002; Greenhalgh et al., 2005). It is encouraging though that there is evidence from several exploratory studies that both physicians and patients embrace the concept of measurement. Of great interest would be the somewhat different question of whether such information can be used to make a difference at the system level when used to assess quality and performance of services. Such evidence is even more lacking.

For PHI's to be implemented successfully there needs to be a cultural acceptance and strong clinical leadership; financial resources; specific training for staff; and time allocated.

More generally, it needs to be emphasised that users, clinicians, patients, managers and service providers will have legitimate and valuable views about the relative importance and appropriateness of instruments for any given application. In addition to the evidence assembled in reviews such as those reported here, a full appraisal of the relative value of instruments for a given task must take account of stake-holders judgements of the fit between instruments and specific intended uses. This element of judgement about appropriateness and relevance to a given context cannot be assessed in the same way as formal measurement properties. 'Appropriateness' is one of the key criteria emphasised by Fitzpatrick and colleagues (1998). This criterion has to rely on users' judgements of the degree of fit of the content of an instrument to a specific intended application; something that, a priori, cannot be determined by reviews of formal measurement properties.

This report has also assessed instruments to involve individuals with long term conditions in assessing the quality of their care. Although the field of involving patients and users in assessing quality of care (measures of patient satisfaction, patient experience etc) is as long-standing as that of health status measurement, it has not evolved in the same way. Few if any instruments have emerged to dominate and there is fairly constant flux of instruments. It is not surprising that this should be the main pattern also to emerge from the current review, focusing on instruments for use with patients with long term conditions. A likely reason for the absence of instruments

emerging with clearly superior measurement properties is that instruments are required to address patients' experiences of diverse, specific contexts and services (which in turn constantly change their forms). Instruments end up being developed for specific dedicated purposes and cannot be applied to different settings. The specific questions and concerns of those who commission such work may also vary enormously.

Nevertheless the review did provide important, useful insights for how patients and users with long term conditions might be involved in assessment of service quality. There are commonalities in the domains and topics important to individuals with long-term conditions across the more promising instruments found in the review. Also it is clear that such experiences can be addressed via standard self-complete survey instruments.

Somewhat similar observations may be made about carer impact. Although no instruments emerged clearly to dominate assessments of relative performance, it is clear that there are some possibly promising instruments. There is also some convergence in terms of the range of domains and topics of concern to carers of individuals with long term conditions.

Overall it is hoped that this report provides a clear and encouraging body of evidence to increase the contribution that patients, service users, carers and the public can make to the evaluation of services in the UK.

REFERENCES

Preben Bendtsen, Matti Leijon, Ann Sofie Somme, Margareta Kristenson. Measuring health-related quality of life in patients with chronic obstructive pulmonary disease in a routine hospital setting: Feasibility and perceived value *Health and Quality of Life Outcomes* 2003, **1:**5

Caro JJ, Caro I, Caro J, Wouters F, Juniper EF. Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research* 2001; **10**:683-91.

Casey R, Tarride JE, Keresteci MA, Torrance GW. The Erectile Function Visual Analog Scale (EF-VAS): a disease-specific utility instrument for the assessment of erectile function. *Can J Urol.* 2006;**13**:3016-25

Dorman P, Slattery J, Farrell B et al., A randomised comparison of the EuroQol and Short-Form 36 after stroke. *BMJ* 1997; 315-461.

Duncan P, Reker D, Kwon S, Lai S, Studenski S, Perera S et al.,. Measuring stroke impact with the Stroke Impact Scale: telephone versus mail administration in veterans with Stroke. *Medical-Care* 2005; **43**: 507-15

Edwards P, Roberts I, Clarke M, et al., Increasing response rates to postal questionnaires: a systematic review *BMJ* 2002; 324: 1183-5.

Edwards P, Roberts I Sandercock P, Frost C, Follow-up by mail in clinical trials: does questionnaire length matter? *Controlled Clinical Trials* 2004; 25: 31-52.

Espallargues M, Valderas JM, Alsonso J. Provision of Feedback on Perceived Health Status to Health Care Professionals. *Medical Care***.** 38(8):877-878, August 2000; 175-183

Garratt AM, Hutchinson A, Russell IT. Patient-assessed measures of health outcome in asthma: a comparison of four approaches. *Respiratory Medicine* 2000; **94**:597-606.

Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ*. 2002;**324**:1417

Gilbody S, House A, Sheldon T, Routine administration of Health Related Quality of Life (HRQoL) and needs assessment instruments to improve psychological outcome – a systematic review *Psychological Medicine* 2002: **32**: 1345-1356

Greenhalgh J, Long AF, Brettle AJ, Grant MJ.Reviewing and selecting outcome measures for use in routine practice.*J Eval Clin Pract*. 1998 Nov;4(4):339-50.

Greenhalgh J, Long AF, Flynn R. The use of patient reported outcome measures in routine clinical practice: lack of impact or lack of theory? *Soc Sci Med*. 2005; **60:833-**43

Iglesias C, Torgeron D, Does length of questionnaire matter? A randomised trial of response rates to a mailed questionnaire *J Health Serv Res & Policy* 2000; 5: 219-21.

Jenkinson CP, Mant JW, Carter J, Wade DT, Winner S. The London Handicap Scale: a re-evaluation of its validity using standard scoring and simple summation. *Journal of Neurology, Neurosurgery and Psychiatry* 2000; **68**:365-7.

McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, Thomas R, Harvey E, Garratt A, Bond J. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess.* 2001;**5**(31):1-256

Mallinson S The Short-Form 36 and older people: some problems encountered when using postal administration. *J Epidem Comm Health* 1998; 52: 324-28.

McHorney CA, Bricker ED. A qualitative study of patients' and physicians' views about practice-based functional health assessment. *Medical Care*, 2002;40(11):1113-1125..

Nakesh R, Hutton J, Jorsted-Stein E, et al., Maximising response to postal questionnaires--a systematic review of randomised trials in health *research BMC Med Res Methodolog* 2006; 1471-2281.

Parker S, Bechinger-Engish D, Jagger C et al., Factors affecting completion of the SF-36 in older people. *Age Ageing* 2006; 35: 376-81.

Smeeth L, Fletcher A, Stirling S, et al., Randomised comparison of three methods of administering a screeing questionnaire to elderly people. *BMJ* 2001; 323: 1403-6.

Sitzia J and Wood N, Response rate in patient satisfaction research: an analysis of 210 published studies. *International Journal for Quality in Health Care* 1998; 10:311-17

Steinwachs DM, Wu AW, Skinner EA. How will outcomes management work? *Health Affairs* 1994.
Wasserman J, Aday LA, Begley CE, Ahn C, Lairson DR. Measuring health state preferences for hemophilia: development of a disease-specific utility instrument. *Haemophilia.* 2005;**11**:49-57

Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol*. 2003;56:52-60