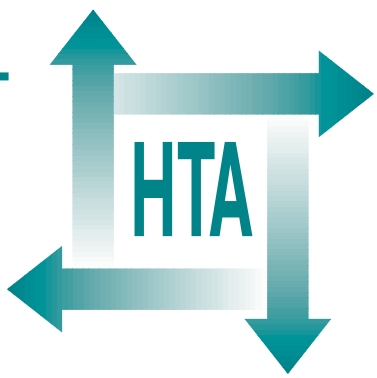
Health Technology Assessment 1998; Vol. 2: No. 14

Review

Evaluating patient-based outcome measures for use in clinical trials

Ray Fitzpatrick
Claire Davey
Martin J Buxton
David R Jones

Health Technology Assessment
NHS R&D HTA Programme



Standing Group on Health Technology

Chair: Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Dr Sheila Adam,
Department of Health

Professor Martin Buxton,
Professor of Economics, Brunel University †

Professor Angela Coulter,
Director, King's Fund, London

Professor Anthony Culyer,
Deputy Vice-Chancellor, University of York

Dr Peter Doyle,
Executive Director, Zeneca Ltd,
ACOST Committee on Medical Research & Health

Professor John Farndon,
Professor of Surgery, University of Bristol †

Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee †

Professor John Gabbay,
Director, Wessex Institute for Health Research & Development †

Professor Sir John Grimley Evans,
Department of Geriatric Medicine, Radcliffe Infirmary, Oxford †

Dr Tony Hope,
The Medical School, University of Oxford †

Professor Howard Glennester,
Professor of Social Science & Administration, London School of Economics & Political Science

Mr John H James,
Chief Executive, Kensington, Chelsea & Westminster Health Authority

Professor Richard Lilford,
Regional Director, R&D, West Midlands †

Professor Michael Maisey,
Professor of Radiological Sciences, UMDS, London

Dr Jeremy Metters,
Deputy Chief Medical Officer, Department of Health †

Mrs Gloria Oates,
Chief Executive, Oldham NHS Trust

Dr George Poste,
Chief Science & Technology Officer, SmithKline Beecham †

Professor Michael Rawlins,
Wolfson Unit of Clinical Pharmacology, University of Newcastle-upon-Tyne

Professor Martin Roland,
Professor of General Practice, University of Manchester

Mr Hugh Ross,
Chief Executive, The United Bristol Healthcare NHS Trust †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York

Professor Trevor Sheldon,
Director, NHS Centre for Reviews & Dissemination, University of York †

Professor Mike Smith,
Director, The Research School of Medicine, University of Leeds †

Dr Charles Swan,
Consultant Gastroenterologist, North Staffordshire Royal Infirmary

Dr John Tripp,
Department of Child Health, Royal Devon & Exeter Healthcare NHS Trust †

Professor Tom Walley,
Department of Pharmacological Therapeutics, University of Liverpool †

Dr Julie Woodin,
Chief Executive, Nottingham Health Authority †

† Current members

HTA Commissioning Board

Chair: Professor Charles Florey, Department of Epidemiology & Public Health,
Ninewells Hospital & Medical School, University of Dundee †

Professor Ian Russell,
Department of Health, Sciences & Clinical Evaluation, University of York *

Dr Doug Altman,
Director of ICRF/NHS Centre for Statistics in Medicine, Oxford †

Mr Peter Bower,
Independent Health Advisor, Newcastle-upon-Tyne †

Ms Christine Clark,
Honorary Research Pharmacist, Hope Hospital, Salford †

Professor David Cohen,
Professor of Health Economics, University of Glamorgan

Mr Barrie Dowdeswell,
Chief Executive, Royal Victoria Infirmary, Newcastle-upon-Tyne

Professor Martin Eccles,
Professor of Clinical Effectiveness, University of Newcastle-upon-Tyne †

Dr Mike Gill,
Director of Public Health and Health Policy, Brent & Harrow Health Authority †

Dr Jenny Hewison,
Senior Lecturer, Department of Psychology, University of Leeds †

Dr Michael Horlington,
Head of Corporate Licensing, Smith & Nephew Group Research Centre

Professor Sir Miles Irving
(Programme Director), Professor of Surgery, University of Manchester, Hope Hospital, Salford †

Professor Alison Kitson,
Director, Royal College of Nursing Institute †

Professor Martin Knapp,
Director, Personal Social Services Research Unit, London School of Economics & Political Science

Dr Donna Lamping,
Senior Lecturer, Department of Public Health, London School of Hygiene & Tropical Medicine †

Professor Theresa Marteau,
Director, Psychology & Genetics Research Group, UMDS, London

Professor Alan Maynard,
Professor of Economics, University of York †

Professor Sally McIntyre,
MRC Medical Sociology Unit, Glasgow

Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield †

Professor Gillian Parker,
Nuffield Professor of Community Care, University of Leicester †

Dr Tim Peters,
Reader in Medical Statistics, Department of Social Medicine, University of Bristol †

Professor David Sackett,
Centre for Evidence Based Medicine, Oxford

Professor Martin Severs,
Professor in Elderly Health Care, Portsmouth University †

Dr David Spiegelhalter,
MRC Biostatistics Unit, Institute of Public Health, Cambridge

Dr Ala Szczepura,
Director, Centre for Health Services Studies, University of Warwick †

Professor Graham Watt,
Department of General Practice, Woodside Health Centre, Glasgow †

Professor David Williams,
Department of Clinical Engineering, University of Liverpool

Dr Mark Williams,
Public Health Physician, Bristol

Dr Jeremy Wyatt,
Senior Fellow, Health and Public Policy, School of Public Policy, University College, London †

* Previous Chair
† Current members

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Evaluating patient-based outcome measures for use in clinical trials

Ray Fitzpatrick[1]
Claire Davey[1]
Martin J Buxton[2]
David R Jones[3]

[1] Division of Public Health and Primary Health Care, University of Oxford
[2] Health Economics Research Group, Brunel University
[3] Department of Epidemiology and Public Health, University of Leicester

Published October 1998

This report should be referenced as follows:

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

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel and funded as project number 93/47/09.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for the recommendations for policy contained herein. In particular, policy options in the area of screening will, in England, be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

Contents

List of abbreviations

EORTC	European Organisation for Research and Treatment of Cancer
FLIC	Functional Living Index-Cancer
FLP	Functional Limitations Profile
HAQ	Health Assessment Questionnaire
HUI	Health Utilities Index
MACTAR	McMaster–Toronto Arthritis Patient Preference Disability Questionnaire
MCID	minimal clinically important difference
NHP	Nottingham Health Profile
QALY	quality-adjusted life year
QoL	quality of life
QWB	Quality of Well-Being Scale
SEIQoL	Schedule for the Evaluation of Individual Quality of Life
SF-36	Short Form 36-item questionnaire
SIP	Sickness Impact Profile
SRM	standardised response mean

Executive summary

Background

'Patient-based outcome measure' is a short-hand term referring to the array of questionnaires, interview schedules and other related methods of assessing health, illness and benefits of health care interventions from the patient's perspective. Patient-based outcome measures, addressing constructs such as health-related quality of life, subjective health status, functional status, are increasingly used as primary or secondary end-points in clinical trials.

Objectives

- To describe the diversity and reasons for diversity of available patient-based outcome measures.
- To make clear that criteria investigators should have in mind when they select patient-based outcome measures for use in a clinical trial.

Methods

Data sources

Literature was identified by a combination of electronic searches of databases, handsearching of selected journals and retrieval of references cited in available literature. Databases used included MEDLINE, EMBASE, CINAHL, PsychLIT and Sociofile.

Study selection

A set of explicit criteria were used for selection of literature. Articles were included if they focused on any methodological aspect of patient-based outcome measures (for example, methods of evaluating such measures, psychometric evaluation of measures, comparative studies of measures, studies reporting validation of measures). Studies were excluded if they only reported use of a measure without evaluation, focused only on cross-cultural issues, focused only on clinician-based outcome measures or discussed economic utility theory only without considering measurement.

A total of 5621 abstracts and articles were identified by initial searches as potentially relevant. However, after assessment, 391 key references were selected as useful to the objectives of the review. A further 22 references were incorporated into the final version as a result of comments from external experts and referees.

Data synthesis

A first draft synthesising the evidence was produced by the first author of this review (RF) and extensively critiqued by the other three authors. A revised version was then submitted for evaluation to a panel of ten experts recruited to represent a wide range of areas of expertise (including clinical medicine, clinical trials, health economics, health services research, social sciences and statistics). Feedback from this panel was read and discussed by the authors of the review and a third version of the review drafted. The final version is a quasi-consensus view from individuals with a wide range of expertise.

Results

Diversity of patient-based outcome measures

- Seven major types of instrument can be identified in the literature: disease-specific, site-specific, dimension-specific, generic, summary item, individualised, utility.
- Concepts, definitions and theories of what such instruments measure are generally not clearly or consistently used. For example, there is little consistency of use or agreement as to the meaning of key terms such 'quality of life' and 'health-related quality of life'.
- The intended purpose and content of types of instruments vary. There are advantages and disadvantages to each of the different type of instrument when used in a particular clinical trial.

Criteria for selecting patient-based outcome measures

- There are eight criteria that investigators should apply to evaluate candidate patient-based outcome measures for any specific clinical trial: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability, feasibility.

- These criteria are not consistently defined and the literature associated with the criteria cannot be summarised in clear, explicit and unambiguous terms.
- It is not possible from available evidence to rank order the relative importance of the eight criteria in relation to decisions about selection of measures to include in a trial.
- Appropriateness requires that investigators consider the match of an instrument to the specific purpose and questions of a trial.
- Reliability requires that an instrument is reproducible and internally consistent.
- Validity is involved in judging whether an instrument measures what it purports to measure.
- Responsiveness in this context addresses whether an instrument is sensitive to changes of importance to patients.
- Precision is concerned with the number and accuracy of distinctions made by an instrument.
- Interpretability is concerned with how meaningful are the scores from an instrument.
- Acceptability addresses how acceptable is an instrument for respondents to complete.
- Feasibility is concerned with the extent of effort, burden and disruption to staff and clinical care arising from use of an instrument.

Conclusions and recommendations

- Investigators need to make their choice of patient-based outcome measures for trials in terms of the criteria identified in this review.
- Developers of instruments need to make evidence available under the same headings.
- By means of the above criteria, further primary research and consensus-type processes should be used to evaluate leading instruments in the different fields and specialties of health care to improve use of patient-based outcome measures in research. Primary research is needed either in the form of methodological additions to substantive clinical trials (for example comparing the performance of two or more measures) or studies of leading measures with methodology as the primary rationale.

Chapter 1

Purpose and plan of this review

For the purpose of this review, by patient-based outcome measures we mean questionnaires or related forms of assessment that patients complete by themselves or, when necessary, others on their behalf complete, in order that evidence is obtained of their experiences and concerns in relation to health status, health-related quality of life (QoL) and the results of treatments received. Although these measures have been developed for a number of other applications, this review is concerned with their use in clinical trials. There is now an enormous array of such measures that can be used in clinical trials. The purpose of this review is to make clear the criteria investigators should have in mind when they select patient-based outcome measures at the stage of designing a clinical trial.

The first purpose of the review is that the diversity and reasons for diversity of available instruments are made clear to the reader. Patient-based outcome measures have been developed to serve a variety of different functions, and it is therefore important to appreciate the range, types and intended uses of such instruments. These issues are the subject matter of chapter 2 of this review.

The second purpose of the review, covered in chapter 3, is explicitly to identify the criteria whereby instruments should be evaluated and selected for use in any given trial. We distinguish eight different criteria or considerations that are relevant to the such a selection. The reader is then provided with a summary of currently available evidence and thinking behind each of the eight criteria.

In appendix 1, an explanation can be found of how the relevant literature was identified and assessed, and how we approached this review of evidence. It is not the purpose of a review such as is reported here to find and synthesise the contents of every article written on patient-based outcome measures. The format of the review is more akin to a 'structured review' with as explicit a search strategy as is feasible combined with what seem inevitably qualitative methods of describing and summarising material. In many ways, such a review is more accurately described as a 'scoping' or 'mapping' exercise. The authors recognise that bias may be involved both in the search and selection procedures used to assemble evidence, and probably, more seriously and realistically, in how the diverse literature assembled was summarised, interpreted and reported. As discussed more extensively in the appendix describing the methods of the review (appendix 1), the most substantial check against such biases was the recruitment of a panel of experts as diverse in scientific interests and approach as possible to critique an earlier draft of the review. Every effort was made by the authors as a group to revise the review in the light of the expert panel's comments. A list of the references that were used to inform the review is provided.

The criteria that we have identified can most directly be expressed in terms of eight questions that investigators should have in mind when they are choosing a patient-based outcome measure for a trial. These questions are listed in *Box 1*. In our view, if investigators give explicit attention to each of these questions, they will make more appropriate choices of patient-based outcome measure for trials. Some of the questions are relatively simple and, in principle, evidence should be readily available to help investigators evaluate

BOX 1 Eight questions that need to be addressed in relation to a patient-based outcome measure being considered for a clinical trial (see chapter 3)

- Is the content of the instrument appropriate to the questions which the clinical trial is intended to address? (Appropriateness)
- Does the instrument produce results that are reproducible and internally consistent? (Reliability)
- Does the instrument measure what it claims to measure? (Validity)
- Does the instrument detect changes over time that matter to patients? (Responsiveness)
- How precise are the scores of the instrument? (Precision)
- How interpretable are the scores of the instrument? (Interpretability)
- Is the instrument acceptable to patients? (Acceptability)
- Is the instrument easy to administer and process? (Feasibility)

the merits of an instrument. For example, data on the response rate associated with a questionnaire, that is, the proportion of individuals who are asked to complete a questionnaire and actually do so, may be of direct relevance to judging the acceptability of a questionnaire and ought to be relatively easy to interpret. By contrast and as the literature review in chapter 3 will demonstrate, for some of the other criteria, there are much greater inherent ambiguities and much less consensus. Thus the criterion of validity is concerned with the beguilingly simple question of whether a questionnaire is truly assessing what it purports to assess. Although there is unanimity in the literature that this is a fundamental question with patient-based outcome measures, there is no agreement on how exactly validity should be assessed. The purpose of this review is therefore to draw together the different dimensions and approaches to validity so that, ultimately, investigators can be clearer and better informed when they decide whether an instrument does have validity for a particular question addressed in a trial.

It should be noted that the questions and criteria we have identified are not rank ordered in terms of importance. Nor is there any reason to think that they need to be approached in the order with which they have been presented here. Above all, in practice investigators may find they have to make trade-offs between criteria when faced with choices between instruments. For example, a questionnaire may ask such a large number of relevant questions of patients that it may appear to have considerable validity as a measure. However, its very detail and length may reduce its acceptability and feasibility. There is no evidence in the literature to assist researchers in assigning priority to the criteria we have discussed. The selection of a patient-based measure for a trial therefore remains to some extent a matter of judgement and as much an art as a science. It is our hope that the reader of this review will be clearer about the principles involved in such judgements.

Chapter 2

What are patient-based outcome measures?

The emergence of patient-based outcome measures

A number of trends in health care have resulted in the development and growing use of patient-based outcome measures to assess matters such as functional status and health-related QoL (Bergner, 1985; Ebbs *et al.*, 1989). It is increasingly recognised that traditional biomedically defined outcomes such as clinical and laboratory measures need to be complemented by measures that focus on the patient's concerns in order to evaluate interventions and identify more appropriate forms of health care (Slevin *et al.*, 1988). Interest in patient-based measures has been fuelled by the increased importance of chronic conditions, where the objectives of interventions are to arrest or reverse decline in function (Byrne, 1992). In the major areas of health service spending, particularly in areas such as cancer, cardiovascular, neurological and musculo–skeletal disease, interventions aim to alleviate symptoms and restore function, with major implications for QoL (de Haes and van Knippenberg, 1985; Fowlie and Berkeley, 1987; Devinsky, 1995). In many new and existing interventions, increased attention also has to be given to potentially iatrogenic effects of medical interventions in areas such as well-being and quality of life. Patient-based outcome measures provide a feasible and appropriate method for addressing the concerns of patients in the context of controlled clinical trials.

At the same time, increased attention is given to patients' preferences and wishes in relation to their health care (Till *et al.*, 1992). Patients increasingly expect with good reason to be involved in decisions about their care and to be given accurate information to facilitate their involvement (Siegrist and Junge, 1989). More evidence, and more relevant evidence, is therefore needed by patients about how illnesses and their treatments are likely to affect them.

Continuing difficulties experienced by all governments and health authorities in finding financial resources to meet demands on health care increase pressures for evidence to assess benefits in relation to costs of health care so that better use is made of resources. Evidence is needed of such benefits as perceived by patients, carers, health care professionals and by society as a whole (Epstein, 1990; Anonymous, 1991a; O'Boyle, 1995).

For all these reasons, much greater effort is now required to assess the impact upon the individual of illness and treatments by means of accurate and acceptable measures. An enormous array of instruments in the form of questionnaires, interview schedules, rating and assessment forms has emerged that have in common the objective of assessing states of health and illness from the patient's perspective. Because their purpose is to assess the impact of health care interventions from the view-point of the patient, this review refers collectively to such instruments as patient-based measures of outcome.

Accompanying the mounting interest in patient-based measures is an explosion of literature. One MEDLINE search on QoL revealed 1000 articles (Rosenberg, 1995) and another retrieved three times as many QoL papers in 1994 as in 1990 (Editorial, 1995). In part, this vast and rapidly expanding literature reflects a huge growth in the number of new questionnaires and other instruments to assess health status and related concepts. In response to these developments, a number of volumes have appeared which provide guides to the different types of instruments, their content and range of applications (Wilkin *et al.*, 1993; Bowling, 1995a; McDowell and Newell, 1996; Spilker, 1996; Bowling, 1997). These volumes provide an excellent resource for the investigator wishing to examine the range of instruments available in, for example, a particular condition such as cancer, or to assess a particular aspect of QoL such as social support. From such sources, the reader can review the range of instruments in any field and also the history of their development and use, to date.

There are also now available increasingly clear and informative guidelines about how to develop and report the development of patient-based outcome measures (Sprangers *et al.*, 1993; McDowell and Jenkinson, 1996) and how such measures should be used and reported in clinical trials (Staquet *et al.*, 1996; Fayers *et al.*, 1997).

This review is intended to be a resource with a somewhat different purpose. It aims to provide explicit guidance on how to select from the array of available instruments. It makes as explicit as possible the considerations relevant to choosing a patient-based outcome measure for use in research. It is primarily intended for use in the fields of clinical trials and related evaluation studies where a questionnaire assessing health status might be included as a measure of outcome. This distinctive focus is upon the assessment of changes in health in groups of patients that may be detected in clinical trials and may be due to the treatment under investigation. Later in this chapter, other applications of patient-based measures (in areas such as health needs assessment and screening) are briefly discussed, but a detailed consideration of other uses is beyond the scope of this review. A number of general discussions have already been published with the intention of helping the trialist to select and use patient-based outcome measures (Aaronson, 1989; Fitzpatrick *et al.,* 1992; Guyatt *et al.,* 1993b; Guyatt, 1995; European Research Group on Health Outcomes Measures, 1996; Testa and Simonson, 1996). Guidance on choosing an instrument has also been published in a number of more specialist fields including; rheumatology (Tugwell and Bombardier, 1982; Deyo, 1984; Bombardier and Tugwell, 1987; Bell *et al.,* 1990; Fitzpatrick, 1993; Bellamy *et al.,* 1995; Peloso *et al.,* 1995), cancer (Clark and Fallowfield, 1986; Maguire and Selby, 1989; Moinpour *et al.,* 1989; Skeel, 1989; Fallowfield, 1993; Selby, 1993), cardiovascular disease (Fletcher *et al.,* 1987), neurology (Hobart *et al.,* 1996), surgery (Bullinger, 1991), and in relation to particular applications such as, drug trials (Jaeschke *et al.,* 1992; Patrick, 1992) and rehabilitation (Wade, 1988). This review builds on and synthesises this body of literature. It is intended to make as explicit as possible the different properties that are expected of patient-based outcome measures. They are presented in terms of the criteria whereby we should judge instruments when selecting the most appropriate one for a particular trial. Where important differences of views exist in the published evidence on any point, the review attempts to reflect this diversity.

Concepts and definitions

This is a review of a field in which there is no precise definition or agreement about subject matter (McDaniel and Bach, 1994). We are concerned with questionnaires and related instruments that ask patients about their health. However with regard to more precise definitions of what such instruments are intended to assess, there is no agreed terminology and reviews variously refer to instruments as being concerned with 'QoL', 'health-related QoL', 'health status', 'functional status', 'performance status', 'subjective health status', 'disability', 'functional well-being'. To some extent, this diversity reflects real differences of emphasis between instruments. Some questionnaires focus exclusively upon physical function, for example, assessing mobility and activities of daily living without reference to social and psychological factors, and might appropriately be described as functional status instruments. Other instruments may ask simple global questions about the individual's health. Other instruments again are concerned with the impact of health on a broad spectrum of the individual's life, for example, family life and life satisfaction, and might reasonably be considered to assess QoL. In reality the various terms such as 'health status' and 'QoL' are used interchangeably to such an extent that they lack real descriptive value (Spitzer, 1987). It is unusual in the current literature for terms such as 'QoL' to be selected with any specific intent. The term 'patient-based outcome measure' is here used wherever possible as the most all-embracing term to encompass all of the types of instruments conveyed by other terms such as 'health status', or 'QoL'.

Some of the terms used to describe this field can actually be unhelpful. In particular, the frequently used phrase 'QoL' to describe instruments, misleadingly suggests an abstract or philosophical set of judgements or issues relating to life in the broadest sense of factors outside the person, such as living standards, political or physical environment. Because, rightly or wrongly, hardly any of the vast array of so-called QoL measures used in health settings address matters beyond the health-related (Meenan and Pincus, 1987), we avoid using this terminology as much as possible.

The common denominator of all instruments considered relevant to this review is that they address some aspect of the patient's subjective experience of health and the consequences of illness. Such instruments ask for patients to report views, feelings, experiences that necessarily are as perceived by the respondent (Mor and Guadagnoli, 1988). Respondents are asked about experiences such as satisfaction, difficulty, distress or symptom severity that are unavoidably subjective phenomena. It has to be accepted that such experiences cannot be objectively 'verified' (Albrecht, 1994). In some

cases questionnaire items may ask for reports of very specific behaviours, for example, ability to walk a certain distance, or use of physical aids, that observers such as carers or therapists can in principle readily verify from observation. Even with such behavioural items, the questionnaire still largely elicits perceptual information. It is this reporting of the personal and the subjective by the patient that uniquely identifies patient-based outcome measures from other health information used as outcomes, such as laboratory data. Clinical scores and scales are a different kind of subjective perceptual evidence; they are the perceptual judgement of doctors or of other health professionals. It is the inherently subjective source of patient-based material that leaves grounds for anxiety in some minds about the 'hardness', robustness and ultimately scientific value of such evidence (Fries, 1983; Deyo, 1991). Such concerns are addressed when we consider desirable measurement properties of patient-based measures in chapter 3.

Dimensions such as 'QoL' and 'subjective health status' can be assessed by what may be considered a continuum of methods. At one extreme, health professionals or others make judgements with minimal input from the patient, and, at the other extreme, assessments are largely determined by the patient with minimal influence from other observers. This review is largely concerned with instruments of the latter kind because there is a *prima facie* case that such measures more directly elicit the respondent's perspective rather than the observer's (O'Brien and Francis, 1988; Rothman *et al.,* 1991; Berkanovic *et al.,* 1995). However there is a continuum of approaches and much of what is discussed in this review may be relevant to assessments such as disability scales or standardised psychiatric assessments which are completed by observers on the basis of evidence from a patient, but without the patient himself or herself literally selecting the items or description that most fit their view.

There are circumstances where patients are unable to provide their unique report of their perceptions, due to ill-health, physical or cognitive problems, or some other incapacity. In these cases, proxy reports may be necessary because of the need for some assessment to inform a clinical trial. Because there is consistent evidence of lack of agreement with patients' judgements of their QoL by observers such as health professionals, informal carers, and other so-called 'proxy' judges, this is increasingly considered a second-best solution to be used only when the patient cannot contribute (Mosteller *et al.,* 1989; Clarke and Fries, 1992; Sprangers and Aaronson, 1992). However, there is also substantial evidence that patients with poorer health are less likely to complete patient-based outcome measures (Bleehen *et al.,* 1993). Since such patients are an important group in relation to assessment of outcomes in trials and their omission may result in bias, effort is required to examine the extent to which proxy ratings of outcome are valid. Whilst there is clear evidence of discrepancies in judgements between patients' and proxy reports from others, it is important to examine closely the scope for obtaining proxy reports when patients' are unable to contribute. Sneeuw and colleagues (1997) used a simple QoL instrument with relatively few distinctions between levels of QoL with patients with a range of cancer diagnoses, and their informal carers and physicians also rated the sample of patients on the same instrument. For five out of six dimensions, there was broad agreement between patient and physician or carer in 85% of patients, and 75% agreement on a sixth dimension ('social activities'). Agreement also increased for some dimensions at a follow-up assessment. Thus there is some support for using evidence from proxies at least when relatively simple judgements are required.

Theories and concepts

It is sometimes argued that this field lacks a rigorous underpinning theory and clear and precise definitions that flow from theory (Schipper and Clinch, 1988; Ventegodt, 1996). There is some basis for this criticism; much of the work stems from very applied and pragmatic problem solving, rather than deriving from an explicit theoretical framework. However it is not entirely true that the field of patient-based outcome measures lacks theories of the phenomena that investigators wish to measure. Psychometric theory provides a well established foundation for most patient-based outcome measures (Nunnally and Bernstein, 1994). This scientifically rigorous field is concerned with the science of assessing the measurement characteristics of scales and involves such properties as validity, reliability and responsiveness (Hays *et al.,* 1993). More recently developed is the field of 'clinimetrics' (Feinstein, 1987; Feinstein, 1992; Wright and Feinstein, 1992). Closely associated with psychometric theory, this field focuses on the clinical challenges of constructing scales that clinicians use for measuring health status of patients (Fava, 1990). Similarly, economic contributions to this field have also a broad range of theoretical literature on which to draw from evidence such as decision-theory (Torrance, 1986).

Thus strictly speaking, a far greater difficulty than the dearth of theory is that there are a large

number of such discussions of the theoretical basis of, say, QoL (Rosenberg, 1995; Rogerson, 1995). These theories also generate definitions, each with distinctive emphases. This can be seen from an illustrative list of definitions and discussions of health and health-related QoL that have been cited as useful in this field *(Box 2)*. The literature is replete with such definitions accompanied by theoretical justification. None has commanded greater attention than others.

BOX 2 Illustrations of range of definitions and discussions of health and QoL

- Health as a 'state of complete physical, mental and social well-being and not merely the absence of disease or infirmity.' (WHO, 1947)
- 'Quality of life is an individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns.' (WHOQOL Group, 1993)
- 'Quality of life refers to patients' appraisal of and satisfaction with their current level of functioning as compared to what they perceive to be ideal.' (Cella and Tulsky, 1990)
- 'Health-related quality of life is the value assigned to duration of life as modified by the impairment, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy.' (Patrick and Erickson, 1993a)
- 'Health-related quality of life refers to the level of well-being and satisfaction associated with an individual's life and how this is affected by disease, accidents and treatments from the patient's point of view.' (Lovatt, 1992)
- 'Quality of life is enhanced when the distance between the individual's attained and desired goals is less.' (Bergner, 1989)
- 'Quality of life measures the difference, or the gap, at a particular period of time, between the hopes and expectations of the individual and that individual's experiences (Calman, 1984)

There is therefore an enormous array of concepts and definitions. Farquhar (1994, 1995) reviewed the range of definitions of QoL in the field of health and developed a typology. She distinguished 'global definitions' which express QoL in general terms such as degree of satisfaction with life, 'component definitions' that break down QoL into specific parts or dimensions, such as health, life satisfaction and psychological well-being; and 'focused definitions' that emphasise only one or two of the range of possible component parts of life.

Schipper and colleagues (1996) assess the array of different perspectives that inform definitions of the term QoL in medical research and distinguish five different concepts or emphases *(Table 1)*. They suggest that the following simple definition captures much that is important across the five different perspectives:

> '*"Quality of life" in clinical medicine represents the functional effect of an illness and its consequent therapy upon a patient, as perceived by the patient' (Schipper* et al., *1996:16).*

Such a definition makes a very important point very simply with its emphasis upon the perception of the patient. In view of the competing array of such definitions, it would not be productive to attempt to devise a more convincing or more authoritative version. The result of any such exercise would add another competing definition to the abundance of already existing attempts. In any case, it is our view that very substantial progress may be made in the assessment of patient-based outcome measures without imposing a (somewhat

TABLE 1 *Alternative perspectives underlying competing definitions of QoL in health care*

Perspective	Illustration
The psychological view	The patient's perceptions of the impact of disease; for example, how symptoms are experienced and labelled
The utility view	The values attached to health states; the trade-offs individuals make between survival against QoL
The community centered view	The extent to which illness impacts on the individual's relations to a community in terms of employment, home making etc.
Reintegration into normal life	The extent to which, following illness, the individual can resume normal life in terms of self care, social activities etc.
The gap between expectations and achievements	The more the patient is able to realise his or her expectations, the higher the QoL

Adapted from Schipper et al. *(1996)*

arbitrary) theoretical stance in relation to such work.

Although no single definition or theory can plausibly be promoted as clearly more useful than others, it has been argued that the WHO's classification of Impairments, Disabilities and Handicaps (WHO, 1980) provides the most coherent and comprehensive framework for considering the consequences of health and disease (Wade, 1992; Ebrahim, 1995). Impairment refers to any loss or abnormality of psychological, physiological or anatomical function. Disability is any restriction or lack of ability to perform an activity in ways considered normal for an individual. Handicap is resulting from impairment or disability that limits the fulfilment of a role that is normal for that individual. Whilst there is no simple or straightforward mapping of the typical content of patient-based outcome measures onto this schema, as will be seen from the next section, items of most measures correspond to one or other of the headings of the WHO model.

Content of instruments

The content of the instruments with which we are concerned varies enormously, and in general a researcher will usually be able to find an instrument with questionnaire items that at least approximate to the issues of concern to his or her research question (Jenkinson *et al.,* 1996). Every instrument attempts to provide an assessment of at least one dimension of health status, either the respondent's global assessment of health or more specific dimensions such as mobility, pain or psychological well-being.

In reality, it is increasingly the case that instruments provide assessments of several dimensions of health status (Bice, 1976; Hall *et al.,* 1989; Jenkins, 1992; Hughes *et al.,* 1995). There are a large number of attempts to enumerate the full range of dimensions potentially implicated in constructs such as health status and health-related QoL. Lists of dimensions have been drawn up by investigators in at least three different ways. Some discussions of health-related QoL have drawn on consensus conferences to identify dimensions (Bergner, 1989). A second approach is to identify dimensions of health-related QoL by means of content analysis of the subscales of existing measures (van Knippenberg and de Haes, 1988; McColl *et al.,* 1995). The third approach is with minimal prompting to elicit from patients or members of the general public their views of the dimensionality of QoL (Sutherland *et al.,* 1990; Farquhar, 1994; Bowling, 1995b). Finally, statistical methods such as factor analysis have been used to identify the dimensionality of concepts such health status and QoL (Segovia *et al.,* 1989).

An attempt has been made in *Table 2* to draw together the dimensions of health status most commonly identified in the literature as relevant to patient-based outcome measures. It is apparent that the range is substantial. This increases the complexity of the choice faced by the individual who wishes to select an instrument for a clinical trial (Spilker, 1992). Dimensions range from those which are most obviously related to a patient's health status such as the patient's global view of their health, experiences of symptoms or

TABLE 2 *Range of dimensions assessed by patient-based outcome measures*

I Physical function	
Mobility, dexterity, range of movement, physical activity	
Activities of daily living: ability to eat, wash dress	
II Symptoms	
Pain	Energy, vitality, fatigue
Nausea	Sleep and rest
Appetite	
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	Memory
Alertness	Confusion
Concentration	Ability to communicate
VII Role activities	
Employment	Financial concerns
Household management	
VIII Personal constructs	
Satisfaction with bodily appearance	
Stigma and stigmatising conditions	
Life satisfaction	
Spirituality	
IX Satisfaction with care	

psychological illness through to dimensions that increasingly reflect the broader impact of illness on the individual's life such as social function, role activities and impact on paid income. Some dimensions such as spirituality may seem rather too ill-defined, subjective or remote from health care but may be important when, for example, judging the outcomes of palliative care (Joyce, 1994). Some dimensions have still received very little attention, for example, the sense of embarrassment or stigma that may be associated with many health problems.

Types of instruments

One of the main decisions to be made in selecting an instrument for a clinical trial is to choose among the different kinds of instrument that exist. The different major types of instrument are identified with examples in *Box 3*. They differ in content and also in the primary intended purpose. Whilst the distinction between types is a useful means of considering the range of options in patient-based outcome measures, the classification should not be interpreted too rigidly. Some instruments have elements of more than one category or evolve over time in their intended uses.

BOX 3 Different types of instruments and examples

- *Disease-specific:* the Asthma Quality of Life Questionnaire, the Arthritis Impact Measurement Scales
- *Site or region-specific:* the Oxford Hip Score, the Shoulder Disability Questionnaire
- *Dimension-specific:* Beck Depression Inventory, McGill Pain Questionnaire
- *Generic:* SF-36, FLP
- *Summary items:* question about limiting long-standing illness in the General Household Survey
- *Individualised:* MACTAR, SEIQoL
- *Utility:* EuroQoL EQ-5D, Health Utility Index (HUI)

In this section, we consider briefly the advantages and disadvantages claimed for different types of instruments. It should be emphasised that to a large extent these are postulated rather than firmly established advantages and disadvantages. Generalisations about advantages and disadvantages of broad types of instrument are difficult to substantiate because too little evidence is available particularly from direct comparisons of their use.

Disease/condition-specific

As the title implies, these instruments have been developed in order to provide the patient's perception of a specific disease or health problem. An example of such a questionnaire is the Asthma Quality of Life Questionnaire (Juniper *et al.,* 1994). It contains 32 questions assessing four dimensions (activity limitations, symptoms, emotional function and exposure to environmental stimuli). Another example is the Arthritis Impact Measurement Scale, a self administered questionnaire for use in rheumatic diseases (Meenan *et al.,* 1980; Meenan, 1982). There are 45 questionnaire items covering nine dimensions: dexterity, physical activity, mobility, household activities, activities of daily living, depression, anxiety, pain and social activities. Both instruments clearly are intended to have a quite specific range of applications in terms of disease.

A distinctive approach has been developed in the area of cancer. The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Study Group has developed and tested a cancer-specific questionnaire, the EORTC QLQ-C30, which has 30 items assessing five aspects of function, global QoL, and various areas of symptoms for use with patients with any form of cancer (Aaronson *et al.,* 1993). To this core instrument may be added one of the supplementary questionnaires that they have also developed, to provide more specific assessments of for example breast cancer (Sprangers *et al.,* 1996) or head and neck cancer (Bjordal *et al.,* 1994). This 'modular' approach provides a core instrument with which comparisons across cancer groups may be made together with more specific instruments intended to be particularly relevant to a more specific group.

Advantages and disadvantages

Several advantages are claimed for disease specific measures. Firstly, they are intended to have very relevant content when used in trials for a specific disease. All of the items in the instrument should have been developed specifically to assess the particular health problem being studied in the trial. A related but distinct advantage is claimed, namely that disease-specific instruments are more likely to detect important changes that occur over time in the particular disease studied (Patrick and Deyo, 1989). An arthritis-specific instrument should be particularly sensitive to important changes in patients with arthritis because it should contain few if any irrelevant items. It might also be argued that the acceptability to patients and therefore completion rates should be high as the instrument has clear relevance to the patient's presenting problem.

The most salient potential disadvantage is that it is generally not possible to administer disease-specific instruments on samples who do not have

the relevant condition. This is a problem when investigators want data from a general sample of well individuals with which to compare health status scores of a study sample. This is a common procedure to provide some form of standard comparison with which to gauge the health of the study sample. In the most obvious sense, it is not possible to ask individuals about the experience of various problems arising from a condition that they do not have. A related disadvantage is that disease-specific instruments do not allow any obvious or easy comparison to be made between outcomes of different treatments for patients with different health problems. This is only a problem when some comparative judgement is required of effectiveness of different treatments for different diseases for purposes such as resource allocation (Cairns, 1996). Finally disease-specific instruments may not capture health problems associated with a disease and its treatment that have not been anticipated. An instrument with a broader range of items may be more likely to detect such unexpected effects (Read, 1993).

Site-specific

In some areas of medicine and surgery, instruments assessing the impact on the individual of a disease have come to be considered too broad in their coverage. Instruments have therefore been developed that assess health problems in a more specific part of the body. The Oxford Hip Score is a 12-item questionnaire designed to be completed by patients having total hip replacement surgery (Dawson *et al.,* 1996a). The items are summed to produce a single score of level of difficulties arising from the diseased hip. The Shoulder Disability Questionnaire is a 22-item questionnaire to assess degree of disability arising from shoulder symptoms (Croft *et al.,* 1994).

Advantages and disadvantages

The primary intended advantage is that the site-specific instrument should contain items that are particularly relevant to patient groups experiencing treatment for a very specific region of the body. They should also be particularly sensitive in trials of interventions to changes experienced by patients in that region. For example, a number of hip scores have been produced because of the need for outcome measures in orthopaedic surgery. Differences in outcome between different arms of a trial of total hip replacement surgery are quite difficult to detect and questions about pain due to osteo-arthritis in general may fail to detect specific problems in the one part of the body of concern in the evaluation (Dawson *et al.,* 1996a).

The principle disadvantage is the consequence of the relatively narrow focus of such instruments, namely that such instruments are unlikely to detect any changes in broader aspects of health or overall QoL. They are unlikely to be of value in detecting, for example, unexpected side-effects of an intervention in a trial.

Dimension specific

Dimension-specific instruments assess one specific aspect of health status. By far the most common type of dimension-specific measure is one that assesses aspects of psychological well-being. An example is the Beck Depression Inventory (Beck *et al.,* 1961). It contains 21 items that address symptoms of depression. The scores for items are summed to produce a total score. It was largely developed for use in psychiatric patients but is increasingly used more widely to assess depression as an outcome in physically ill populations. Another commonly assessed dimension of outcome in trials of physically ill patients is pain (Cleeland, 1990). The McGill Pain Questionnaire is an example of a dimension specific instrument developed for this use in this area (Melzack, 1975). It has several different versions, but the core of the instrument is formed by a series of lists of adjectives to describe pain, from each of which lists the patient selects adjectives that best describe his or her pain. Individual adjectives are ranked in terms of severity on the basis of prior research with patients treated for pain, and the items chosen by patients are summed to produce scores for three aspects of pain experience.

Advantages and disadvantages

The principal advantage of such instruments is that they provide a more detailed assessment in the area of concern, for example pain or psychological well-being, than is normally possible with the short scales usually used in disease-specific or generic instruments. Many of the instruments have been widely used in a range of clinical populations so that there is a wide range of comparative data with which to compare results (Wiklund and Karlberg, 1991). They are appropriate to medical as well as psychiatric conditions, although some instruments assessing psychological well-being need to be slightly modified either in content or scoring; items asking about physical problems but intended to assess somatic aspects of psychological distress may actually reflect underlying physical disease (Pincus *et al.,* 1986). This range of instruments is clearly of particular importance where psychological well-being is a key concern in a trial. Many of the other kinds of instruments we are

considering either omit this dimension or include only superficial assessments.

A potential problem is that assessments of psychological well-being in particular were often developed more to measure inter-patient differences for purposes of diagnosis or needs assessment than as outcome measures. Evidence for their appropriateness as an outcome measure requiring sensitivity to changes over time therefore needs to be examined carefully. Clearly in the context of a trial, obtaining a more detailed assessment of one dimension such as depression or pain must involve some reduction of depth on other dimensions if the overall burden of data collection on patients is not to be too great, so careful thought is required about the significance of the proposed specific dimension.

Generic instruments

Generic instruments are intended to capture a very broad range of aspects of health status and the consequences of illness and therefore to be relevant to a wide range of patient groups. The content of such questionnaires has been deliberately designed to be widely appropriate. They may provide assessments of the health status of samples of individuals not recruited because they have a specific disease, for example, from primary care or the community as a whole. One of the most widely used of such instruments is the SF-36 (Ware and Sherbourne, 1992). It is a 36-item questionnaire which measures health status in eight dimensions: physical functioning, role limitations due to physical problems, role limitations due to emotional problems, social functioning, mental health, energy/vitality, pain and general perceptions of health. An additional single item asks the respondent about health change over 1 year which is not scaled. Item responses are summed to give a score for each dimension. Evidence has also been presented that SF-36 can be used in several other forms. The items have been summed into two summary measures: the physical component summary and mental component summary (Ware *et al.,* 1995). The SF-36 has also been further reduced into a 12-item version (Ware *et al.,* 1996b).

A more lengthy generic instrument is the Functional Limitations Profile (FLP) which is the English version of Sickness Impact Profile (SIP) developed in the United States (Patrick and Peach, 1989). The FLP measures sickness related behavioural dysfunction by assessing individual perceptions of the effect of illness upon usual daily activities. It consists of 136 items grouped into 12 dimensions covering ambulation, bodycare and movement, mobility, household management, recreation and pastimes, social interaction, emotion, alertness, sleep and rest, eating, communication and work. Unlike the SF-36, the FLP uses weights expressing the severity of individual items that have been derived from prior research. The FLP has a summary score for physical and psychosocial dimensions and a total score can also be calculated.

Advantages and disadvantages

The main advantage of generic instruments is that they can in principle be used for a broad range of health problems. This means that they may be of use if no disease-specific instrument exists in a particular area (Visser *et al.,* 1994), although this is increasingly unlikely to be the case. Because it can be widely used, it enables comparisons across treatments for groups of patients with different groups, to assess comparative effectiveness. Because of their broad range of content and more general applicability, such instruments have been used more frequently than disease-specific instruments to assess the health of non-hospital samples in the general population. This has led to the use of such data to generate 'normative values', that is scores in the well with which patients with health problems can be compared. Because generic instruments are intended to be broad in scope, they may have value in detecting unexpected positive or negative effects of an intervention, whereas disease-specific instruments focus on known and anticipated consequences (Cox *et al.,* 1992; Fletcher *et al.,* 1992). Another potential advantage is that by covering a wide range of dimensions in a relatively economic format, they reduce the patient burden entailed by using a number of questionnaires. A less tangible advantage to any individual user is that, if trials generally converged on the use of a small number of generic instruments, a more general body of experience and comparative evidence could emerge to enhance the value and interpretability of patient-based outcome measures (Hadorn *et al.,* 1995; Ware, 1995).

Against these postulated advantages have to be weighed some potential disadvantages. In particular, it may be argued that by including items across a broad range of aspects of health status, generic instruments must sacrifice some level of detail in terms of relevance to any one illness. The risk is therefore of some loss of relevance of questionnaire items when applied in any specific context. A particularly important potential consequence for clinical trials is that generic instruments would have fewer relevant items to the particular disease and intervention and therefore be less sensitive to changes that might occur as a result of an intervention.

Summary items

Single questionnaire items have an important role in health care research. They invite respondents to summarise diverse aspects of their health status by means of one or a very small number of questions. The General Household Survey has, in annual surveys since 1974, used two questions that together provide an assessment of chronic illness and disability: 'Do you have any long-standing illness or disability?' and 'Does this illness or disability limit your activities in any way?' A positive answer to the two questions provide an indication of chronic disability.

An even simpler summary item is the question used in the Health Survey for England: 'How is your health in general? Would you say it was 'very good', 'good', 'fair', 'bad', 'very bad'?'

The item 'How would you rate your general feelings of well-being today' with answers indicated on a single visual analogue scale has been advocated for use in cancer (Gough *et al.,* 1983). Transition items are another form of summary health item, in this case asking the respondent to assess the state of their health currently compared with a specific point in the past such as their last clinic visit. Thus a transition item for use in arthritis asks patients: 'Thinking of any overall effects your arthritis may have on you, how would you describe yourself compared to the last time I interviewed you in (*month*)?' 'Do you feel you are 'much better', 'slightly better', 'the same', 'slightly worse' or 'much worse'?' (Fitzpatrick *et al.,* 1993b).

Advantages and disadvantages

The most obvious advantage of all such items is their brevity. They make the least demands on respondents' time. In the case of summary health items, some have also been used widely and for a long time on large samples of the general population so that there is a considerable range of potential comparable evidence. Despite their obvious simplicity, there is evidence of summary item validity; negative answers to such single items are given by individuals with poorer health (Leavey and Wilkin, 1988; Anderson *et al.,* 1990). The single item visual analogue scale for use in cancer was validated by showing cross-sectional agreement with more specific and established QoL scores (Quality of Life Index) in patients with advanced cancer (Gough *et al.,* 1983). Equally, there is evidence of the predictive value of single items; individuals providing negative answers are more likely to have poorer health in the future (Mossey and Shapiro, 1982). Idler and Angel (1990) found that, amongst middle-aged men, self rated health status was predictive of mortality over 12 years, after controlling for its association with medical diagnoses, demographic variables and health behaviour. There is also very favourable evidence of the reproducibility of self-rated health (Lundberg and Manderbacka, 1996). With regard to sensitivity to change, a visual analogue item has been used in a series of randomised controlled trials of various treatments for breast cancer and been shown to be a responsive measure of well-being (Hurny *et al.,* 1996). Similarly, transition items have been shown to have good validity by producing scores consistent with independent evidence of the direction of change in health experienced by respondents between separate assessments (MacKenzie *et al.,* 1986a; Fitzpatrick *et al.,* 1993b; Garratt *et al.,* 1994a). In these studies, patients reporting deterioration or improvement in relation to a baseline assessment show corresponding patterns of change in other baseline and current data on their health (Deyo and Diehl, 1986). As pointed out below amongst the disadvantages, summary items cannot reveal contradictory trends in different dimensions of health, for example an improvement in physical function that coincides with deteriorating psychological well-being. However by inviting the respondent to summarise their health, they do offer a potential method for weighing up the significance of such contradictory trends. How a sample views a gain in, say, mobility and reduced pain from an anti-arthritic drug, if it is at a cost in terms of nausea or some other side-effect, may be best assessed by their global judgements of overall health change.

Disadvantages of summary and transition items mainly relate to their brevity. Respondents are invited to make a summary judgement of dimensions of health and it is usually not possible to make more specific inferences about particular aspects of their health from these answers. The numbers of distinctions made by the response categories of simple summary items are few ('excellent', 'good' etc.) and these may be too crude to detect subtle but important changes observed by more detailed assessment (Jenkinson *et al.,* 1994). Because of the inevitably general nature of such questions, they may be considered particularly prone to the influence of expectations, transient aspects of mood, and variations between respondents in criteria for answering such questions (Krause and Jay, 1994). In the context of a trial, investigators are often interested in opposing trends in different dimensions of health; for example, improvements in physical health at the expense of mood. Summary or transitional items by themselves do not permit the detection of such trends. There is also evidence

that individuals completing summary transition items may recall poorer health states than actually experienced so that degree of improvement is exaggerated (Mancuso and Charlson, 1995). Respondents may also be unduly influenced by their current health state when asked to compare current with past health (Bayley *et al.*, 1995).

Individualized measures

Individualized measures are instruments in which the respondent is allowed to select issues, domains or concerns that are of personal concern that are not predetermined by the investigator's list of questionnaire items. By a variety of means, the respondent is encouraged to identify those aspects of life that are personally affected by health, without imposing any standardised list of potential answers (Ruta and Garratt, 1994). Individualized measures are still in their infancy but have attracted interest precisely because they appear to offer considerable scope for eliciting respondents' own concerns and perceptions. One example is the Schedule for the Evaluation of Individual Quality of Life (SEIQoL) (O'Boyle *et al.*, 1992). It is completed in three phases by semi-structured interview in order to produce an overall QoL score for sick or healthy people. The first stage asks the individual, with structured interviewer prompting when necessary, to list five areas of life most important to their QoL. Secondly, each of the five nominated areas is rated on a visual analogue scale from 'as good as it could be' to 'as bad as it could be'. The individual patient also rates overall QoL. The last stage uses 30 hypothetical case vignettes which vary systematically in terms of the properties respondents have already identified as important to them. Judgement analysis of respondents' ratings of these vignettes allows the investigator to produce weights for the five chosen aspects of life and an index score is calculated between 0 and 100. This exercise can then be repeated at subsequent assessments. A shorter method of deriving weights has recently been published (Hickey *et al.*, 1996). The SEIQoL is intended to be used rather like generic measures for the widest possible range of health problems.

A simpler example of an Individualized instrument is the McMaster–Toronto Arthritis Patient Preference Disability Questionnaire (MACTAR), primarily intended for use in arthritis (Tugwell *et al.*, 1987). Individuals with arthritis are asked to identify without prompting up to five activities adversely affected by their disease. They then rank order their selected areas in terms of priority. Assessment of change over time is simpler than with SEIQoL because individuals rate degree of change in nominated areas by transition questions or simple visual analogue scales. The MACTAR has been successfully incorporated into a randomised controlled trial of methotrexate for rheumatoid arthritis, in which it proved at least as sensitive to important changes as other conventional clinical measures included in the trial (Tugwell *et al.*, 1990, 1991).

Advantages and disadvantages

The main advantage claimed for individualised measures is that they particularly address individuals' own concerns rather than impose standard questions that may be less relevant. In this sense, they may have a strong claim for validity in terms of the content of items addressed by the instrument.

The principal disadvantage is that because respondents' concerns are addressed in some depth, the interview that is involved has to be personally administered, most likely by well trained personnel. This necessitates greater resources than are required by self-completed questionnaires. There is a greater time commitment for both investigators and respondents. Overall, the greatest potential disadvantage is therefore in terms of lower practical feasibility than simpler self-completed instruments. It is less easy to produce population-based comparative or normative data for such instruments although it has been possible to produce some comparative evidence of judgements made by relatively healthy individuals with SEIQoL (O'Boyle *et al.*, 1992).

Utility measures

This review follows the approach of some previous overviews in considering utility measures as a distinct type of measure contrasting with those already described, such as generic and disease-specific measures (Sutherland *et al.*, 1990; Zwinderman, 1990; Chalmers *et al.*, 1992). However another view is that utility measures are not a distinct class of measure but should be considered as a generic health status measure with one particular form of numerical weighting or valuation of health states (Torrance, 1986). Because important and distinctive properties are claimed for approaches based on preferences or utilities as weights, compared to all previous approaches considered in this review, detailed attention is given to this approach in this review.

Utility measures have been developed from economics and decision theory in order to provide an estimate of individual patients' overall preferences for different health states (Drummond, 1993; Bakker and van der Linden,

1995). This form of measure may therefore be described as using preference-based methods in contrast to non-preference approaches, which would describe many of the other types of instrument we have already reviewed (Gold *et al.,* 1996). The former is concerned as far as possible to obtain the respondent's own overall value of the different dimensions of his or her health status whereas the latter, as has already been described, mostly derives scores for dimensions of health status based on summing responses to questionnaire items, with the possibility of dimension scores being in turn summed.

Utility measures therefore elicit the personal preferences of individuals regarding health states. This kind of measure has also been regarded as a means of obtaining important evidence of the overall value to society of health states. Data from utility measures are applied to assess in turn the social value of health care interventions by means of cost–utility analysis (Patrick, 1976). Data regarding costs and utilities for different health care interventions have been used to inform decisions about resource allocation between competing interventions (Gold *et al.,* 1996). Whilst most attention has been given to utility measures because of their role in cost–utility analyses to inform decisions about resource allocation, there is some research on their use as decision-aids in individual patient care where patients face difficult choices between treatment options (McNeil *et al.,* 1982).

In the context of a clinical trial, there are two basically different methods of assessing the preferences or utilities of the patients involved. The most direct way of assessing patients' utilities associated with health states is for them to be elicited directly from patients who are in the health states of interest by means of an interview in which respondents take part in experimental tasks such as standard gamble or time trade-off to elicit their values and preferences (Read *et al.,* 1984; Torrance, 1986, 1987; Drummond, 1987). In simplistic terms, the experimental method employed with standard gamble elicits respondents' values regarding health states by finding out how ready an individual would be hypothetically to undergo varying levels of risk associated with treatment to avoid a given health state. The greater the level of risk acceptable to the individual, the more severe the health state. The analogous experimental task with time trade-off is for subjects to judge the equivalence of periods of time in a particular health state with varying shorter periods in perfect health. The shorter the period of perfect health considered equivalent, the more severe the health state.

The use of experimental tasks such as standard gamble or time trade-off may be considered forms of direct utility measurement in that patients in a trial directly report their own values through responses to experimental tasks in an interview. Alternatively, utilities may be assessed by obtaining information from the patients in a trial by means of self-completed questionnaires that assess health status more or less in the same way as other patient-based outcome measures already reviewed. That is, patients select items that most describe their health state. However, in this second approach, questionnaire items have weighted utility scores attached that have been derived from prior survey data in which utilities have been measured from, as far as possible, appropriate samples of respondents (Feeny *et al.,* 1995; Brooks *et al.,* 1996). This second approach may be considered indirect utility measurement in that whilst patients directly report their health states, utility values attached to these states are derived from prior research on the preferences of other samples. A variant of indirect utility measurement is to elicit values of a specific patient group, say patients with arthritis, that can then be used in other clinical trials of patients with arthritis from whom it may not always be feasible to perform full interviews.

It should be emphasised that the utilities approach to patient-based outcome measures (whether considered as a type of measure or as one form of weighting the scores of measures) is distinctive in the extent to which it draws on specific theoretical assumptions. In particular the concept of utility itself is central to utility measures. It is fundamental to economic theory but, partly because of its axiomatic status, it is hard to define (van Praag, 1993). Richardson (1994) refers to four different concepts or uses employed by the literature when referring to 'utilities'. In one sense, it has been used to refer to a psychological concept of well-being; measurable levels of satisfaction and desirability of individuals in relation to matters such as health. A second usage refers to utility as the ordinal ranking individuals have about options such as health states. Thirdly, utility may be used to refer to the intensity of preferences regarding options. The fourth sense of the term refers to preferences between options under conditions of risk. These important differences of emphasis remain in the current literature and cannot be resolved from research evidence. A somewhat simplistic approximation to the concept in the field of health states that 'the more 'utility' an individual expects to obtain from a particular good or service the more he will be willing to pay for it' (Hurst and Mooney, 1983). However, willingness to pay is only a

behavioural indicator of a more fundamental concept of personal well-being, pleasure, desire fulfilment and preferences (Weymark, 1991).

Most importantly, in the context of health, measures of utility have been pursued that provide a single figure or estimate of the overall value, quality, outcome or benefit obtained from a treatment. Utility measures are based on the assessment of health but attempt to summarise the value of such states. The significance to this approach of a single figure is two-fold. Firstly, a single index particularly and most directly elicits the individual's overall preference for a health state. Secondly, this global preference provides a simpler figure for analyses of the net benefit in health from an intervention, compared with the many outcomes produced by multi-dimensional measures more characteristic of most other health status measures.

The methodological considerations involved in measuring health state preferences in the context of experimental interviews are beyond the scope of this review. They have been usefully summarised by Froberg and Kane (1989a,b,c,d) and in other reviews (Llewellyn-Thomas *et al.,* 1982, 1984; Sutherland *et al.,* 1983).

The most familiar example in Europe of an indirect measure of utilities that can be simply administered to patients in the form of a self-completed questionnaire is the EuroQol EQ-5D (Rosser and Sintonen, 1993; Kind *et al.,* 1994; Brooks *et al.,* 1996). The part of the EQ-5D questionnaire to elicit health status comprises five questions each of which has three alternative response categories. The five items assess mobility, self care, usual activity, pain/discomfort and anxiety/depression. These items can be used by themselves as descriptions of respondents' health states. Responses are also scored by means of weights obtained from the valuations that other samples from the general population have assigned to health states using visual analogue scales.

Another indirect utility instrument is the Health Utilities Index (HUI), which has been developed to assess preferences via eight health status attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, self care and pain (Torrance *et al.,* 1995). There are three versions. The HUI-III is available for administration by both self complete questionnaire and interview. Scores for these attributes have been elicited by both visual analogue and standard gamble methods, although the weights for items have not yet been published. To date, it has not been extensively used in the UK.

With utility-based measures, one very basic choice has to be addressed by investigators that does not exist for other forms of patient-based outcome measures discussed in this review. The investigator must decide whose values are primarily to be reflected in the assessment of outcomes of a trial, a choice which in turn requires a judgement about the decisions a trial is intended to inform (Torrance, 1973; Gold *et al.,* 1996). The choice can be put simply as being between the values of patients themselves (which would suggest the need for direct utility measures, and the values of society as a whole (which would suggest the appropriateness of indirect measures such as EQ-5D which reflect broader population values). A secondary choice also arises as, if patients' preferences are the focus, either the preferences of participants in the current trial may be directly elicited or values of patients with the same health problem can be used. It is beyond the scope of this review to consider this choice in detail. In a very simple sense, this choice would be determined by whether a trial is primarily intended to address a clinical question about effects upon health status or a question about the social use of health care resources.

The decision about whose preferences are to be reflected in utility measures to a large extent reflects a decision about the purposes of a trial, but may also be influenced by more pragmatic decisions about who can best give informed, unbiased and competent judgements about the value of health states. In certain circumstances the ill may not be able to provide such judgements (Gold *et al.,* 1996). This would not be a substantial problem if it were the case that values in relation to health are stable. There is some evidence that values and preferences are consistent; for example, pretreatment ratings of the utilities of health states did not alter when they actually entered those states (Llewellyn-Thomas *et al.,* 1993). However, against such evidence are those studies that find that patient rating the utility of health states do so far more favourably than those who are invited experimentally to imagine the states (Sackett and Torrance, 1978; Slevin *et al.,* 1990; Fitzpatrick, 1996). It might be argued that the use of community or indirect utilities would disadvantage the ill and disabled for the very reason that more general samples of the well would not have insight into the preferences of the ill. However, the very process whereby patients with health problems make positive adjustments over time could actually result in the value of interventions being underestimated if their own rather than community preference values were used (Fitzpatrick, 1996; Gold *et al.,* 1996).

Advantages and disadvantages

Several advantages are claimed for utility measures over other forms of patient-based outcome assessments (Bennett *et al.,* 1991). Firstly, utility measures provide a quantitative expression of the individual's values and preferences regarding overall health status. The value to an individual of his health state is here distinguishable from descriptions of different aspects of that health state such as level of pain or degree of immobility. A second, related advantage is that a utility measure expresses one single overall value for an individual's preferences regarding health. Utility measures require the integration into one figure of the overall preference for a health state, whereas typically health status measures provide more multi-dimensional data (Feeny and Torrance, 1989). A single summary figure of health benefit is viewed as an advantage particularly when comparisons and choices are needed between the costs and benefits of different treatments. For example, if a patient obtains some relief from pain as a result of a treatment but as a side-effect of treatment is made more tired or depressed, this approach would aim to judge the overall value to the patient of these experiences. A third, and again related, advantage of utility measures is that they are designed to provide numerical values relative to states of perfect health and death (Jette, 1980). This has the consequence that outcome measures such as the quality-adjusted life year (QALY) (Torrance, 1986), can be calculated as a single figure of health benefit which numerically expresses on a single continuum this full range of states. There are other measures such as quality-adjusted time without symptoms (Feldstein, 1991; Johnson, 1993), which are not considered to produce utility measures as such, but do attempt a single figure for health states. The argument for single measures is that mortality and morbidity or health status are otherwise incommensurable making single expressions of health benefit impossible.

Other advantages have been claimed for utility measures which are less easy to test or inspect. In particular, as has already been discussed, it is argued that utility measures derive from a 'rigorous theoretical foundation' (Feeny and Torrance, 1989). By comparison, many other patient-based measures are atheoretical and excessively pragmatic. A body of theory emerged from the work of Von Neumann and Morgenstern about the rational choices individuals make in circumstances of uncertainty and risk (von Neumann and Morgernstern, 1953). Methods of experimentally identifying individuals' utilities such as in standard gamble are considered robust because they conform to the classic axioms of von Neumann and Morgenstern (Gafni, 1994). However, the axioms of rational choice are themselves contested and much empirical evidence suggest that individuals do not behave consistently according to the axioms of decision theory (Sen, 1970; Kahneman and Varey, 1991). Moreover the derivation of measures of health utility from axioms are difficult to demonstrate (Richardson, 1992). It is therefore not easy to consider this a clear advantage of utility-based approaches given the current level of support for theoretical under-pinnings.

There are counterbalancing disadvantages (Feeny and Torrance, 1989). Firstly there is a problem with regard to feasibility. Interview based techniques of eliciting preferences and utilities are labour-intensive and time consuming (Torrance, 1995). Some respondents do not understand the nature of the experimental tasks they are required to perform. Well trained interviewers are therefore needed. This problem of feasibility may be dealt with by using questionnaire-based utility measures such as EQ-5D because this instrument provides indirect utility measures from prior evidence and can be postally administered (Brooks *et al.,* 1996). EQ-5D is short and unlikely to impose the burden on patients that direct elicitation of preferences via an interview may impose. A second problem that arises for indirect measures of utility, as for any explicitly weighted health status measure such as SIP or NHP, is that the value attached to any single health state is a mean or median value around which there is variance. The indirect value may not reflect those of the individual patient being assessed in a trial (Hadorn and Uebersax, 1995). Thirdly, the principle of summarising preferences by a single number is not universally accepted, particularly when individuals' preferences are summed to produce a single figure for the social value of an intervention (Drummond, 1992; Spiegelhalter *et al.,* 1992; Smith and Dobson, 1993). It does not provide information on outcomes that have an intuitive clinical meaning in the context of a clinical trial, such as may be provided by an expression of, for example, a particular percentage reduction in pain or depression levels. By presenting overall utilities in a single value, the direct approach to the measurement of utilities cannot provide the disaggregated evidence on specific dimensions so that it cannot detect or express contradictory trends in different dimensions of outcome. Again, this problem may be overcome if an indirect measure such as EQ-5D is used because this questionnaire provides descriptions of five different dimensions of health status.

When the indirect method of assessing patients' utilities is used, as has been explained, values attached to health states are those obtained from other more general samples. This has required the use of statistical modelling to infer the values attached to some of the possible states of health described by such instruments because samples have only been asked directly to value a core subset. Adequacy of the modelling has been contested (Brooks *et al.,* 1996). Thus indirect methods may not yet provide a complete set of directly elicited values for all combinations of health states (Rutten van Molken *et al.,* 1995a).

Using instruments in combination

Before considering different applications of patient-based outcome measures, it is helpful to note a recommendation that has been made by some authors that the optimal strategy is to use a combination of types of measure in a clinical trial. Most commonly it is recommended that trialists include a generic together with a disease-specific measure (Guyatt *et al.,* 1991; Fletcher *et al.,* 1992; Bombardier *et al.,* 1995). The main argument for such an approach is that the two kinds of measures are likely to produce complementary evidence, with, for example, the disease-specific measure producing evidence most relevant to the clinician and also being most responsive to main effects of an intervention while the generic measure may produce information relevant to a broader policy community (including those requiring comparisons across interventions and disease groups) and may also detect unexpected positive or negative effects of a novel intervention. A further refinement of this strategy is to include a generic instrument with a disease-specific measure as supplement, making efforts to ensure that the disease-specific measure contains items that minimally overlap with those of the generic measure (Patrick and Deyo,1989; Patrick and Erickson, 1993).

However, such a strategy cannot be recommended without caveats. In the first place, the addition of questionnaire items may impose a burden on patients that reduces overall compliance. This effect may be increased if respondents have to answer items with overlapping content. The repetitiveness that may attend such an approach may appear insensitive on the part of investigators. Secondly, the addition of each scale or instrument increases the number of statistical analyses and therefore significant effects arising by chance, although this can problem can be managed by disciplined identification of prior hypotheses. A compromise strategy is to include a battery of selected questionnaire items from different types of measures, rather than whole scales. The clear danger with this strategy is that items removed from their context of whole instruments may not retain the measurement properties (such as reliability and validity) of the whole instrument, so that this approach has least to recommend it.

Applications

As already stated, this text is intended to be a guide in the use of patient-based outcome measures for clinical trials. However, it is important to recognise that such measures have been developed for a wide range of different uses (Hunt, 1988; Fitzpatrick, 1994; Fitzpatrick and Albrecht, 1994). Some instruments are considered to be applicable not just as outcome measures in clinical trials but as instruments that can also be used to assess the health care needs of populations and assist health professionals in assessing and caring for individual patients. However insufficient attention has been given to the different kinds of uses to which instruments can be put (Sutherland and Till, 1993; Till *et al.,* 1994). This is a serious omission because a questionnaire may have been established as having considerable validity in, for example, assessing health problems as a screening instrument in hospital clinics whilst having less relevance as a measure of outcome assessing changes in the health status of the same patient group. The range of alternative applications is briefly considered.

Clinical trials and cost–utility studies

The current review has been written with this application in mind. There is far more agreement about the potential and appropriateness of patient-based outcome measures as endpoints in clinical trials (Pocock, 1991). It is increasingly argued that clinical trials should incorporate patient-based outcome measures such as health status and QoL except in circumstances where it is clear that these issues are not relevant outcomes (Ganz *et al.,* 1992; Kaasa, 1992; Ganz, 1994). In some fields such as cancer trials and surgery, thought has been given to the circumstances when it is or is not relevant to include such outcomes (Neugebauer *et al.,* 1991; Gotay *et al.,* 1992; Hopwood, 1992; Nayfield *et al.,* 1992; Osoba, 1992). The clearest role for such outcome measures is in the 'gold standard' form of randomised controlled trial. Patient-based outcome measures have been used as the primary outcome, in randomised controlled trials, in a variety of fields including cancer, rheumatology and heart disease.

Improvement in QoL was used to compare intermittent and continuous chemotherapy treatment in women with advanced breast cancer and found in favour of a continuous strategy (Coates *et al.,* 1987). QoL has been used in a similar manner to compare different treatment strategies in prostate cancer (Keoghane *et al.,* 1996), small-cell lung cancer (Gower *et al.,* 1995) and acute myeloid leukaemia (Stalfelt, 1994). Olsson *et al* (1986) compared metoprolol to placebo in patients with myocardial infarction and found that the treatment improved QoL. The effect of drug treatment on QoL has also been evaluated in heart failure (Bulpitt, 1996) and hypertension (Applegate *et al.,* 1994). QoL was also used as the primary outcome in a clinical trial to compare surgical techniques used in hip arthroplasty, which found no difference between cement versus cementless total hip arthroplastry (Rorabeck *et al.,* 1994).

When investigators also need to obtain evidence of the overall value of a health care intervention in a way that permits comparison with other interventions, whether in the same treatment area or across areas, then outcomes that provide evidence of the overall value to patients of outcomes in the form of utilities are required. The most widely known form of summary value of treatments for comparative purposes is the QALY (Torrance, 1986). The debate about the validity of QALY is beyond the scope of this review and is considered elsewhere (Carr-Hill, 1989; Carr-Hill and Morris, 1991; Coast, 1992; Drummond, 1993; Nord, 1993; Petrou *et al.,* 1993; Smith and Dobson, 1993). In relation to the current review, it is increasingly argued that data for such analysis should be obtained from patients participating in a clinical trial in order that they provide responses to utility-based assessments as well as other data on health status.

Patient-based outcome measures may also be used in non-randomised research designs, although the interpretation of results will be more complex, as is the case with any other form of outcome measure. The overall objective of such uses is similar to that of the randomised clinical trials, to detect differences between groups experiencing different interventions, but for one of a number of reasons observational evidence is used (Ware *et al.,* 1996a). Patient-based outcome measures make such large-scale studies more feasible. It is beyond the scope of the review to address broader questions as to whether observational studies of outcomes of interventions ever fully address issues of bias as successfully as do randomised designs.

In summary, the use of patient-based outcome measures is far more developed than other applications. That instruments have been shown to have validity, appropriateness and other desirable properties for use in randomised controlled trials does not mean that they can be automatically transferred to other uses. The third section of this review primarily has in mind randomised clinical trials and cost–utility studies associated with trials in outlining the criteria in terms of which patient-based outcome measures should be evaluated and chosen by investigators. However, there are two other different types of use that have been argued for patient-based outcome measures: assessing the health of populations and as an aid in individual patient care.

Assessing the health care needs of populations

Health authorities and those responsible for purchasing or providing health care are increasingly expected to base their decisions about the allocation of health care resources on evidence (Scrivens *et al.,* 1985; Kelly *et al.,* 1996). Evidence of health care needs comes from epidemiological data. These may take the form of conventional data on mortality and morbidity or be derived from social, demographic and other indirect measures that may indicate health needs. It has been argued that patient-based outcome measures provide a feasible and valid measure of health status that complements existing approaches, especially in so far as they focus upon felt and experienced health problems (Hunt *et al.,* 1985; Ventegodt, 1996). Particularly if such assessments are based on self-completed questionnaires with proven acceptability, this approach offers the possibility of using social survey methods to assess aspects of health. Surveys have been conducted on particular geographical populations (Curtis, 1987) and specific social groups such as ethnic minorities and the unemployed (Ahmad *et al.,* 1989). A related use of patient-based outcome measures is in combination with mortality data, for example in measures such as health life expectancy, and disability free life expectancy (Robine and Ritchie, 1991). To be most useful in population settings, a questionnaire, as well as being feasible and acceptable, needs to provide information that indicates needs for particular kinds of health or other services. The main problem with this use is that such instruments provide only general indications of health problems. Although there is growing evidence that poor scores on health status measures may be associated with and predictive of elevated rates of subsequent health service use and mortality, they do not provide evidence of more

specific needs to be addressed (Frankel, 1991). There is therefore little evidence in the literature of patient-based outcome measures adding to existing sources of health status in informing population-level decision-making.

Individual patient care

It has been argued that patient-based outcome measures offer an important adjunct to clinicians in the care of their patients (Tarlov *et al.,* 1989; Anonymous, 1991b). Self-completed questionnaires, if proved reliable and valid, offer a quick and reliable way for patients to provide evidence of how they view their health that complements what the clinician collects from clinical and other evidence (Nelson and Berwick, 1989). The value of this additional information is partly because time pressures increasingly constrain health professionals and limit the amount of direct contact. The primary purposes of inviting patients to complete health status questionnaires are to enable health professionals to screen for health problems that may not otherwise become apparent and to monitor the progress of problems identified in the patient and the outcomes of any treatments. Patient-based outcome measures can provide prognostic information about the cause of illness independent of diagnosis (Mauskopf *et al.,* 1995). It is also been argued that such measures can be used to select patients for treatment, for example identifying patients able to undergo surgery (O'Boyle, 1992). Reports have appeared arguing that it is feasible to incorporate shorter measures into the routines of clinical practice (Nelson *et al.,* 1990; Wolfe and Pincus, 1991). When patients are asked about the value of such requests, the majority are positive and consider the information conveyed by questionnaires important for health professionals to know about them (Nelson *et al.,* 1990; Street *et al.,* 1994). Doctors also find the information of value (Young and Chamberlain, 1987; Williams, 1988). However, clinicians report that whilst they regard the issues raised by such instruments as very important, they are not able to make systematic and regular use of information about the QoL of their patients as provided by questionnaires (Taylor *et al.,* 1996). In some fields of medicine, more systematic trials to evaluate the impact of providing clinicians with information from patients in this way have found little evidence that clinical decisions are changed because of the additional data about their patients and health status is not improved (Rubenstein *et al.,* 1989; Kazis *et al.,* 1990). One likely explanation for the apparent lack of impact of patient-based outcome measures on clinical practice is that it is still not clear how to present the data in useful forms and how clinicians should make use of the evidence (Deyo and Patrick, 1989).

From a more formal perspective, the precision of the score from an individual (as in the context of clinical care) is less than that obtained for a group of patients. There is considerable measurement error in the numerical value of an individual respondent's report. This has the consequence that usefulness in the context of individual patient care will be more difficult.

In summary, there is little evidence to date to support the use of patient-based outcome measures in routine practice and more trials are needed to examine their usefulness in this context (Long and Dixon, 1996). It may be that the existence of 'normative' data from representative samples of the general population will facilitate the interpretation of some instruments (Jenkinson *et al.,* 1996).

Chapter 3

Criteria for selecting a patient-based outcome measure

The third section of this review examines the ways in which patient-based outcome measures need to be assessed. In summary, it is argued that there are eight dimensions in terms of which a patient-based measure can be examined. Evidence about a measure that is being considered for inclusion in a trial needs to be considered in terms of the following issues: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability and feasibility. As has already been pointed out, these criteria are not rank ordered in terms of importance and do not follow any sequential logic in terms of how they should be approached. For each of the criteria, the evidence and nature of current views is summarised in order to make clearer to the reader what is meant by a criterion. Three of the criteria, appropriateness, precision and interpretability, have increasingly been discussed in the literature but are less likely to appear on check-lists in many standard discussions. For these criteria, although it is clear from the literature that they are important, there is no uniform language or framework in terms of which they are discussed. The remaining criteria; reliability, validity, responsiveness, acceptability and feasibility are more often cited on standard lists and discussions. In the case of reliability, validity and responsiveness, this in part reflects their widespread usage in the field of psychometrics. For none of the criteria are there absolutely explicitly defined and universally accepted understandings of the terms; in many areas there remain uncertainties and differences of view. The purpose of this section is steer the reader through current debates about the eight criteria in a helpful way.

Appropriateness

Is the content of the instrument appropriate to the questions which the clinical trial is intended to address?

The first and most fundamental consideration to be faced when selecting a patient-based outcome measure is how to identify one that is most appropriate to the aims of the particular trial. This requires careful consideration of the aims of the trial, with reference to the QoL research questions, i.e. which dimensions will be primary and secondary end points, the nature of the study intervention and of the patient group and about the content of possible candidate instruments. For this reason, it is particularly difficult to give specific recommendations about what in general makes an outcome measure appropriate to a trial, because this is ultimately a judgement of the fit between investigators' specific trial questions and content of instruments. However, it is clear from a number of reviews already carried out in this field that it is an absolutely fundamental issue.

There have been several previous reviews that have discussed appropriateness of outcome measures in clinical trials in general terms. Some of the early reviews are mainly critical of clinical trials for failing to use any kind of patient-based outcome measure where the subject matter seemed to indicate that such an approach was needed. Thus Brook and Kamberg (1987) conducted a MEDLINE review of clinical trials and concluded that, from a sample of 73 clinical trials in which they considered health status or QoL was likely to be a major issue, in only two trials was an appropriate patient-based outcome measure used. Najman and Levine (1981) reached the same conclusions from a range of trials in an earlier review. A third review finds evidence of trialists failing to use appropriate outcome measures even where title, keywords or abstract include 'quality of life'. Schumacher *et al.* (1991) reviewed 67 such trials in the fields of oncology and cardiology and found that 43% of studies included no serious assessment of QoL at all and a further 24% assessed a limited single aspect that the reviewers considered inadequate.

A more formal evaluation of outcome measurement in trials is reported by Guyatt and colleagues (1989b). In their study, two raters independently examined all clinical trials published in a range of journals in 1986. Of the 75 trials they evaluated, they considered QoL as crucial or important in 55 (73%) of trials. However, in 44% of this subgroup of trials, no effort was made to assess this dimension of outcome. In a further 38% of

the 55 trials, an untested measure was used that the reviewers considered inappropriate. They concluded that appropriate measures would have considerably strengthened the basis for recommendations emerging from the trial.

Fundamental to such critiques is the view that measures of outcome used in trials and intended to assess the patient's perspective are often limited or superficial. The strongest expression of this criticism can be found in a recent review by Gill and Feinstein (1994). They reviewed 75 articles selected randomly from medical journals that include QoL measurement in the context of health care research. They rated the use of QoL measures in studies on a range of explicit criteria with overall scores for articles ranging from 0 to 100. Only 11% of articles achieved scores of 50 or more (i.e. 'satisfactory' by at least half their criteria). They were particularly critical that in 85% of articles, authors had not defined QoL for the purpose of the study and in 52% of articles had not explained or justified their selection of QoL measure. They also considered unsatisfactory the fact that in 83% of studies, patients were not invited, in addition to other questions, to respond to a global overall rating of their QoL. Their overview reveals quite stringent criteria for what constitutes an appropriate patient-based outcome measure. Guyatt and Cook (1994) re-examined 15 randomly selected papers from the sample reviewed by Gill and Feinstein according to their own criteria. By their criteria, only 33% failed to use questionnaire items reflecting matters of importance to patients and only 27% of studies used measures that omitted important items. It is apparent from the much more favourable ratings by Guyatt and Cook that they employed different criteria to judge appropriateness of outcomes. This is clearly seen if the full set of criteria of the two reviews are examined in full *(Box 4)*.

Both reviews clearly emphasise the need for appropriate QoL measures to incorporate questionnaire items that clearly matter to patients. However they also differ in important respects. Gill and Feinstein (1994) argue for the need for measures that are based on an explicit definition of QoL, that provide a single global score, that allow patients to state the relative importance of issues, that allow patients to give supplementary answers not included in the questionnaire and to globally rate QoL and health-related QoL. Underlying their arguments is a view that instruments have been too dominated by what they term 'psychometric' principles of reliability and validity and insufficient attention has been given to clinical 'face validity', which is largely established by judgements and statements made by patients unconstrained by the format of fixed questionnaire items.

BOX 4 Two competing conceptions of requirements for judging appropriateness of outcome measures

1. Is QoL conceptually defined?
2. Are domains intended to be measured explicitly stated?
3. Are selected outcome measures explained or justified?
4. Are scores aggregated into a single overall score?
5. Are patients able to offer a separate global rating of QoL?
6. Is a distinction made between overall versus health-related QoL?
7. Are patients able to add supplemental comments?
8. Are patients able to rank the importance of individual items?

(Adapted from Gill and Feinstein, 1994)

1. Do the authors show that aspects of patients' lives measured are important to patients?
2. Have previous studies demonstrated their importance?
3. Do investigators examine aspects of patients' lives that, from clinical experience, it is known that patients value?
4. Are there aspects of health-related quality of life that are important to patients but have been omitted?
5. Are individual patients asked directly to place a value on their lives?
6. Are instruments used demonstrated to have reliability, validity and responsiveness?
7. Do instruments used have interpretability (i.e. distinguish trivial from important differences)?

(Adapted from Guyatt and Cook, 1994)

By contrast, Guyatt and Cook, whilst accepting that the primary focus of instruments should be on matters of importance to patients, argue that Gill and Feinstein's requirements are too stringent. They appear to argue that Gill and Feinstein place too much emphasis upon eliciting the values and priorities within a given study and insufficient attention to use of established instruments in which prior development and use have identified matters of importance. At the heart of this dispute are two different ways of ensuring that patients' preferences and concerns are fully incorporated into trial study design, by extensive use of global and

supplementary questions, favoured by Gill and Feinstein, or by use of previously validated instruments, favoured by Guyatt and Cook. With the former method, the patient is directly asked his or her judgements in an open ended way, minimally constrained by predetermined questionnaire items. The latter method largely relies on predetermined questions validated in previous research. To some extent, the dispute reflects differences of philosophy in how to assess patients' experiences that cannot be resolved by current evidence. Their common ground is that appropriate measures in a trial are those that particularly address patients' concerns.

If clinical trials do use instruments that have been developed by psychometric criteria, they may still be flawed. Psychometric principles (reliability, validity and responsiveness) are further explored in later sections of this chapter. Coste and colleagues (1995) reviewed 46 studies published in six medical journals over the period 1988 to 1992, in which scales or indices were used to measure constructs such as QoL and physical function. In less than a fifth of studies did they regard construct and content validity to have been adequately addressed. In only a quarter of studies was adequate attention given to reliability. They describe many of the instruments as being '*ad-hoc*'. Another review, in which independent assessment of studies was made, found that in randomised clinical trials that used patient-based outcome measures, only 10 out of 55 trials used instruments with established validity and responsiveness (Veldhuyzen van Zanten, 1991).

Instruments do need to be clearly focused on patients' concerns and to be psychometrically sound to be considered as appropriate based measures of outcome in trials. However, these properties do not exhaust the list of considerations in determining whether an instrument is appropriate for any particular trial.

Most obviously, an instrument needs to fit the purpose of a trial. This purpose needs to be specified as precisely as is reasonable and outcome measures selected accordingly (Liang *et al.,* 1982; Fallowfield, 1996). A *Lancet* editorial reiterates part of Gill and Feinstein's critique and argues that the rationale for selection of outcome measures is often not clear (Editorial, 1995). Investigators are uncritically inserting questionnaires into their trials without careful consideration of content and relevance to the purpose of the trial. This will primarily mean that the instrument selected must be particularly relevant to the health problem and proposed intervention as possible.

As already stated, this judgement involves simultaneous examination of, on the one hand, the specific treatment and patient group being investigated and on the other the content of instruments in order optimally to match instrument to objective (Guyatt *et al.,* 1991). Investigators have to determine how narrow or broad a measure of health they require. An intervention may be evaluated in which only very accurate assessment of, say, mobility or pain is needed. More often, investigators are uncertain of all the likely consequences of their intervention and opt for a broader measure or set of measures to capture more unexpected consequences.

A useful distinction can also be made between 'proximal' and 'distal' outcome measures (Brenner *et al.,* 1995). Brenner and colleagues suggest that it is helpful to think of a continuum of outcomes in relation to any disease and its treatment. Outcomes that are proximal most closely represent manifestations of the disease itself, for example, pain and stiffness in arthritis. Slightly less proximal and removed from disease are aspects of physical functioning. Distal outcomes are those most removed from disease such as, for example, life satisfaction. The value of the continuum is in making explicit that as one incorporates more distal outcome measures in a trial, the less likely it is that the intervention will have greater effects on those outcomes in the study group compared to controls. On the other hand, they suggest that the more effective an intervention, the greater the likelihood will be that more distal outcome measures will be relevant. Circumstances of a trial will dictate whether distal as well as proximal effects are of interest and therefore important to monitor. It is a useful discipline to consider this continuum in selecting outcome measures.

It is impossible to be clear simply from the titles of instruments or of their constituent scales and dimensions what precisely is being measured. Titles of instruments and constituent scales of instruments cannot be taken at face value, and cannot therefore be assumed to be appropriate on the basis of title alone (Ware, 1987). This is most obviously the case for dimensions of instruments which refer very broadly to, for example, 'social function'. Dimensions of instruments assessing this aspect of patients' experiences may refer to quite disparate issues. Thus two patient-based outcome measures for cancer provide 'social scores' but only weakly agree with each other when patients

completed both (King *et al.,* 1996). The reason for the low level of agreement is that the items of one scale focus upon companionship of family and friends, whilst the other instrument's social scale focuses upon impact of disease on social activities. The same degree of disparate content was found in social dimensions of instruments used to assess well-being in patients with rheumatoid arthritis (Fitzpatrick *et al.,* 1991). Instruments focusing on physical function may also differ in less obvious ways in their content when assessing dimensions such as physical function about which more agreement might be expected. For example, the physical function of patients with rheumatoid arthritis is assessed in one health status instrument by items that ask respondents how much help they need to perform particular tasks, another instrument addresses similar tasks but questionnaire items elicit the degree of difficulty experienced by respondents with tasks (Ziebland *et al.,* 1993).

One commonly recommended solution to ensure that a trial will have an appropriate set of outcome measures is that one disease-specific and one generic instrument be used to assess outcomes (Cox *et al.,* 1992; Bombardier *et al.,* 1995). In this way, it is reasonably likely that both important proximal and distal effects of a treatment will be captured; detecting the most immediate effects upon disease as well as possible consequences that are harder to anticipate.

Summary

In more general terms, appropriateness of an instrument for a trial will involve considering the other criteria we have identified and discuss below; evidence of reliability, feasibility, and so on. In the more specific terms with which we have summarised the rather disparate literature on appropriateness, the term requires that investigators consider as directly as possible how well the content of an instrument matches the intended purpose of their specific trial.

Reliability

Does the instrument produce results that are reproducible and internally consistent?

Reliability is concerned with the reproducibility and internal consistency of a measuring instrument. It assesses the extent to which the instrument is free from random error and may be considered as the amount of a score that is signal rather than noise. It is a very important property of any patient-based outcome measure in a clinical trial because it is essential to establish that any changes observed in a trial are due to the intervention and not to problems in the measuring instrument. As the random error of such a measure increases, so the size of the sample required to obtain a precise estimate of effects in a trial will increase. An unreliable measure may therefore underestimate the size of benefit obtained from an intervention. The reliability of a particular measure is not a fixed property, but is dependent upon the context and population studied (Streiner and Norman, 1995).

The degree of reliability required of an instrument used to assess individuals is higher than that required to assess groups (Williams and Naylor, 1992; Nunnally and Bernstein, 1994). As is described below, reliability coefficients of 0.70 may be acceptable for measures in a study of a group of patients in a clinical trial. However, Nunnally and Bernstein (1994) recommend that a reliability level of at least 0.90 is required for a measure if it is going to be used for decisions about an individual on the basis of his or her score. This higher requirement is because the confidence interval around an individual's true score are wide at reliabilities below this recommended level (Hayes *et al.,* 1993). For a similar reason Jaeschke and colleagues (1991) express extreme caution about the interpretation of QoL scores in N of one trials. Our concern is with group applications such as in trials where the confidence interval around an estimate of the reliability of a measure is increased as sample size increases.

In practice, the evaluation of reliability is in terms of two different aspects of a measure: internal consistency and reproducibility (sometimes referred to as 'equivalence' and 'stability' respectively (Bohrnstedt, 1983). The two measures derive from classical measurement theory which regards any observation as the sum of two components, a true score and an error term (Bravo and Potvin, 1991).

Internal consistency

Normally, more than one questionnaire item is used to measure a dimension or construct. This is because of a basic principle of measurement that several related observations will produce a more reliable estimate than one. For this to be true, the items all need to be homogeneous, that is all measuring aspects of a single attribute or construct rather than different constructs (Streiner and Norman, 1995). The practical consequence of this expectation is that individual items should highly correlate with each other

and with the summed score of the total of items in the same scale.

Internal consistency can be measured in a number of different ways. One approach – split-half reliability – is randomly to divide the items in a scale into two groups and to assess the degree of agreement between the two halves. The two halves should correlate highly. An extension of this principle is Coefficient alpha, usually referred to as Cronbach's alpha, which essentially estimates the average level of agreement of all the possible ways of performing split-half tests (Cronbach, 1951). The higher the alpha, the higher the internal consistency. However, it is also possible to increase Cronbach's alpha by increasing the number of items, even if the average level of correlation does not change (Streiner and Norman, 1995). Also if the items of a scale correlate perfectly with each other, it is likely that there is some redundancy among items, and also a possibility that the items together are addressing a rather narrow aspect of an attribute. For these reasons, it is suggested that Cronbach's alpha should be above 0.70 but not higher than 0.90 (Nunnally and Bernstein, 1994; Streiner and Norman, 1995).

Another approach to establish internal consistency of items is simply to examine the correlation of individual items to the scale as a whole, omitting that item. Steiner and Norman (1995) cite a normal rule of thumb that items should correlate at least 0.20 with the scale.

A balance needs to be struck between satisfactory internal consistency and a measure that is too homogeneous because it measures a very restricted aspect of a phenomenon. Kessler and Mroczek (1995) provide an important argument with illustrative evidence against excessive emphasis upon internal reliability. Essentially, they advocate a shift toward selecting items in a scale in a way that maximises the additional information content of each item, a principle of minimal redundancy. As Kessler and Mroczek argue, investigators usually use factor-analytic techniques to identify items that particularly correlate and therefore yield high internal reliability. They argue for the use of regression and related techniques to replace factor analytic methods of developing scales. Their argument begins with a hypothetical long list of questionnaire items that together may be considered the complete and true measure of some phenomenon, say pain, or mobility. The objective of scale development is to produce a small sub-set that reliably measures the full set. Factor analysis will identify the items from the longer set that most correlate with each other to produce an internally reliable scale. If, however, one conceives of the full set of items as the dependent variable and uses regression analysis with individual items as the independent variables, it is possible to identify a small sub-set of items that explains most of the variance in the full set. As they express it, 'most of the variance in the long form can then usually be reproduced with a small subset of the scale items' (Kessler and Mroczek, 1995: AS112). The argument is then illustrated with data comprising 32 items used to screen for psychological distress in a population survey. They select items to form a short version of the scale from the full set by two methods: factor analytic methods to maximise internal reliability and regression to minimise redundancy. Results from regression produce consistently higher correlations of the sub-scale with the total variance in the full set of 32 items. On the other hand, factor analytic techniques produce consistently higher internal reliability.

It has been argued that excessive attention to internal reliability can result in the omission of important items, particularly those that reflect the complexity and diversity of a phenomenon (Donovan *et al.,* 1993). Certainly, obtaining the highest possible reliability coefficient should not be the sole objective in developing or selecting an instrument because the reductio ad absurdum of this principle would be an instrument with high reliability produced by virtually identical items.

Reproducibility

Reproducibility more directly evaluates whether an instrument yields the same results on repeated applications, when respondents have not changed on the domain being measured. This is assessed by test–retest reliability. The degree of agreement is examined between scores at a first assessment and when reassessed. There is no exact agreement about the length of time that should elapse; it needs to be a sufficient length of time that respondents are unlikely to recall their previous answers, but not so long that actual changes in the underlying dimension of health have occurred. Streiner and Norman (1995) suggest that the usual range of time elapsed between assessments tends to be between 2 and 14 days. One way of checking whether the sample has experienced underlying changes in health that would reduce the apparent reproducibility of an instrument is also to administer a transition question at the second assessment ('Is your health better, the same or worse than at the last assessment?').

Test–retest reliability is commonly examined by means of a correlation coefficient. This is often the Pearson product moment correlation coefficient. This approach is limited and may exaggerate reproducibility because results from two administrations of a test may correlate highly but be systematically different. The second test may result in every respondent having a lower score than their first response, yet the correlation could be 1.0. For this reason, an intra-class correlation coefficient is advocated. This uses analysis of variance to determine how much of the total variability in scores is due to true differences between individuals and how much due to variability in measurement.

It has been argued that correlation coefficients measure the strength of association between two measures and not the extent of agreement (Bland and Altman, 1986). Bland and Altman advocate graphically plotting scores from the two administrations of a test, so that, for example, it is possible to identify areas in the range of scores of an instrument which are less reproducible.

The confidence we may have in any estimate of the reliability of an instrument is influenced by the sample size from which the estimate was obtained (Eliasziw and Donner, 1987). The greater the sample size, the greater our confidence. Some authorities suggest that sample sizes required to test reliability accurately are in the range 200–300 (Kline, 1986; Nunnally and Bernstein, 1994). However, Streiner and Norman (1995) estimate that sample sizes needed are less than 200 provided that investigators accept a confidence interval of ± 0.10.

Commonly cited minimal standards for reliability coefficients are 0.7 for group data, although some experts set much higher requirements (Scientific Advisory Committee of the Medical Outcomes Trust, 1995). It can also be argued that absolute minimally acceptable coefficients are not meaningful, since larger sample sizes for a trial permit more measurement error in an instrument. As any statement of the reliability of an instrument is based on sample statistics, the more frequently this property is studied and reported in different populations, the greater will be confidence in estimates of its reliability (Williams and Naylor, 1992).

Inter-rater-reliability

Inter-rater reliability is not formally considered in this review as an aspect of reliability because it is concerned with agreement between observers or interviewers rather than self-administered instruments. It is an important issue although outside the scope of this review when proxy reports are required because of the incapacity of the subject. High levels of inter-rater reliability may be obtained from training or specialist expertise of raters that may not apply when an instrument is more widely used (Cox *et al.*, 1992).

Summary

Overall, we are here concerned with how reproducible an instrument is, and, where relevant, how internally consistent items are in scales. Reproducibility can be assessed by fairly specific methods so that it is a relatively straightforward aspect of an instrument to assess when such evidence is available. In reality, internal consistency is more frequently reported although very high internal consistency is not always considered desirable.

Validity

Does the instrument measure what it claims to measure?

The validity of a measure is an assessment of the extent to which it measures what it purports to measure. There are a number of different ways of establishing the validity of a measure. As with reliability, it is not a fixed property of a measure; its validity is assessed in relation to a specific purpose and setting (Jenkinson, 1995). It is therefore meaningless to refer to a validated measure; it should be considered a measure validated for use in relation to a specific purpose or set of purposes. For example, a valid measure of disability for patients with arthritis cannot automatically be considered valid for use for patients with multiple sclerosis; a measure considered validated for individuals with mild impairment may not be valid for those with severe impairments.

Criterion and predictive validity

Criterion validity is involved when a proposed new measure correlates with another measure generally accepted as a more accurate or criterion variable. However, in the field of application of health status measures with which we are concerned, as outcome measures in clinical trials, it is rarely if ever that a perfect 'gold-standard' measure exists against which to test the validity of new health status measure, and a number of different and more indirect approaches are recommended to judge instruments' validity (Patrick and Erickson, 1993b). One exception may be when a longer version of a questionnaire is used as the 'gold standard' to develop a shorter version of the same established instrument (Hickey *et al.*, 1996b; Ware *et al.*,

1996b). When the new measure correlates with future values of the criterion variable, then we are concerned with predictive validity. For example, a new measure of psychological well-being may be predictive of individuals' future medical consultations for psychological problems. Again, this is not as important or relevant an issue with patient-based outcome measures used in clinical trials.

Face and content validity

Face, content, and (below) construct validity are far the most relevant issues for the use of patient-based outcome measures in trials. It is vital to inspect the content of a measure in relation to its intended purpose. This inspection largely involves qualitative matters of judgement that contrast with more statistical criteria that also need to be considered in the context of construct validity (discussed below). Judgement of the content of an instrument contributes to what has been termed face validity and content validity. The two terms are related but have been distinguished in the following way: face validity refers to 'what an item appears to measure based on its manifest content' (Ware *et al.,* 1981:623). Content validity refers to 'how well a measurement battery covers important parts of the health components to be measured' (ibid). Guyatt and colleagues make the distinction thus: 'Face validity examines whether an instrument appears to be measuring what it is intended to measure, and content validity examines the extent to which the domain of interest is comprehensively sampled by the items, or questions, in the instrument.' (Guyatt *et al.,* 1993b:624). Together, they address whether items clearly address the intended subject matter and whether the range of aspects are adequately covered. Face validity can overlap with judgements of the interpretability of items, but these aspects are kept separate here. Face and content validity need to be inspected, literally by examining the questionnaire. Because they cannot be so readily measured statistically, these aspects of validity tend, wrongly, to be dealt with more cursorily than is construct validity (Feinstein, 1987). Another important source of evidence can be obtained from evidence of how the questionnaire was developed in the first place. How extensively did individuals with relevant clinical or health status methodology expertise participate in generating the content (Guyatt and Cook, 1994)? Even more importantly, to what extent did patients with experience of the health problem participate in generating and confirming the content of an instrument (Lomas *et al.,* 1987). It is still quite common for the content of questionnaires to be determined by 'experts' alone (Chambers *et al.,* 1982). Whilst knowledgeable about an illness, they cannot substitute completely for the direct experience that patients have of health problems. Guyatt *et al.,* (1986) describe different degrees of effort to establish validity by an analogy to cars. The Rolls-Royce model takes extensive steps in the construction of the questionnaire and involves patients at every phase of its development. By contrast, the Volkswagen model reduces the process, usually for resource reasons, for example by relying solely on expert opinion to determine content, thereby leaving validity relatively untested.

Construct validity

A more quantitative form of assessing the validity of an instrument is also necessary. This involves construct validity. A health status measure is intended to assess a postulated underlying construct, such as pain, isolation or disability rather than some directly observable phenomenon. The items of a questionnaire represent something important other than a numerical score but that 'something' is not directly observable. This construct, for example, pain or disability, can be expected to have a set of quantitative relationships with other constructs on the basis of current understanding. Individuals experiencing more severe pain may be expected to take more analgesics; individuals with greater disability to have less range of movement in their environment. Construct validity is examined by quantitatively examining relationships of a construct to a set of other variables. No single observation can prove the construct validity of a new measure; rather it is necessary to build up a picture from a broad pattern of relationships of the new measure with other variables (Bergner and Rothman, 1987). Patient-based outcome measures are sometimes presented as 'validated' because they have been shown to agree with clinical or laboratory evidence of disease severity. Whilst such evidence provides an aspect of construct validity, it is not sufficient. As Streiner and Norman observe (1995:9) 'the burden of evidence in testing construct validity arises not from a single powerful experiment, but from a series of converging experiments.'

There are no agreed standards for how high correlations should be between an instrument or scale being assessed and other variables in order to establish construct validity (Avis and Smith, 1994). It is very unlikely that correlations of a new measure, of, for example, mobility, would reach 1.00. In reality, that is only likely to be achieved by measuring the same thing twice which would undermine the very point of the new measure. Also in statistical terms the upper limit of the correlation between two variables is set by

the product of these variables' reliability coefficients. Therefore, given typical levels of reliability of patient-based variables, a correlation coefficient of 0.60 may be strong evidence in support of construct validity (McDowell and Newell, 1996). Because there is considerable vagueness and variability in the levels of correlation coefficients that authors cite as evidence of construct validity of new instruments, McDowell and Jenkinson (1996) recommend that expected correlations should be specified at the outset of studies to test instruments' validity in order that it be possible for validity to be disproved.

The most sophisticated form of testing construct validity is so-called 'convergent and discriminant validity' (Campbell and Fiske, 1959). This approach requires postulating that an instrument that we wish to test should have stronger relationships with some variables and weaker relationships with others. A new measure of mobility should correlate more strongly with existing measures of physical disability than with existing measures of emotional well-being. Essentially correlations are expected to be strongest with most related constructs and weakest with most distally related constructs. Typically, construct validity is examined by inspecting correlations of a new measure against a range of other evidence such as, disease staging, performance status, clinical or laboratory evidence of disease severity, illness behaviour, use of health services and related constructs of well-being (Spitzer *et al.,* 1981; Fletcher, 1988; Sullivan *et al.,* 1990; Aaronson *et al.,* 1993). As an example of convergent and discriminant validation, Sullivan and colleagues (1990) examined validity of the scales of the SIP for use in rheumatoid arthritis with the expectation that the SIP physical function score should correlate most with various measures of disease severity and the SIP psychosocial scale should correlate most with other measures of mood and psychological well-being. Conversely, measures of physical function were expected to correlate less with measures of physical mood. Similarly, Morrow and colleagues (1992) examined the convergent-discriminant validity of the Functional Living Index-Cancer (FLIC) by examining relationships to other variables. As predicted the subscales of FLIC to measure gastrointestinal symptoms was significantly related to other ratings by patients of nausea and vomiting, but correlations were close to zero between psychological and social subscales of FLIC and patients' separate reports of nausea and vomiting.

The most demanding form of convergent and discriminant validity is the multitrait-multimethod matrix (Campbell and Fiske, 1959) in which two unrelated constructs are measured by two or more methods. In essence, it is expected that different measures of a single underlying trait should correlate most and different measures of different constructs least. Whilst potentially powerful in psychometric test development, it is difficult to apply to patient-based outcome measures because of problems of obtaining alternative methods of measuring constructs.

Most instruments to assess outcomes from the patient's point of view are multi-dimensional. They, for example, assess physical, psychological and social aspects of an illness within one questionnaire. This internal structure of an instrument can also be considered a set of assumed relationships between underlying constructs. At the very least, an instrument with sub-scales has implied that the instrument measures different underlying constructs by providing different sub-scales, rather than requiring that all items should simply be added to produce one score of one underlying construct. Normally instruments with multiple scales also assume particular underlying relationships for the constructs measured by the instrument; for example, scales of different aspects of emotional response to illness will correlate more with each other than with scales assessing physical function. This internal structure of instruments has also to be established by construct validation. The most common of methods for this purpose is statistical, particularly the use of factor analysis. Thus, factor analysis is often considered an aspect of construct validity.

To simplify, factor analysis is the analysis of patterns of, in this field, items that go together to assess single underlying constructs. Typically, statistical analysis of answers of a sample of respondents to a pool of questionnaire items is used to reveal two or more sub-scales assessing distinct dimensions. In essence, the data can be checked to see whether individual questionnaire items correlate more with the scale of which they are a part and less with other scales to which they do not belong. Examples of instruments in which factor analysis has played a key role in establishing the internal structure of sub-scales include the Profile of Mood States (McNair *et al.,* 1992) and the St George's Respiratory Questionnaire (Jones *et al.,* 1992).

However, there are problems with factor analytic methods used in this context. Fayers and Hand (1997) provide important arguments and evidence against excessive reliance upon factor analysis alone to determine or evaluate the construct

validity of instruments. In an analysis of quality of life of patients participating in a drug trial for colorectal cancer, they show that factor analysis of pooled results for the Hospital Anxiety and Depression Scale produces a satisfactory solution with the two expected dimensions of anxiety and depression clearly emerging. In other words, items tapping these two psychological experiences cluster together and factor analysis proved an appropriate method of identifying constructs. By contrast, when factor analysis was carried out for the same sample of patients' results for the Rotterdam Symptom Checklist, a four factor solution emerged, with one factor addressing a heterogeneous list of disease-related symptoms, such as loss of appetite and decreased sexual interest. This factor also appears unstable across studies. They argue that this 'factor' probably reflected specific treatment effects associated with one of the randomised drug regimes. They argue, more generally, that experiences such as symptoms in particular, whilst of major importance to patients, are causally unrelated to the more psychological factors emphasised in QoL, so that they may not be associated with or contribute to a factor. In studies in which items of importance to patients such as symptoms and side-effects of treatments do not cluster together or with other items, they may therefore be omitted altogether from the development of an instrument if principles of factor analysis are too strictly adhered to. Fayers and Hand advocate supplementing statistical analysis of factors with other techniques such as directly asking patients to identify important or omitted issues in the development of appropriate instruments.

Usually investigators use exploratory factor analysis to examine whether there is any underlying pattern of scales amongst a set of questionnaire items. However, it is also possible, although rarely applied, to perform confirmatory factor analysis in which a model of a factor–analytic structure is pre-specified and the purpose of further analysis is to examine how well the data fit this model (Fayers and Machin, 1998). One reason that this technique has not been widely applied in the field of patient-based outcomes is that investigators are rarely confident to specify a model to fit multiple questionnaire items in advance.

The contrast between development of instruments by formal methods from psychometrics, such as factor analysis and more informal methods involving patients more directly is illustrated by Juniper and colleagues (Juniper *et al.,* 1997). Adults with asthma completed a questionnaire with 152 items regarding QoL and asthma. Patients rated how frequent and how important each item was to them. The investigators reduced the items to a more manageable length by two methods, factor analysis and selecting items that had the greatest impact in terms of frequency and importance to patients. The former method resulted in a 36-item and the latter in a 32-item questionnaire. Only 20 items were common to both. The researchers note that both methods require elements of judgement and argue that the decision as to which method is better depends on investigators beliefs about the relative significance of importance to patients compared with statistical consistency in developing instruments.

Validity in relation to specific purposes

Although difficult, the range of observations needed to validate a measure of health-related QoL for a particular disease in the context of a trial is manageable. The issue of validity is far more complex if a measure is considered to serve a number of different purposes. The validity of an instrument can only ever be a judgement about how well an instrument measures something. In other words, does a measure of physical disability truly measure that construct? This issue is particularly salient if we consider some current issues surrounding generic and utility measures. These types of measures have tended to be used for a wide range of purposes (Revicki and Kaplan, 1993; Revicki *et al.,* 1995). It is more demanding to find clear evidence for the validity of each such purpose (Mulley, 1989). The issue is most complex where instruments are considered to be measures of the health status and health-related QoL of patients, measures of preferences and utilities of these same patients but also indicators of the social value of different health outcomes, and consequently of the social value of interventions (Nord *et al.,* 1993; Kind *et al.,* 1994; O'Hanlon *et al.,* 1994; Brooks *et al.,* 1996). It is important to examine the evidence for how well an instrument has been validated across such a range of purposes.

The Quality of Well-Being Scale (QWB) is an example of an instrument which has been put to such a wide range of uses. As with other generic instruments, the QWB has been used in a range of clinical areas such as AIDS, cystic fibrosis and arthritis (Kaplan *et al.,* 1989, 1992; Kaplan, 1993). However, it is also presented as encompassing a range of purposes as a measure: a measure of health status, of individual preferences and utilities with regard to health and, when combined with mortality, a measure of community preferences regarding benefits and health care priorities. When instruments are used for such

wide-ranging purposes, they require more extensive validation.

There are potentially three ways in which such measures need to be assessed for validity: as measures of (i) health status, (ii) of personal preferences and utilities, and (iii) of the social value. It is important to recognise that these are distinct constructs. Thus, Nord (1992) suggests that if measures are intended to provide assessments of the social value of interventions, as opposed to a measure of the individual utilities of patients, then one important component of validation would need to be **reflective equilibrium** whereby respondents are directly invited to consider and accept the implications in terms of resource allocation of judgements made about weightings. To the extent that public opinion accepted decisions about resource allocation of health services based on evidence of costs and benefits of treatments and where benefits are in part measured by utilities, according to Nord, this would be an indirect form of validation of utility measures used. An Australian study found very little public support for health policies that aimed to maximise health benefit without egalitarian considerations (Nord *et al.,* 1995).

Nord and colleagues has examined the validity of generic measures by more direct methods. In a series of surveys of Norwegian and Australian samples, Nord and colleagues (1993) examined respondents' ratings of specific health states for an instrument in comparison with the same health states judged by 'Person Trade Off'. The latter method involves respondents being asked to state what numbers of patients receiving one treatment are equivalent to a specific number receiving a second different treatment and is intended more directly to assess the social value of different health states. By comparing respondents' ratings from the two methods, it emerged that the instrument weighted by conventional rating scales produced much lower values for health states than did the Person Trade Off method, so that it appeared to produce lower social value for interventions. As Nord and colleagues argue, measures need to be examined separately to assess their validity as measures of individuals' utilities and as measures of social value. The two are distinct constructs and a measure may be valid as a measure of one but not the other and this is an area of continued methodological debate as to the validity of instruments for these different purposes (Carr-Hill, 1992; The EuroQol Group, 1992; Gafni and Birch, 1993; Dolan and Kind, 1996).

Summary

The apparently simple question as to whether an instrument measures what it purports to measure has to be considered by means of a range of different kinds of evidence including how content was determined, inspection of the content, and of patterns of relationships to other variables. Because no single set of observations is likely to determine validity and different kinds of evidence are needed, judgement of this property of an instrument in relation to a specific trial is not straightforward.

Responsiveness

Does the instrument detect changes over time that matter to patients?

For use in trials, it is essential that a health status questionnaire can detect important changes over time within individuals, that might reflect therapeutic effects (Kirshner and Guyatt, 1985; Kirshner, 1991). This section addresses sensitivity to change, or responsiveness. The latter term is preferable because sensitivity has a number of more general uses in epidemiology. As it is conceivable for an instrument to be both reliable and valid but not responsive, this dimension of a health status measure is increasingly essential to evaluate. Potential confusion is caused by the fact that in order to emphasise its importance, some authors treat it as an aspect of validity (Hays and Hadorn, 1992). Guyatt and colleagues (1989a) define responsiveness as the ability of an instrument to detect clinically important change. They provide illustrative evidence of the importance of this aspect of instruments with data from a controlled trial of chemotherapy for breast cancer. Four health status instruments considered to be validated were completed by women. However only one of the four instruments showed expected differences over time as well as providing valid evidence of women's' health status. Guyatt and colleagues (1987, 1992a,b) have emphasised an important distinction between **discriminative** instruments intended to be particularly valid in distinguishing between respondents at a point in time and **evaluative** instruments that need to be particularly sensitive to changes within individuals over time in the context of clinical trials. The question at the beginning of this section emphasises that patient-based outcome measures changes of importance **to patients**. Whilst responsiveness as such is not defined in terms of importance to patients, this would seem an important specification in relation to patient-based outcome measures.

TABLE 3 *Statistical methods of evaluating responsiveness*

Method	Summary of distinctve features
Correlation with other change scores (Meenan *et al.,* 1984)	Significant correlations with changes in other variables considered as evidence of responsiveness
Effect size (Kazis *et al.,* 1989)	Change score for an instrument is divided by standard deviation of baseline measure of instrument
Standardised response mean (SRM) (Liang *et al.,* 1990)	Change score for an instrument is divided by standard deviation of change score
Modified standardised response mean (Guyatt *et al.,* 1987b)	SRM as above except denominator is standard deviation of change score for individuals otherwise identified as stable
Relative efficiency (Liang *et al.,* 1985)	Square of the ratio of paired *t*-test for instrument relative to another instrument
Sensitivity and specificity (Deyo and Inui, 1984)	Transforms change scores into categorical data 'improved', 'stable' etc., and tests sensitivity and specificity of categories against independent evidence
Receiver operating characteristics (Deyo and Centor, 1986)	Plots sensitivity and specificity data as receiver operating characteristics

Rather like validity, there is no single agreed method of assessing or expressing an instrument's responsiveness and a variety of statistical approaches have been proposed *(Table 3)*. The literature on responsiveness is not as well developed as it is for reliability and validity. The various methods to assess responsiveness are now considered in turn.

Change scores

The simplest method to use is to calculate change scores for the instrument over time in a trial or longitudinal study and to examine the correlations of such change scores with changes in other available variables. For example, Meenan and colleagues (1984) examined the correlations of changes over time in a health status measure with changes in physiological measures in a trial of patients with arthritis. Correlations were significant and the health status measure considered responsive. This approach provides important evidence of whether a health status measure provides changes over time that are consistent with other available data. It does not provide a formal statistic of responsiveness.

Effect size

A number of methods, now discussed, have been proposed to provide quantitative expressions of the **magnitude** and **meaning** of health status changes. These same approaches may also be considered expressions of the responsiveness of health status instruments. Just as with reliability and validity, the estimates provided for responsiveness are strictly speaking confined to specific uses in particular populations and not an inherent property of the instrument.

One common form of standardised expression of responsiveness is the effect size. The basic approach to calculation of the effect size is to calculate the size of change on a measure that occurs to a group between assessments (for example before and after treatment), compared with the variability of scores of that measure (Kazis *et al.,* 1989). Most commonly this is calculated as the difference between mean scores at assessments, divided by the standard deviation of baseline scores. The effect size is then expressed in standardised units that permit comparisons between instruments (Lydick and Epstein, 1993; Jenkinson *et al.,* 1995a,b; Rutten van Molken *et al.,* 1995b). Effect size is more commonly used than methods such as standardised response mean (SRM) (below) because data are usually more readily available for baseline standard deviations in the scores of an instrument (Liang, 1995). Kazis and colleagues (1989), in their original discussion of the role of effect size in evaluating responsiveness, acknowledged that frequently the data from which effect sizes are calculated are not normally distributed. They propose that investigators, instead of using parametric statistics, consider using medians and interquartile ranges. However, there is little evidence of this suggestion being taken up.

It has been proposed that effect sizes can be translated into benchmarks for assessing the relative size of change; an effect size of 0.2 being

considered small, 0.5 as medium and 0.8 or greater as large (Cohen, 1977, Kazis *et al.,* 1989).

Standardised response mean

An alternative measure is the SRM. It only differs from an effect size in that the denominator is the standard deviation of change scores in the group in order to take account of variability in change rather than baseline scores (Liang *et al.,* 1990). Because the denominator in the SRM examines response variance in an instrument whereas the effect size does not, Katz and colleagues (1992) consider that the SRM approach is more informative.

Modified standardised response mean

A third method of providing a standardised expression of responsiveness is that of Guyatt and colleagues (1987b). As with effect sizes and SRMs, the numerator of this statistic is the mean change score for a group. In this case, the denominator is the standard deviation of change scores for individuals who are identified by other means as stable (Tuley *et al.,* 1991). This denominator provides an expression of the inherent variability of changes in an instrument, 'an intuitive estimate of background noise' (Liang, 1995). Unlike the two other expressions, this method requires independent evidence that patients are indeed stable, in the form of a transition question asked at follow-up. MacKenzie and colleagues (1986b) used a transition index and the modified SRM to test the responsiveness of the SIP.

Relative efficiency

Another approach is to compare the responsiveness of health status instruments when used in studies of treatments widely considered to be effective, so that it is very likely that significant changes actually occur. As applied by Liang and colleagues (1985) who developed this approach, the performance of different health status instruments is compared to a standard instrument amongst patients who are considered to have experienced substantial change. Thus they asked patients to complete a number of health status questionnaires before and after total joint replacement surgery. Health status questionnaires were considered most responsive that produced the largest paired *t*-test score for pre and post surgical assessments. Liang and colleagues (1985) produce a standardised version of the use of *t*-statistics (termed 'relative efficiency'), the square of the ratio of *t*-statistic for two instruments being compared. As noted earlier, much of the data in this field is non-parametric, so that *t* statistics are not appropriate and non-parametric forms of relative efficiency need to be used.

Sensitivity and specificity of change scores

Another approach to assessing responsiveness is to consider the change scores of a health status instrument as if they were a screening test to detect true change (Deyo and Inui, 1984). In other words, data for a health status measure are examined in terms of the sensitivity of change scores produced by an instrument (proportion of true changes detected) and specificity of change scores (proportion of individuals who are truly not changing, detected as stable). This method requires that investigators identify somewhat arbitrarily a specific change score that will be taken as of interest to examine, say an improvement of five points between observations. The sample is then divided according to whether or not they have reported five points improvement or not. Some external standard is also needed to determine 'true' change; commonly it is a consensus of the patient's and or clinician's retrospective judgement of change (Deyo and Inui, 1984). Essentially, the extent of agreement is then examined between change as defined in this hypothetical case as more than five points change and independent evidence of whether individuals have changed. The same analysis is then provided for specificity. Scores of less than five points are counted as 'unchanged'. Individuals scores are then examined to determine how much this classification of individuals as unchanged agrees with independent evidence that they have not changed.

This approach permits an assessment of the sensitivity and specificity of different amounts of change registered by an instrument. For example, an instrument which, for the sake of simplicity, has five possible scores (1–5), when applied on two occasions over time, may either produce identical scores on both occasions or register up to four points of change (for example a change from '5' to '1'). If one has an external standard of whether change truly occurred, such as a consensus of patient's and doctor's opinion, one can assess how sensitive and specific to the external evidence of change are changes of one point, two points etc. As sensitivity improves, specificity may deteriorate.

Receiver-operating characteristics

Deyo and Centor (1986) extend the principle of measuring the sensitivity and specificity of an instrument against an external criterion by suggesting that information be synthesised into receiver-operating characteristics. Like the simpler version just described in the previous section, this

method depends upon having an external gold-standard assessment of whether change has actually occurred. Their method plots the true positive rate (i.e. a true change has occurred) for an instrument against the false positive rate for all possible cut-off points (i.e. where change is taken successively as 'one point', 'two points' and so on). The most responsive instrument would have a plot where the true positive rate sharply increases whilst the false positive rate remains low. The greater the total area under a plotted curve from all cut-off points, the greater the instrument's responsiveness. It has been suggested that the plotting of an instrument's sensitivity and specificity with different cut-off points in this way represents the best way of finding the optimal cut-off point for an instrument (Deyo *et al.,* 1991). By 'cut-off' is meant whether a change of, say, two, rather than four or five points should be regarded as the minimal evidence of a 'real' change having occurred. The optimal cut-off point is identified as that which produces the highest sensitivity rate for the lowest specificity rate. Beurskens *et al* (1996) compared the responsiveness of four instruments (Oswestry Questionnaire, Roland Disability Questionnaire, main complaint scale and pain severity scale) within the setting of lower back pain, using receive–operator characteristics and global perceived effect as the external gold standard. The graphic presentation revealed that the Roland Disability Scale and the pain severity scale were the most responsive as those curves were closed to the upper left corner (the true positive rates rose sharply whilst the false-positive rate remained low) and had the greatest area under the plotted curve.

In general, these various methods express subtly different aspects of change scores produced by instruments. It is not surprising, therefore, that when several instruments are compared in terms of their responsiveness, somewhat different impressions can be formed of relative performance depending on which methods is used to assess responsiveness (Deyo and Centor, 1986; Fitzpatrick *et al.,* 1993a). Wright and Young (1997) found that the rank order of responsiveness of different patient-based outcome measures varied according to which of five different methods they used in a sample of patients before and after total hip replacement surgery. They note that there are no agreed external 'gold-standards' of extent of 'real' change against which to judge the competing expressions of responsiveness.

In the section below, the concept of 'minimum clinically important difference', which is a method aiding the interpretability of numerical change scores from patient-based outcome measures, is discussed in more detail. This approach may also be considered relevant to responsiveness in that this approach also defines meaningful minimum levels of change that instruments are capable of detecting (Juniper *et al.,* 1994).

Ceiling and floor effects

The previous section describes different statistical expressions of the responsiveness of an instrument. Here one of the main limitations on the responsiveness of an instrument is examined. The actual form of questionnaire items in an instrument may reduce the likelihood of further improvement or deterioration being recorded beyond a certain point. Put another way, the wording of questionnaire items does not make it possible to report most favourable or worst health states. The terms 'ceiling' and 'floor' effects are usually used to refer to the two forms of this problem. Such problems are quite difficult to detect but have been illustrated in research (Brazier *et al.,* 1993). A study administered the MOS-20 scale to patients in hospital at baseline and again 6 months later (Bindman *et al.,* 1990). At the follow-up survey respondents also completed a 'transition question' in which they assessed whether their health was better, the same or worse than at baseline assessment. A number of respondents who reported the worst possible scores for the MOS-20 at baseline reported further deterioration in their follow-up assessment in their answers to the transition question. It was clearly not possible for such respondents to report lower scores on the MOS-20 than at baseline.

Similarly, a series of patients with rheumatoid arthritis were assessed by means of the HAQ at baseline and 5 years later, with the follow-up questionnaire also including a transition question (Gardiner *et al.,* 1993). There was a general trend towards deterioration in HAQ scores across the sample which is expected with rheumatoid arthritis. However, the group who at baseline reported the worst HAQ score (i.e. most severe disability) showed significantly less deterioration than other groups over 5 years despite reporting the worst changes in their transition question at follow-up. Both Bindman and colleagues and Gardiner and colleagues interpret their studies in terms of the limited scope of the instruments to permit very ill respondents to report further deterioration because of floor effects. Essentially, more severe items are not available on the questionnaire. Similar observations have been made of ceiling effects in instruments, in that questionnaires appeared unable to detect improvements in

patients beyond a certain level (Ganiats *et al.,* 1992).

Distribution of baseline scores

The responsiveness of an instrument may also be influenced by the relationship of items in the instrument to the distribution of levels of difficulty or severity in the underlying construct. As a hypothetical example, it is possible to imagine an instrument designed to measure mobility where items mainly reflected 'easy' tasks; that is the majority of respondents could be expected to report no problem, for example, in walking a very short distance. Because most items in the scale reflect 'easy' items, a large amount of change could be produced (i.e. the patient reports change over the majority of items) even when only a small amount of real improvement had occurred. Stucki and colleagues (1995) show that the problem of the relationship of items to an underlying range of degrees of difficulty or seriousness is not entirely hypothetical. They provide evidence that many items from the physical ability scale of the SF-36 reflect intermediate rather than extremes of level of difficulty for patients undergoing total hip arthroplasty. Thus patients experiencing improvements at this intermediate level of physical difficulty can be expected to experience high levels of gain according to SF-36 at least in part because of the range of items. As Stucki and colleagues argue, this problem can arise from the ways in which scales are often developed, as described in earlier sections of this report, with emphasis upon high levels of agreement between items on a scale (internal reliability), rather than requiring items that reflect a full range of difficulty or severity of an underlying problem. We have already seen arguments against excessive reliance on inter-item agreement to develop instruments rehearsed by Kessler and Mroczeck (1995) in the context of reliability, above. Here it is possible to see problems arising from excessive emphasis upon internal reliability in the context of responsiveness.

Summary

The need for an instrument to be responsive to changes that are of importance to patients should be of evident importance in the context of clinical trials. Whilst there are no universally agreed methods for assessing this property, at a more general level all discussions require evidence of statistically significant change of some form from observations made at separate times and when there is good reason to think that changes have occurred that are of importance to patients.

Precision

How precise are the scores of the instrument?

This review is primarily concerned with the use of patient-based outcome measures in the context of clinical trials. Investigators will need to examine the pattern of responses to health status measures in a trial to determine whether there are clear and important differences between the arms of a trial. They therefore need to examine a number of aspects of candidate instruments' numerical properties which have not been clearly delineated in the literature, but which relate to the precision of distinctions made by an instrument. Testa and Simonson (1996) refer to this property as 'sensitivity':

> *'Although a measure may be responsive to changes in Q (quality of life), gradations in the metric of Z (the instrument) may not be adequate to reflect these changes. Sensitivity refers to the ability of the measurement to reflect true changes or differences in Q' (1996: 836).*

Stewart (1992) also refers to this property as 'sensitivity'. In particular, she refers to the number of distinctions an instrument makes; the fewer, the more insensitive it is likely to be. Kessler and Mroczek (1995) refer to this property as 'precision', which is probably less confusing since sensitivity has a number of other uses and meanings in this field. As Kessler and Mroczek argue, an instrument may have high reliability but low precision if it makes only a small number of crude distinctions with regard to a dimension of health. Thus at the extreme one instrument might distinguish with high reliability only between those who are healthy and those who are ill. For the purposes of a trial, such an instrument would not be useful because it is degrees of change within the category of 'unwell' that are likely to be needed to evaluate results of the arms of the trial.

There are a number of ways in which the issue of precision has been raised in relation to patient-based outcome measures. This is fairly disparate evidence and it is reviewed under a number of more specific headings.

Precision of response categories

One of the main influences on the precision of an instrument is the format of response categories; i.e. the form in which respondents are able to give their answers. At one extreme answers may be given by respondents in terms of very basic distinctions,

'yes' or 'no'. Binary response categories have the advantage of simplicity but there is evidence that they do not allow respondents to report degrees of difficulty or severity that they experience and consider important to distinguish (Donovan *et al.,* 1993). Many instruments therefore allow for gradations of response, most commonly in the form of a Likert set of response categories:

– strongly agree
– agree
– uncertain
– disagree
– strongly disagree

or some equivalent set of ordinally related items:

– very satisfied
– satisfied
– neither satisfied nor dissatisfied
– dissatisfied
– very dissatisfied

Alternatively, response categories may require that respondents choose between different options of how frequently a problem occurs.

There is some evidence that there is increased precision from using seven rather than five response categories. A sample of older indviduals with heart problems were assigned to questionnaires assessing satisfaction with various domains of life with either five or seven item response categories (Avis and Smith, 1994). The latter showed higher correlations with a criterion measure of QoL completed by respondents. However there is little evidence in the literature of increased precision beyond seven categories.

The main alternative to Likert format response categories is the visual analogue scale, which would appear to offer considerably more precision. Respondents can mark any point on a continuous line to represent their experience and in principal this offers an extensive range of response categories. However, the evidence is not strong that the apparent precision is meaningful (Nord, 1991). Guyatt and colleagues (1987a) compared the responsiveness of a health-related QoL measure for respiratory function, using alternate forms of a Likert and visual analogue scale. They found no significant advantage for the visual analogue scale. Similar results were found in a randomised trial setting, showing no advantage in responsiveness for visual analogue scales (Jaeschke *et al.,* 1990). An additional concern cited earlier is the somewhat lower acceptability of visual analogue scales as a task. Overall, firm empirical evidence of superiority of visual analogue scales over Likert scales is difficult to find (Remington *et al.,* 1979).

Precision of numerical values

To be of use in clinical trials, what patients report in health status measures is generally transformed into numerical values or codes that, on the one hand, most accurately reflect differences between individuals and changes within individuals over time and, on the other hand make possible statistical analysis of the size and importance of results. Clearly philosophical and epistemological issues can be raised about this process of assigning numerical values to subjective experience (Nordenfelt, 1994). These issues must be acknowledged but are beyond the scope of this review to address. Instead, we need to examine how the field has drawn upon psychometric, social scientific and statistical principles to produce pragmatically plausible numerical values as accurately as possible to capture subjective experiences that may in some way be related to health care interventions.

Two basically different methods of numerical scoring can be found amongst health status measures. On the one hand, the majority of instruments use somewhat arbitrary but common-sense based methods of simple ordinal values. For example, many instruments use Likert format response categories where degrees of agreement with a statement are given progressively lower values:

strongly agree = 1; agree = 2; neither agree nor disagree = 3; disagree = 4, strongly disagree = 5.

The direction of such values is entirely arbitrary, and can be reversed so that greater agreement is given higher numerical value.

It is worth noting that some instruments such as SF-36 recode numerical values so that items are expressed as percentages or proportions of the total scale score. To take a hypothetical example, an instrument may have six alternative responses for an assessment of pain, ranging in severity from, let us say, 'no pain at all' through to 'severe pain all of the time'. Instead of scoring responses '1', '2', '3' and so on, the scores may be transformed into percentages of a total: '17%', '33%', '50%'. Although this approach produces a range of values between 0 and 100, the simple and limited basis from which values are derived should be kept in mind. In particular, while it might appear that

an instrument has a high level of precision because scores are expressed as percentages, the range of actual possible values may still be quite small and scores are in no sense interval.

By contrast to such common-sense based methods of weighting are efforts directly to assess the relative severity or undesirability of different states. The SIP is an example of an instrument with a more sophisticated and more explicitly based weighting system. Once the questionnaire items for the instrument had been identified, a panel of patients, health professionals and pre-professional students used category scaling to assign weights to items by making judgements of the relative severity of dysfunction of items (Bergner *et al.,* 1976). To illustrate the impact of this weighting approach to questionnaire items, in the English version of the instrument, the most severe items in the body care and movement scale are 'I am in a restricted position all the time' (–124) and 'I do not have control of my bowels' (–124), whereas the least severe items are 'I dress myself but do so very slowly' (–043) and 'I am very clumsy' (–047). Separate weighting exercises on American and English versions by separate panels in the two language communities arrived at very similar weightings for items for the SIP (Patrick *et al.,* 1985). Other instruments that include such explicitly derived weighting systems include the Nottingham Health Profile (NHP), QWB and EQ-5D.

There are two particularly striking problems if the numerical values used in different patient-based outcomes are examined. On the one hand, many instruments use methods of scoring items that are deceptively simple. Although apparently simple, such scoring nevertheless may require strong assumptions; for example that the difference between the first and second responses is regarded as the same as the difference between the fourth and fifth response in a five-point Likert scale, if scores are analysed as interval scale scores.

On the other hand, the other most striking problem is that scoring methods that attempt directly to estimate the values of such response categories such as in the SIP by weighting systems, risk being deceptively precise. Their numerical exactness might lend pseudo-precision to an instrument. For investigators examining the numerical values of instruments, it is sensible to treat all scoring methods as weighted, differing only in how transparent weights are, and to look beyond superficial aspects of precision to examine how weightings have been derived and validated.

More pragmatically, it is appropriate to ask whether weighting systems make a difference (Björk and Roos, 1994). Sensitivity analysis may reveal that they make no significant difference to results. For example, Jenkinson and colleagues (1991) analysed patterns of change over time in health status for patients with rheumatoid arthritis by means of the FLP and NHP. Sensitivity to change as indicated by a battery of other clinical and laboratory measures was very similar, whether weighted or unweighted (items valued as '1' or '0') versions of the instruments were used. Other studies have similarly suggested that weighted scales may not improve upon the sensitivity of unweighted scales (O'Neill *et al.,* 1996).

The response format of a patient-based outcome measure to some extent determines the kinds of statistical tests that may be used on it. This is here considered an aspect of precision in the sense that many instruments contain items that are at best ordinal in form (i.e. questionnaire items where there is an implied rank to responses: 'very often', 'quite often' etc.) but not interval (i.e. where the interval between responses is of known value) or ratio (where there is a meaningful zero point). It might be argued that instruments that have only ordinal level measurement properties are capable of less precision (Haig *et al.,* 1986). Certainly, a review of the statistical properties of a series of health status scales published in the literature concluded that the majority of scales were presented and analysed as if based on interval-level when this property was not established (Coste *et al.,* 1995). Whilst it might be argued that an advantage of visual analogue scale over Likert format answers is that it would enable more extensive use of parametric statistics, this needs to be balanced against the lower acceptability of visual analogue scale techniques and the risk of pseudo-precision that this technique involves (Aaronson, 1989).

Mackenzie and Charlson (1986) reviewed trials employing ordinal scales in three medical journals over a 5-year period and found that many measures purporting to be ordinal were not. For example, values for the items of a scale were not truly hierarchical, so it was not clear whether lower numerical scores truly reflected worse underlying states.

As Streiner and Norman (1995) point out, there is a large and unresolved literature as to the propriety of using interval level statistics when it is unclear that there is a linear relationship of a measure to the underlying phenomenon. In practice, there may be many circumstances where cautious

assumption of interval properties with ordinally based data does not seriously mislead.

One quite practical illustration of the need for caution is in the calculation of sample sizes for trials using patient-based outcome measures as a primary end-point. Using the SF-36 as example, Julious and colleagues (1995) show that for those dimensions of SF-36 where the distribution of scores are highly skewed, sample size calculations are very different if parametric and non-parametric methods are used to estimate required sample size.

Distribution of items over true range

The items and scores of different instruments may vary in how well they capture the full underlying range of problems experienced by patients. It is not easy to examine the relationship between the distinctions made by a measuring instrument and the true distribution of actual experiences, for the obvious reason that one usually does not have access to the true distribution other than through one's measuring instrument. Nevertheless a number of arguments have been put forward that show that this is a real issue. Kessler and Mroczek (1995) have illustrated this problem by means of an instrument to measure psychological distress. They showed that it was possible to select short form versions of small numbers of items taken from a full set of 32 items measuring distress that, whilst all having the same reliability as the full 32-item scale and agreeing strongly with the total scale, differed markedly in ability to discriminate between distressed and not distressed individuals at different levels in an overall continuum of severity of psychological distress. One short form version was most discriminating at low levels of distress, and so on. A comparison that makes this point more intuitively understandable would be a range of intelligence tests, with, at one extreme a test that could distinguish the very cleverest as a category from all others who would be grouped together. At the opposite extreme, tests would sensitively distinguish those with very low intelligence from all others. The ideal test, whether of health or intelligence would have equal precision at every level.

Another illustration of the problematic relationship between items and the 'true' distribution of what is being measured is provided by Stucki and colleagues' (1996) analysis of SF-36 physical ability scores in patients undergoing total hip replacement surgery. They showed that many of the items of this scale represent moderate levels of difficulty for patients to perform (e.g. 'bending, kneeling or stooping'); by contrast, there are only a few items that almost everyone could do with no difficulty (e.g. 'bathing and dressing yourself') and only a few items that were difficult for the majority to perform (e.g. 'walking more than a mile'). A direct consequence of this is that patients passing a difficulty level in the middle of this scale of the SF-36 are more likely to have larger change scores than patients undergoing change at either the top or bottom of the range of difficulty of items, simply because of the larger number of items assessing moderate levels of difficulty. A similar set of observations about the 'maldistribution' of items of this scale of SF-36 was made by another group of investigators (Haley *et al.,* 1994). The most obvious consequences of this effect are two-fold: (i) the meaning of change scores for instruments may need to be interpreted in the knowledge of baseline scores of patients and (ii) instruments may need to ensure a more even distribution of items across the range of levels of severity or difficulty.

The distribution of items for the physical scale of SF-36 was examined by Stucki and colleagues (1996) by a variety of statistical methods including Rasch analysis (discussed on page 37) to address the issue of distribution. It should also be possible for investigators to inspect instruments at a more informal and intuitive level to consider whether there may be problems of the distribution of items in relation to the intended trial and patient group, to see whether particular levels of severity of illness are under-represented.

Ceiling and floor effects in relation to precision

The problem of ceiling and floor effects has already been considered in the context of responsiveness. They are mentioned again here because, essentially they may be viewed as problems arising from the precision and distribution of items in questionnaires. Studies were cited above in the context of responsiveness (Bindman *et al.,* 1990; Gardiner *et al.,* 1993) in which convincing evidence was found that some instruments did not allow patients with poor health status to report further deterioration. Questionnaires were found not to include items to capture the poorest levels of health. Potential solutions, depending on the overall format of the instrument, include adding a response category such as 'extremely poor' to questions and increasing the range of items, particularly addressing more severe experiences. Bindman and colleagues (1990) suggest adding transition questions which directly invite respondents to say whether they are worse or better than at a previous assessment. However, this is an unwieldy solution in terms of the formatting of questionnaires.

It has been argued that a commonly used instrument, the NHP, suffers from the opposite problem of having a ceiling effect. From population data, it was found that the modal response to the NHP was zero (i.e. no stated health problems) (Kind and Carr-Hill, 1987). However, the data were drawn from a survey of the general population most of whom were likely to be well, and ceiling effects need most urgently to be identified in patients with confirmed health problems. As Bindman and colleagues (1990) argue, ceiling effects are less of a concern generally because, in practice, researchers are less likely to search for improvements in health amongst those who already have excellent health.

Dimensionality and precision of scales

An important aspect of the precision of an instrument or of scales within it is the extent to which items clearly and precisely assess the one construct that is intended rather than unrelated and unintended aspects. For example, does a scale intended to assess depression actually include unintended items assessing symptoms of physical disease? Ideally, the scales of an instrument should be in this sense uni-dimensional. As we have seen in the context of reliability earlier, to examine the precision of a scale we need to look carefully at how it was developed. Detailed accounts of methodologies for assessing scales are beyond the scope of this report, but an understanding of basic principles will enable investigators to make informed choices between instruments based on different types of scales.

As was noted above in relation to reliability, the most common way to establish whether items in a questionnaire represent a scale with clear and precise content is by use of factor analytic techniques which identify whether items in an instrument load onto a smaller number of underlying dimensions. With scales based on Likert and similar principles, the emphasis is upon particular forms of statistical analysis such as factor analysis and tests of reliability, that demonstrate the internal consistency of items. Much of the discussion of internal reliability earlier in this report depends on this approach. Most of the scales which investigators will encounter in the context of clinical trials are likely to have been developed by means of factor analysis if any formal statistical approach was used in scale development.

A quite different technique – Thurstone scaling – has been used in the development of a minority of patient-based outcome measures. A form of this approach was used, for example in the development of the NHP. Essentially, samples are asked to judge lists of statements about health in terms of the degree of severity indicated, the task being achieved by means of paired comparisons, and sampling of comparisons so that every member of the panel does not have to make every possible comparison. Their rankings are used to give items their numerical value in a final instrument. This approach has attracted several criticisms. Kind and Carr-Hill (1987) argue that the dimensionality of the NHP is determined *a priori;* their analyses suggest that scales of the NHP overlap more than is desirable. Jenkinson (1994) argues that Thurstone scaling is designed for attitude measurement and inappropriate when factual or objective information such as regarding physical function is assessed. He also argues that principles of Thurstone scaling are broken if items do not reflect the full range of intensity of a trait, whereas the NHP appears to address only the more severe levels of subjective health (Kind and Carr-Hill, 1987; Brazier *et al.,* 1992). In general, Thurstone scaling has not been widely used as a method of scaling for patient-based outcome measures.

Another method of scaling, known as Guttman scaling, examines whether, in addition to being internally consistent, items are hierarchically related. Questionnaire items are tested for conformity to a model which requires that they assess increasing amounts or degrees of a trait or property, for example increasing difficulty in performing different daily tasks such as washing and eating. The hierarchical order of items means that, in a hypothetical example of a scale to assess ability to perform daily tasks, if the individual scores as having difficulty in performing one item, say, getting out of bed, then it can be assumed that the individual will have difficulty with all more difficult items, for example, getting around the house. One of the main areas where this approach has been tried is in rehabilitation medicine, where it has been widely believed that functions such as activities of daily living are both lost and recovered in a hierarchical sequence. To some extent, scales in this area have been shown to conform to the Guttman model (Spector *et al.,* 1987). The field of disability assessment has been the most promising for scales using this approach although the evidence is mixed (Williams, 1983). However, in health care more generally, it is uncommon for scales to have hierarchical properties according to Guttman scaling, for the simple reason that most problems addressed by patient-based outcome measures do not occur or are not experienced in a strictly hierarchical or strictly ordered manner.

Most recently, Rasch models have been used to assess the extent to which items in patient-based outcome measures are uni-dimensional, hierarchical and contain items that cover adequately the range of levels of the underlying construct (health, mobility etc.). Essentially, Rasch models test how well instruments conform to uni-dimensionality, hierarchy and interval location of items by examining patterns of individuals' performances on the range of items in a scale and patterns of items' difficulty or severity. Tennant and colleagues (1996) examined data from a population survey using the Health Assessment Questionnaire (HAQ) with Rasch methodology. On the positive side, especially for patients with rheumatoid arthritis, the HAQ appears to be both uni-dimensional and to have potential as a hierarchical measure. On the other hand, the results provide interesting evidence that HAQ scores may not test the full range of stages or levels of underlying disability. They also infer from the analyses that, for patients with osteo-arthritis, rather than rheumatoid arthritis, one item (ability to grip) does not contribute to an otherwise unidimensional assessment of disability. This is consistent with clinical evidence that items on grip will be less relevant to assessing disability in predominantly lower limb-affected osteo-arthritis.

Haley and colleagues (1994) used Rasch analyses to examine the physical functioning scale of the SF-36. They found evidence that the scale is uni-dimensional, hierarchical (i.e. knowledge of individuals' scores on any item will reliably predict scores for other items), and contains items over a full range of the underlying continuum of physical activity. They were able to examine these properties in patients across a wide range of clinical conditions and argue for their consistency. One problem that the analysis does identify is possible 'bunching' of items, so that extremes of low or high difficulty are under-represented. As cited earlier, this methodology was used by Stucki and colleague (1996) to examine the distribution of the items of SF-36 physical scale in patients undergoing total hip replacement surgery. They came to the same conclusion as Haley and colleagues; that Rasch analysis reveals lack of coverage at either end of the underlying spectrum.

Rasch analysis appears to offer a very useful way of examining the precision of scales as we have identified the term. In particular, it offers what appear to be more formal methods of addressing uni-dimensionality and range of coverage. There are practical problems because it requires very large sample sizes to be robust, possibly in excess of 1000 (Streiner and Norman, 1995). On the other hand, it is worth remembering that many, if not most instruments were not designed to have the hierarchical (Guttman-like) properties that Rasch methodology tests and indeed in the way that they are used in trials, most scales are not required to have this property. There is also a rather strong assumption required of Rasch methodology, that while items differ in difficulty (i.e. what point in the continuum of, say, level of disability they are assessing), they are considered similar in discriminating ability (let us say, to distinguish 'disability' from 'non-disability') (Streiner and Norman, 1995). The assumptions tested by Rasch models are different from, say, Likert scaling (van Alphen *et al.,* 1994). To date, few instruments have been developed with the intention explicitly to conform to the demanding requirements of measurement required by Rasch analysis, so that it remains to be seen whether it is a useful method of selecting between instruments. However, for our purposes, it is important that users consider the nature of the evidence for the precision of scales. Formal methods can provide statistical evidence of this property. They provide the most precise evidence for what may also be considered informally and qualitatively by inspection of the content of scales, namely the range and uni-dimensionality of items contained therein.

Bias in the assessment of outcome

Randomised allocation is the optimal design of clinical trial for reducing risks of various forms of bias. One way of summarising the thrust of the literature on patient-based outcome measures generally, and of this report specifically, is the need to reduce random error in outcome assessment by means of greater validity, reliability, precision and related efforts. It is perhaps remarkable that less attention has been given to the equally important threat to trials arising from systematic bias in patient-based outcomes (Bouchet *et al.,* 1996). To some extent, systematic forms of bias that might influence health status scores are addressed by more general aspects of study design, for example by making assessments wherever possible blind to intervention. Of course, many trials cannot achieve this aspect of trial design. In many areas of health care, research the patient inevitably knows which arm of a trial he or she has received. It is reasonable, therefore, to ask whether instruments may differ in proneness to systematic bias arising from patients not being blinded. If instruments require that they are personally administered, there are additional risks of more subtle systematic bias if interviewers cannot be blinded to patients' assignment. It must be assumed that the risk of such

biases is greater, the further removed a trial is from a drug trial, so that differences in the processes of care between arms of a trial have greater chances of influencing patients' judgements of outcomes.

The social psychological literature on such forms of bias is quite extensive, particularly with regard to phenomena such as halo effects, social desirability effects and so on. There is evidence from qualitative research based on recordings of subjects talking about completing the NHP, that a considerable amount of cognitive work by patients precedes them selecting response categories (Donovan *et al.,* 1993). Respondents, for example, attempt to work out what investigators' intentions are in asking particular questions. However, there is no research evidence that we have found that considers whether different patient-based outcome measures might be more or less prone to cognitive effects that could bias results.

Other sources of potential bias have been examined in relation to health status questionnaires (Anonymous, 1995). Item bias occurs when background variables such as the respondent's gender or age, affect the response to items in the questionnaire. In comparing groups, item bias analysis tests the influence of variables, such as age, sex or race on patterns of responses and examines whether the possible differences between groups is correctly shown in the score (Groenvold *et al.,* 1995). A QoL questionnaire used with breast cancer patients was analysed and item bias was found in three out of nine dimensions due to age and other factors (Groenvold *et al.,* 1995).

A simple form of bias was identified in a study of outcomes of care for rheumatoid arthritis (Jenkinson *et al.,* 1993). Patients with rheumatoid arthritis were asked either as inpatients or outpatients to complete several health status measures. All were followed up and reassessed with the same measures in outpatient clinics 3 months later. Those who were inpatients at baseline showed very substantial improvements on the mobility scale of the FLP but not on the equivalent scale of the NHP. Outpatients showed little improvement by either measure. Much of the improvement on the FLP scale amongst those who were initially inpatients was attributed to the fact that the FLP, but not the NHP, produced more severe scores for anyone confined to a bed *per se* regardless of health status. Moving out of hospital confinement alone produced substantial changes in one, but not the other instrument. In a trial where randomisation was between, say hospital and outpatient or ambulatory management, it seems likely that one instrument would have much more potential for systematic bias, by 'exaggerating' the degree of improvement of individuals leaving hospital care. Another potential source of bias is the influence of psychological mood upon patient-based outcome measures. There is a range of evidence that factors such as depressed mood have a disproportionate influence upon patterns of answers to health status questionnaires (Spiegel *et al.,* 1988; Sensky and Catalan, 1992). Indeed in the MOS study, individuals with confirmed depression had amongst the poorest general health status scores of any chronically ill group studied (Wells *et al.,* 1989). Disentangling the reasons for such patterns is not easy. Depression may be associated with poor physical health for a variety of reasons (Brooks *et al.,* 1990). However, it is also possible that such patterns reflect cognitive distortion. In a randomised trial design, such effects may not necessarily have important consequences. However, unrecognised depression may distort evidence of overall effectiveness of treatments across dimensions of health status. Some instruments have been shown to be relatively immune to such effects; the HAQ, for example, seems relatively unaltered by depressed mood (Peck *et al.,* 1989).

Again, informal inspection of the content of instruments is as likely to identify the kinds of gross systematic bias just illustrated. More generally, the field has tended to address this issue more by attempting to reduce random error in patient-based outcome measures.

Patient-based outcome measures may therefore vary in how clearly and precisely the numerical values generated by measures relate to underlying distributions of patients' experiences of health status. Investigators need to consider (i) precision of response categories, (ii) precision of numerical values, (iii) distribution of items over true range, (iv) ceiling and floor effects, (v) precision of scales and (vi) sources of potential bias in the scoring of instruments. The degree of precision required of a patient-based outcome measure will depend on other aspects of trial design such as sample size, and also on the differences expected to be found between arms of the trial. However, investigators need to have some sense of the meaning of the scores that will be generated by instruments that they intend to use in a trial and precision is a component of meaning.

Summary

Overall, we are here concerned with how precise are the distinctions made by an instrument, with at

one extreme instruments that make very few rather gross distinctions between levels of health and illness and, at the other extreme, instruments that make many more specific distinctions. Given that clinical trials are frequently concerned with looking for difficult-to-detect differences between treatments, it might appear that the capacity to make numerous distinctions is in itself desirable. However, the literature has suggested a number of ways in which this would be misguided and not reflect accurate precision.

Interpretability

How interpretable are the scores of an instrument?

The issue of the interpretability of scores has only recently begun to receive attention in the literature on patient-based outcome measures. It has often been commented that patient-based outcome measures lack the interpretability that other measures, for example blood pressure, blood sugar levels or erythrocyte sedimentation rate, have for clinicians (Deyo and Patrick, 1989; Greenfield and Nelson, 1992). To some extent, this may be due to lack of familiarity with use. Researchers have also begun to make efforts to make scores more interpretable (Testa and Simonson, 1996). One method used in a trial of antihypertensives was to calibrate change scores on QoL instruments with the changes for the same instruments that have been found with major life events such as loss of a job (Testa *et al.*, 1993). In this way, health status scores could be related to other human experiences that have clear and intuitive meaning.

Another approach to interpreting results is to identify a plausible range within which a minimal clinically important difference (MCID) falls (Jaeschke *et al.*, 1989, 1991; Juniper *et al.*, 1994). Jaeschke *et al.* (1989) define a MCID as 'the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient's management' (1989:408). They examined this concept in relation to patients completing at baseline and follow-up either the Chronic Respiratory Questionnaire in a drug trial for asthma or the Chronic Heart Failure Questionnaire in a drug trial for patients with heart failure. Changes between baseline and follow-up were examined in relation to their bench-mark for a MCID, which was the patient's follow-up assessment in a transition item of whether they were worse, better or the same compared with baseline assessment. They showed that a mean change of 0.5 for a seven-point scale was the minimal change amongst patients reporting a change. Other methods of understanding clinically important changes require the selection of other external benchmarks such as the global judgement of the clinician or laboratory tests or reference to distribution-based interpretations, such as using effect size (Lydick and Epstein, 1993; Deyo and Patrick, 1995).

A different approach to interpretability can be considered if representative data are available from the general population with which to compare scores obtained in a trial. In practice only in the case of a few widely used instruments like SF-36 are such 'normative' data available against which to compare results (Jenkinson *et al.*, 1996). An extension of the logic of using more representative population data is to normalise or standardise the scores for an instrument used in a trial to scores based on those observed for the population as a whole, essentially by relating individuals' scores to the mean and standard deviation of the population as a whole. In this way, units of measurement that otherwise have no inherent meaning could now be transformed to identify a change in a clinical trial sample of, for example, one and a half standard deviations from the population mean (Streiner and Norman, 1995).

Summary

Interpretability is concerned with how meaningful are the scores from an instrument To date, it is not possible to compare patient-based outcome measures in terms of how interpretable developers have managed to make their instruments, although clearly those instruments that are more regularly included in trials and population studies will come to be more widely known and more familiar by use (Greenfield and Nelson, 1992).

Acceptability

Is the instrument acceptable to patients?

It is essential that instruments be acceptable to patients. This is clearly desirable to minimise avoidable distress to patients already coping with health problems. It is also essential in order to obtain high response rates to questionnaires to make results of trials more easy to interpret, more generalisable and less prone to bias from non-response. The acceptability of patient-based outcome measures has far less frequently been

examined than issues such as reliability and validity, and there is less consensus as to what constitutes acceptability. For Selby and Robertson (1987), acceptability is 'a description of the speed of completion of the questionnaire and the proportion of patients who find it difficult, impossible or unacceptable for any reason' (1987:528). Ware and colleagues (1981: 622) subsume these properties under practicality: 'An important aspect of practicality is respondent burden, indicators of which include refusal rates, rates of missing responses, and administration time'. Pragmatically, trialists using patient-based outcome measures are concerned with the end result; whether they obtain as complete data from patients as possible. Methods for increasing completion rates have been addressed in a number of reviews (Yancik and Yate, 1986; Aaronson, 1991; Sadura *et al.,* 1992). Others have considered how to analyse missing data (Fayers and Jones, 1983; Zwinderman, 1990). However, we need to consider the different components of acceptability in turn to identify sources of missing data.

Reasons for non-completion

Patients may either not return a whole assessment or may omit some items in the assessment. If patients either do not attempt to complete an instrument at all or omit particular items frequently, this is potentially a sign that a questionnaire is difficult to understand, distressing, or in some other way unacceptable. It may also be evidence of poor validity of an instrument if the non-response rate is high. However, there may be other reasons for non-completion such as the method of delivery of the questionnaire. Patients may not receive a mailed questionnaire in the first place or may not have a telephone in order to be contacted in this way. Patients may also be unable to complete questionnaires because of their health status or other disabilities, particularly cognitive or visual (Medical Research Council, 1995). In practice, determining the role that an instrument has on completion rates compared to other factors is not easy.

It is beyond the scope of this report to consider broader issues of survey methodology. However, it is important to be aware of the evidence that how a questionnaire is administered can influence response rates regardless of content. Postal surveys are more often used because they are cheaper than alternatives. However, they tend to have lower response rates than personally administered or telephone interviews. It has been argued that, if careful attention is paid to methodology, general postal surveys can expect to achieve 75–80% response rates and a variety of extra steps may be used to increase this level (Dillman, 1978; de Vaus, 1986; Oppenheim, 1992). Surveys using patient-based outcomes are subject to the same effects with postal methods of data collection achieving somewhat lower response rates than other methods (Sullivan *et al.,* 1995; Weinberger *et al.,* 1996).

It may be noted that there is also an unresolved debate in the general survey literature as to whether the method of administration can influence the content of answers to a questionnaire (Bremer and McCauley, 1986; Anderson *et al.,* 1986; Chambers *et al.,* 1987). In the general survey literature, there is substantial evidence that respondents give more favourable reports about aspects of their well-being when personally interviewed than they provide in a self completed questionnaire (Schwarz and Strack, 1991). Topics of a particularly sensitive nature are considered particularly prone to effects of method of data gathering, but the evidence is inconsistent as to whether mailed questionnaire or interview produce more accurate information (Wiklund *et al.,* 1990; Korner Bitensky *et al.,* 1994). Cook and colleagues (1993) showed the significance of this factor in patient-based outcome measures; patients reported more health-related QoL problems on a self completed questionnaire than when personally interviewed.

More general features of the layout, appearance and legibility of a questionnaire are thought to have a strong influence on acceptability. Some instruments such as the COOP Charts have deliberately included extremely simple and short forms of wording of questions together with pictorial representations to add to ease and acceptability of use (Hughes *et al.,* 1995). A rare experimental study to test the benefit of pictorial representation in a QoL study showed that cartoon figures to depict degrees of illness severity improved test–retest reliability compared with responses to conventional formatting (Hadorn *et al.,* 1992).

The health status of respondents can influence the likelihood of completing a questionnaire. Hopwood and colleagues (1994) provide evidence that, in a sample of patients with lung cancer, completion rates of a health status questionnaire were 92% amongst patients independently assessed as in the most favourable health state but 31% amongst those in the poorest health state. Poorer visual function has also been shown to be an influence on non-response in health status surveys (Sullivan *et al.,* 1995). There is conflicting

evidence of the extent to which older individuals have difficulty in completing health status questionnaires (Brazier *et al.,* 1992; Lyons *et al.,* 1994; Hayes *et al.,* 1995; Hill *et al.,* 1996). It is, however, the influence of health status that is the greatest concern, especially in the context of a clinical trial where loss to follow-up of those with poorer ill-health may create important biases in results. In practice, effects of characteristics of the patient group such as health status may be difficult to disentangle from those due to the acceptability of the questionnaire.

There is only limited evidence available comparing the response rates of different health status instruments, rather than the method of their administration. In a series of older patients who had undergone total hip replacement surgery, higher completion rates were obtained from a 12-item condition-specific questionnaire compared with a longer generic instrument, the SF-36 (Dawson *et al.,* 1996b).

Another form of evidence is the differential responses to different subject matters in surveys of health status. Guyatt and colleagues (1993a) found that a sample of elderly respondents were somewhat more likely to complete the section of a questionnaire concerned with physical compared with emotional items, suggesting differential acceptability of topics depending on how personal they were. By contrast, in a qualitative study of patients with small cell lung cancer (Bernhard *et al.,* 1995) it was reported that patients found questions about their psychological well-being more tolerable than questions about tumour-related symptoms.

Time to complete

It is often assumed that one aspect or determinant of the acceptability of a questionnaire is its length; the longer it takes to complete, the less acceptable is the instrument (Ware, 1984). Many instruments are published with claims by those who have developed the instrument about the length of time required to complete it. Far less commonly is this property independently assessed or instruments' time to complete measured comparatively. Amongst instruments requiring the least time to complete are the self-completed COOP charts which have been estimated to take 2–3 minutes (Nelson *et al.,* 1990). Similarly, Wolfe and colleagues (1988) directly assessed the mean length of time required to complete one of the most commonly used of instruments for arthritis – the HAQ – 3 minutes. Most health status instruments are longer than these two examples and probably require more time to complete. Aaronson and colleagues (1993) directly measured time to complete the EORTC QLQ-C30 for a sample on two separate occasions, before and during active treatment (12 and 11 minutes, respectively). The time required may depend upon the characteristics of respondents. Guyatt and colleagues (1993a) estimated that the total length of time, including instructions and eliciting of patient-specific information for the Geriatric Quality of Life Questionnaire was 30 minutes.

A smaller number of studies have examined comparatively the time required for various instruments or methods of administration. Weinberger and colleagues (1996) assessed the time required for SF-36 to be completed by two different methods of administration; self completed the instrument required 12.7 minutes compared with 9.6 minutes for face-to-face interviews. In a elderly group of patients, the SF-36 took 14 minutes by personal interview and 10.2 minutes by telephone administration (Weinberger *et al.,* 1994). Read and colleagues (1987) compared the time to administer of the General Health Rating Index, the QWB and the SIP, which required 11.4, 18.2 and 22.4 minutes, respectively. Generally such evidence is not available. In a comparative study of health status measures of outcomes, Bombardier and colleagues (1991) estimated that the HAQ required 5 minutes to complete, compared with three different utility measures that required administration by interview and between 30 and 60 minutes to complete. The above list of studies are unusual and there is no reliable and objective estimate of the time required for many instruments. This may be a problem because developers of instruments may be over-optimistic in their estimates.

The format of a patient-based assessment can also influence acceptability. At one extreme some tasks requiring respondents to derive utilities can be both distressing and difficult to comprehend (O'Hanlon *et al.,* 1994). Evidence of a less severe form of difficulty is provided by Guyatt and colleagues (1987a) who compared the measurement properties of Likert and visual analogue forms of response categories to a health-related QoL instrument. In explaining the two forms of task to patients they found that patients viewed visual analogue scales as harder to understand. In specific terms, they report that it took up to twice as long to explain.

Shorter forms

It is increasingly argued that, if there are no or minimal costs in terms of validity, responsiveness and other key components of instruments, then

instruments should be reduced in terms of length and number of items in order to increase acceptability (Burisch, 1984). Some comparative studies have shown no loss of responsiveness when such shorter instruments are used (Fitzpatrick *et al.,* 1989; Katz *et al.,* 1992). Thus, the SF-12 has emerged as a shorter version of the SF-36, considered to require only 2 minutes or less to complete whilst reproducing more than 90% of the variance in SF-36 scores in the general population (Ware *et al.,* 1996). Similarly the SIP has undergone a number of attempts to reduce it from its full 136-item version (McDowell and Newell, 1996).

Whilst there are good reasons in particular circumstances to prefer shortened versions of instruments, attention is needed to how an instrument has been shortened. A recent structured review of 42 studies intended to shorten longer original measures found that the majority used statistical methods alone to achieve this objective, typically relying on correlations of shorter with longer versions in the same data, or methods to maximise internal consistency of the shorter version (Cronbach's alpha) (Coste *et al.,* 1997). The authors argue that, whilst there are obvious advantages in shortening a well developed and widely validated longer instrument, there are also methodological pitfalls. In particular, selection of items on the basis of internal consistency will further narrow the scale. They argue that the psychometric properties of the short version need to be examined as if it is a new instrument. Properties such as precision, as discussed in an earlier section of this review may also be jeopardised by an instrument with fewer items.

Direct assessment of acceptability

It is preferable directly to assess patients' views about a new questionnaire. Sprangers and colleagues (1993) argue that patients' views should be obtained at the pre-testing phase prior to formal tests for reliability etc., by means of a structured interview in which they are asked whether they found any questionnaire items difficult annoying or distressing or whether issues were omitted. When the EORTC QLQ-C30 was assessed in this way, 10% of patients reported that one or more items were confusing or difficult to answer and less than 3% that an item was upsetting, whilst more generally patients welcomed the opportunity to report their experiences (Aaronson *et al.,* 1993). Another formal evaluation of acceptability of a questionnaire found 89% enjoyed the task of completing the COOP instrument and 97% reported understanding the questions (Nelson *et al.,* 1990).

Weinberger and colleagues (1996) directly asked patients for their preferences for different forms of administration of the SF-36. Far more positive preference was expressed for face-to-face interview compared with either self complete or telephone based administration.

Not all studies of patient acceptability of instruments are positive. In a qualitative study of patients' views of a QoL assessment that included an early form of EORTC questionnaire, patients complained about length, difficulties in understanding the format of the questionnaire and possible risks that their answers would influence subsequent treatment decisions (Bernhard *et al.,* 1995).

In general, users should expect to see evidence of acceptability being examined at the design stage. Subsequently, the most direct and easy to assess evidence is the length and response rates of questionnaires.

Translation and cultural applicability

One basic way in which a questionnaire may fail to be acceptable is if it is expressed in a language unfamiliar to respondents. This issue has received a large amount of attention in recent literature on patient-based outcomes, mainly because of the increasing need for clinical trials incorporating QoL measures to be conducted on a multi-national basis, especially in Europe (Kuyken *et al.,* 1994; Orley and Kuyken, 1994; Shumaker and Berzon, 1995). As a result, there are quite elaborate guidelines available intended to ensure high standards of translation of questionnaires (Bullinger *et al.,* 1995; Leplege and Verdier, 1995). Amongst procedures to improve translation, according to such guidelines, are: use of several independent translations that are compared; back-translation; testing of the acceptability of translations to respondents. Less attention has been paid to cultural and linguistic variations within national boundaries, but it would seem that similar principles could be applied to increase cultural applicability. Presently, few patient-based outcome measures have been translated into the languages of ethnic minorities in the UK.

An important issue is whether rigorous translation can by itself establish the appropriateness of an instrument to a new cultural context from the one in which it was developed. Such methods may not establish whether subjective experiences differ in terms of salience from one culture to another, or indeed may fail to identify concerns and experiences not anticipated in the culture in which an

instrument was first developed (Hunt, 1998). In this sense even the most thorough observation of translation procedures cannot alone establish the validity of an instrument in a new culture. An unusual solution that attempts to overcome the cultural specificity of questionnaires is the WHOQOL Group's (1998) development of the World Health Organization Quality of Life Assessment (WHOQOL). Instead of the usual practice in which a questionnaire is developed in one culture and then translated into the languages of other cultures, in this case concepts and questionnaire items were developed in 15 different field centres around the world, including developing as well as developed countries. Initial results have appeared regarding basic aspects of reliability and validity (WHOQOL, 1998). Further research will be required to examine the value of this 100-item questionnaire.

Summary

Evidence is required that an instrument is acceptable to patients. The simplest and most direct form of such evidence is that it has consistently been associated with high response rates. Early on in the development of an instrument, this property may have been more directly tested by eliciting views of patients about the instrument.

Feasibility

Is the instrument easy to administer and process?

In addition to patient burden and acceptability, it is important to evaluate the impact of different patient-based outcome measures upon staff and researchers in collecting and processing information (Aaronson, 1992; Lansky *et al.,* 1992; Erickson *et al.,* 1995). Data from patients for clinical trials are often gathered in the context of regular clinical patient care and excessive burden to staff may jeopardise trial conduct and disrupt clinical care. An obvious example is the additional staff effort and costs involved in personally administering questionnaires over postal delivery. To a lesser extent, the length and complexity of instrument are an additional component. Certainly it may require additional staff time to assist and explain how more complex questionnaires are to be filled out by patients. The simplest of instruments such as the nine-item COOP charts require a minimum of time and effort to process (Nelson *et al.,* 1990). Their brevity (one item per domain) and pictorial representation mean that they require less staff supervision than most alternatives. A related component of feasibility is time required to train staff to use an instrument, with questionnaires designed for self completion imposing the least burden in this respect. Where instruments do require interviewer administration, training needs can vary according to the complexity of the tasks. Read and colleagues (1987) compared the training times required for three health status instruments and found that they varied from 1 to 2 hours for the easiest to 1 to 2 weeks for the most complex instrument. Utility measures which involve respondents making complex judgements under unusual experimental conditions almost invariably require highly trained staff (Feeny and Torrance, 1989).

It is sometimes thought that more complex scoring systems reduce feasibility compared to simple scores. However, with computer programmes universally used to process such data, this element is unlikely to be a major component of burden to staff. Far more likely to require time to process are the measurement of physical marks put by patients onto visual analogue scales which have directly to be measured in terms of distance from origin.

Above all, with both acceptability and feasibility, as with other dimensions we have examined, these should not be considered entirely fixed properties of instruments. To some extent, both the content and appearance of instruments can be improved to enhance response rates. Probably more importantly, as Bernard and colleagues (1995) argued in their qualitative study of the use of health status measures, staff attitudes and acceptance of the value of patient-based outcome measures can make a substantial difference to ultimate acceptability by patients.

Summary

The time and resources required to collect, process and analyse a patient-based outcome measure are not often independently reported so that evidence may not be readily available. A judgement of this aspect of an instrument has to be made in the context of clinical trials given that this will be but one component of burden on participants that will determine the overall viability of a trial and therefore the quality of its final results.

Chapter 4
Conclusions

The rapid expansion of efforts to assess outcomes of health care from the patient's perspective has resulted in hundreds of instruments that have in common that they purport to provide standardised assessments of matters of importance to patients such as functional status, subjective health and broader aspects of health-related QoL. Seven major types of instrument can be distinguished: disease-specific, site or region-specific, dimension-specific, generic, global or summary, individualised, and utility. These distinctions between types should not be viewed as rigid since instruments can have properties associated with more than one kind. Given that the vast majority of such instruments are candidates for inclusion in trials, investigators facing the need to select an instrument or instruments to include for any specific trial have quite a daunting decision.

There are substantial areas of uncertainty and dispute regarding outcome measurement. Over a number of issues, gaps and limitations of concepts and measurement have been acknowledged in the literature. This review has built on and attempted to integrate previous efforts to identify desirable properties of patient-based outcome measures.

It is very encouraging that authors from three disciplines of social science, economics and statistics can agree to this document; this is itself an important step in progress to define the field. Broad assent to the principles of the review was also obtained from a wide range of disciplines and expertise relevant to health technology assessment and health services research: comments on a draft were sought from those with expertise in clinical medicine and clinical trials, health economics, health service research, psychology, sociology, statistics. Every effort was made to respond to and integrate expert advisors' suggestions. We feel that the resulting document presents views based on substantial consensus about issues.

Despite clear limitations in the evidence available to date, it is possible to conclude that there are eight criteria that can provide an explicit framework for decisions about selection of patient-based outcome measures in trials. In determining how best to assess outcomes from the patient's perspective in the context of a clinical trial, investigators need to consider candidate patient-based outcome measures in terms of appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability and feasibility.

Chapter 5

Recommendations

For trialists selecting a patient-based outcome measure

We recommend that, on as explicitly stated grounds as possible, and making use of available evidence about instruments, outcome measures for clinical trials should be chosen by evaluating evidence about instruments in relation to the following eight criteria: appropriateness, reliability, validity, responsiveness, precision, interpretability, acceptability and feasibility. Although underlying issues have been widely discussed, three of our criteria, appropriateness, precision and interpretability, are not always included in lists of desirable properties of instruments. The remaining five criteria are widely cited and identified in the same or similar terminology as in this review.

The selection of instruments on the basis of our criteria cannot, given the present state of the field, be a straightforward or mechanical one. This is partly because there is only a moderate level of consensus about what exactly is meant by some criteria. The literature does not provide unambiguous definitions and advice regarding the issues we have reviewed. The evidence for any given instrument will be partial and complex to assimilate. Above all, the criteria themselves cannot be weighted or prioritised given the current state of knowledge.

Investigators need to think of the desirable properties of outcome measures for a specific use in a specific trial question. Instruments do not have properties of being reliable, valid and so on in some universal sense; they are properties in relation to a specific use. This makes selection of instruments a complex process. Investigators need to select outcomes appropriate to the question addressed by a trial. Ideally each instrument is optimally appropriate, valid, reliable and so on, although, in reality, trials may include combinations of outcome measures that together have optimal measurement properties. There are costs as well as benefits to be considered of following the advice sometimes offered to include a combination of generic and disease-specific measures.

Given the incomplete and complex state of knowledge in this field, it may be advantageous for investigators setting up trials to involve those with expertise in outcomes in trial design and analysis.

For developers of patient-based outcome measures

To encourage more appropriate use of outcome measures, those who develop such instruments need to provide as clear evidence as possible of the available evidence of new instruments in terms of the eight criteria emphasised by this review. Standards for documentation of patient-based outcome measures will improve. These developments will make the task in selecting outcome measures for trials much more evidence-based.

Future research

In almost all areas reviewed there are substantial gaps in knowledge and understanding of how best to capture patients' perceptions of illness and outcomes of interventions within clinical trials. There is therefore a strong case for further methodological research in relation to patient-based outcome measures. To facilitate appropriate selection of instruments for clinical trials, two kinds of further research in particular are needed. Firstly, in trials and observational studies, the performance of patient-based outcome measures should be directly compared. There are still too few 'head-on' comparisons of different types of measures completed by the same patients within a trial, especially with regard to the issue of responsiveness. More such studies are needed either in the form of additional methodological components of major clinical trials or as methodological investigations in their own right. It will then be possible to address questions such as whether disease-specific, generic or other kinds of instruments are more responsive in various clinical contexts. Secondly, researchers and clinicians in specific areas, oncology, rheumatology, psychiatry and so on, should carry out assessments of evidence for the comparative performance generally of the more widely used of outcome measures in their field. This process has begun to happen in some specialties and publication of such consensus views would

further promote awareness of the role of patient-based outcomes in clinical trials.

By identifying a set of criteria and making some attempt to be more explicit about their meaning, this review is intended to progress the appropriate use of such methods in order to facilitate the conduct of clinical trials taking full account of patients' judgements about their health and health care.

Acknowledgements

This study was supported by the NHS R&D Executive's Health Technology Assessment Programme. We are indebted to the referees for their perseverance in reading the report and the quality of their comments.

The following individuals are thanked for their generous input of time and effort in commenting upon an earlier draft of the document: John Brazier, Peter Fayers, Jeremy Hobart, Penny Hopwood, Crispin Jenkinson, Paul Kind, Andrew Long, Hannah McGee, Derick Wade, David Wilkin. Every effort has been made to take account of the very extensive and helpful feedback provided by this group of advisors, although the four named authors of the document alone remain responsible for the final draft submitted to the NHS HTA Programme. Thanks are also due to Katherine Johnston for her invaluable help in relation to bibliographic searches and logistics of the review.

The first author (Ray Fitzpatrick) took responsibility for producing drafts of the text which were revised in the light of the comments of the other three authors and of the external experts listed above.

References

Aaronson NK (1989). Quality of life assessment in clinical trials: methodologic issues. *Control Clin* Trials;**10**:195S–208S.

Aaronson NK (1991). Methodologic issues in assessing the quality of life of cancer patients. *Cancer*;**67**:844–50.

Aaronson NK (1992). Assessing the quality of life of patients in cancer clinical trials: common problems and common sense solutions. *Eur J Cancer*;**28A**:1304–7.

Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, *et al* (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*;**85**:365–76.

Ahmad WI, Kernohan EE, Baker MR (1989). Influence of ethnicity and unemployment on the perceived health of a sample of general practice attenders. *Community Med*;**11**:148–56.

Albrecht G (1994). Subjective health assessment. In: Measuring health and medical outcomes (Jenkinson C, editor). London: University College London Press. p. 7–26.

Anderson JP, Bush JW, Berry CC (1986). Classifying function for health outcome and quality-of-life evaluation. Self-versus interviewer modes. *Med Care*;**24**:454–69.

Anderson JS, Sullivan F, Usherwood TP (1990). The Medical Outcomes Study Instrument (MOSI) – use of a new health status measure in Britain. *Fam Pract;***7**:205–18.

Anonymous (1991a). Quality of life [editorial]. *Lancet;***338**:350–1.

Anonymous (1991b). Recognising disability [editorial]. *Lancet;***338**:154–5.

Anonymous (1995). Choosing questions. Introduction. *Med Care*;**33**:AS106–8.

Applegate WB, Pressel S, Wittes J, Luhr J, Shekelle RB, Camel GH, *et al* (1994). Impact of the treatment of isolated systolic hypertension on behavioral variables. Results from the systolic hypertension in the elderly program. *Arch Intern Med*;**154**:2154–60.

Avis NE, Smith KW (1994). Conceptual and methodological issues in selecting and developing quality of life measures. In: Advances in medical sociology (Fitzpatrick R, editor). London: JAI Press Inc. p. 255–80.

Bakker CH, van der Linden S (1995). Health related utility measurement: an introduction. *J Rheumatol*;**22**:1197–9.

Bayley KB, London MR, Grunkemeier GL, Lansky D J (1995). Measuring the success of treatment in patient terms. *Med Care*;**33**:AS226–35.

Beck A, Ward C, Medelson M, Mock J, Erbaugh J (1961). An inventory for measuring depression. *Arch General Psychiatry*;**4**:561–71.

Bell MJ, Bombardier C, Tugwell P (1990). Measurement of functional status, quality of life and utility in rheumatoid arthritis. *Arthritis Rheum*;**33**:591–601.

Bellamy N, Boers M, Felson D, Fries JF, Furst D, Henry D, *et al* (1995). Health status instruments/ utilities. *J Rheumatol*;**22**:1203–7.

Bennett K, Torrance GW, Tugwell P (1991). Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis. *Control Clin Trials;***12**:118S–128S.

Bergner M (1985). Measurement of health status. *Med Care*;**23**:696–704.

Bergner M (1989). Quality of life, health status and clinical research. *Med Care*;**27**:S148–56.

Bergner M, Bobbitt RA, Pollard WE, Martin DP, Gilson BS (1976). The Sickness Impact Profile: validation of a health status measure. *Med Care*;**14**:57–67.

Bergner M, Rothman ML (1987). Health status measures: an overview and guide for selection. *Ann Rev Public Health*;**8**:191–210.

Berkanovic E, Hurwicz ML, Lachenbruch PA (1995). Concordant and discrepant views of patients' physical functioning. *Arthritis Care Res*;**8**:94–101.

Bernhard J, Gusset H, Hurny C (1995). Quality-of-life assessment in cancer clinical trials: an intervention by itself? *Support Care Cancer*;**3**:66–71.

Beurskens A, de Vet H, Köke A (1996). Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain;***65**:71–6.

Bice TW (1976). Comments on health indicators: methodological perspectives. *Int J Health Serv*;**6**:509–20.

Bindman AB, Keane D, Lurie N (1990). Measuring health changes among severely ill patients. The floor phenomenon. *Med Care*;**28**:1142–52.

Bjordal K, Ahlner Elmqvist M, Tollesson E, Jensen AB, Razavi D, Maher E, *et al* (1994). Development of a European Organization for Research and Treatment of Cancer (EORTC) questionnaire module to be used in quality of life assessments in head and neck cancer patients. EORTC Quality of Life Study Group. *Acta Oncologica;***33**:879–85.

Björk S, Roos P (1994). Analysing changes in health-related quality of life. In: Concepts and measurement of quality of life in health care (Nordenfelt L, editor). Dordrecht: Kluwer Academic Publishers. p. 229–40.

Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*;**1**:307–10.

Bleehen NM, Girling DJ, Machin D, Stephens RJ (1993). A randomised trial of three or six courses of etoposide cyclophosphamide methotrexate and vincristine or six courses of etoposide and ifosfamide in small cell lung cancer (SCLC). II: Quality of life. Medical Research Council Lung Cancer Working Party. *Br J Cancer*;**68**:1157–66.

Bohrnstedt G (1983). Measurement. In: Handbook of survey research (Rossi P, Wright J, Anderson A, editors). New York: Academic Press. p. 69–121.

Bombardier C, Tugwell P (1987). Methodological considerations in functional assessment. *J Rheumatol*; **14** Suppl 15:6–10.

Bombardier C, Raboud J and The Auranofin Cooperating Group (1991). A comparison of health-related quality-of-life measures for rheumatoid arthritis research. *Control Clin Trials*;**12**:243S–256S.

Bombardier C, Melfi CA, Paul J, Green R, Hawker G, Wright JG, *et al* (1995). Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Med Care*;**33**:AS131–44.

Bouchet C, Guillemin F, Briançon S (1996). Nonspecific effects in longitudinal studies: impact on quality of life measures. *J Clin Epidemiol*;**49**:15–20.

Bowling A (1995a). Measuring disease. A review of Disease-Specific Quality of Life Measurement Scales. Buckingham: Open University Press.

Bowling A (1995b). What things are important in people's lives? A survey of the public's judgements to inform scales of health related quality of life. *Soc Sci Med;***41**:1447–62.

Bowling A (1997). Measuring health: a review of Quality of Life Measurement Scales. 2nd edn. Buckingham: Oxford University Press.

Bravo G, Potvin L (1991). Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. *J Clin Epidemiol*;**44**:381–90.

Brazier J, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, *et al* (1992). Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ*;**305**:160–4.

Brazier J, Jones N, Kind P (1993). Testing the validity of the EuroQol and comparing it with the SF-36 health survey questionnaire. *Qual Life Res*;**2**:169–80.

Bremer BA, McCauley CR (1986). Quality-of-life measures: hospital interview versus home questionnaire. *Health Psychol*;**5**:171–77.

Brenner MH, Curbow B, Legro MW (1995). The proximal-distal continuum of multiple health outcome measures: the case of cataract surgery. *Med Care;***33**: AS236–44.

Brook RH, Kamberg CJ (1987). General health status measures and outcome measurement: a commentary on measuring functional status. *J Chronic Dis*;**40**:131S–136S.

Brooks RH and the EuroQol Group (1996). EuroQol: the current state of play. *Health Policy*;**37**:53–72.

Brooks WB, Jordan JS, Divine GW, Smith KS, Neelon FA (1990). The impact of psychologic factors on measurement of functional status. Assessment of the sickness impact profile. *Med Care;***28**:793–804.

Bullinger M (1991). Quality of life: definition, conceptualization and implications – a methodologist's view. *Theoretical Surgery*;**6**:143–8.

Bullinger M, Anderson RB, Cella DF, Aaronson NK (1995). The international assessment of health-related quality of life: theory, translation, measurement and analysis (Shumaker S, Berzon R, editors). Oxford: Rapid Communications of Oxford. p. 83–91.

Bulpitt CJ. (1996). Quality of life with ACE inhibitors in chronic heart failure. *J Cardiovasc Pharmacol*; **27** Suppl 2:S31–5.

Burisch M (1984). You don't always get what you pay for: measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality*;**18**:81–98.

Byrne M (1992). Cancer chemotherapy and quality of life [editorial]. *BMJ*;**304**:1523–4.

Cairns J (1996). Measuring health outcomes [editorial]. *BMJ*;**313**:6.

Calman KC (1984). Quality of life in cancer patients – an hypothesis. *J Medical Ethics*;**10**:124–7.

Campbell D, Fiske D (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychol Bull*;**56**:81–105.

Carr-Hill R (1989). Assumptions of the QALY procedure. *Soc Sci Med*;**29**:469–77.

Carr-Hill R (1992). A second opinion. Health related quality of life measurement – Euro style. *Health Policy*;**20**:321–8.

Carr-Hill R, Morris J (1991). Current practice in obtaining the "Q" in QALYs: a cautionary note. *BMJ*;**303**:699–701.

Cella DF, Tulsky DS (1990). Measuring quality of life today: methodological aspects. *Oncology Huntingt*;**4**:29–38.

Chalmers I, Dickersin K, Chalmers TC (1992). Getting to grips with Archie Cochrane's agenda [editorial]. *BMJ*;**305**:786–8.

Chamber, LW, Macdonald LA, Tugwell P, Buchanan WW, Kraag G (1982). The McMaster Health Index Questionnaire as a measure of quality of life for patients with rheumatoid disease. *J Rheumatol*;**9**:780–4.

Chambers LW, Haight M, Norman GR, Macdonald LA (1987). Sensitivity to change and the effect of mode of administration on health status measurement. *Med Care*;**25**:470–80.

Clark A, Fallowfield LJ (1986). Quality of life measurements in patients with malignant disease: a review. *J R Soc Med*;**79**:165–9.

Clarke AE, Fries JF (1992). Health status instruments and physical examination techniques in clinical measurement methodologies. *Curr Opin Rheumatol*;**4**:145–52.

Cleeland CS (1990). Assessment of pain in cancer. Measurement issues. *Advances in Pain Research and Therapy*;**16**:49–55.

Coast J (1992). Reprocessing data to form QALYs. *BMJ*;**305**:87–90.

Coates A, Gebski V, Bishop JF, Jeal PN, Woods RL, Snyder R, *et al* (1987). Improving the quality of life during chemotherapy for advanced breast cancer. A comparison of intermittent and continuous treatment strategies. *N Engl J Med*;**317**:1490–5.

Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press. p. 1977.

Cook DJ, Guyatt GH, Juniper EF, Griffith LE, McIlroy W, Willan A, *et al* (1993). Interviewer versus self-administered questionnaires in developing a disease-specific, health-related quality of life instrument for asthma. *J Clin Epidemiol*;**46**:529–34.

Coste J, Fermanian J, Venot A (1995). Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals. *Stat Med;***14**:331–45.

Coste J, Guillemin F, Pouchot J, Fermanian J (1997). Methodological approaches to shortening composite measurement scales. *J Clin Epidemiol*;**50**:247–52.

Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR (1992). Quality-of-life assessment: can we keep it simple? *J R Statist Soc*;**155**:353–93.

Croft P, Pope D, Zonca M, O'Neill T, Silman A (1994). Measurement of shoulder related disability: results of a validation study. *Ann Rheum Dis*;**53**:525–28.

Cronbach L (1951). Coefficient alpha and the internal structure of tests. *Psychometrica*;**16**:287–334.

Curtis SE (1987). Self reported morbidity in London and Manchester: intra-urban and inter-urban variations. *Social Indicators Research*;**19**:255–72.

Dawson J, Fitzpatrick R, Carr A (1996a). Questionnaire on the perceptions of patients about total hip replacement. *J Bone Jt Surg*;**78-B**:185–90.

Dawson J, Fitzpatrick R, Murray D, Carr A (1996b). Comparison of measures to assess outcomes in total hip replacement surgery. *Quality in Health Care*;**5**:81–8.

de Haes JC, van Knippenberg FC (1985). The quality of life of cancer patients: a review of the literature. *Soc Sci Med*;**20**:809–17.

de Vaus D (1986). Surveys in social research. London: George Allen & Unwin.

Devinsky O (1995). Outcome research in neurology: incorporating health-related quality of life [editorial]. *Ann Neurol*;**37**:141–2.

Deyo RA (1984). Measuring functional outcomes in therapeutic trials for chronic disease. *Control Clin Trials*;**5**:223–40.

Deyo RA (1991). The quality of life, research and care [editorial]. *Ann Intern Med*;**114**:695–97.

Deyo RA, Centor RM (1986). Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*;**39**:897–906.

Deyo RA, Diehl AK (1986). Patient satisfaction with medical care for low-back pain. *Spine*;**11**:28–30.

Deyo RA, Inui TS (1984). Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Serv Res*;**19**:275–89.

Deyo RA, Patrick DL (1989). Barriers to the use of health status measures in clinical investigation, patient care and policy research. *Med Care*;**27**:S254–68.

Deyo RA, Patrick DL (1995). The significance of treatment effects: the clinical perspective. *Med Care*;**33**:AS286–91.

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

Dillman D (1978). Mail & telephone surveys: the total design method. New York: Wiley.

Dolan P, Kind P (1996). Inconsistency and health state valuations. *Soc Sci Med*;**42**:609–15.

Donovan JL, Frankel SJ, Eyles JD (1993). Assessing the need for health status measures. *J Epidemiol Community Health*;**47**:158–62.

Drummond M (1987). Discussion: Torrance's "utility approach to measuring health related-quality of life". *J Chronic Dis*;**40**:601–3.

Drummond M (1992). The role and importance of quality of life measurements in economic evaluations. *Br J Med Econ*;**4**:9–16.

Drummond M (1993). Estimating utilities for decisions in health care. In: Analysing how we reach clinical decisions (Llewellyn-Thomas H, Hopkins A, editors). London: Royal College of Physicians of London. p. 125–43.

Ebbs SR, Fallowfield LJ, Fraser SC, Baum M (1989). Treatment outcomes and quality of life. *Int J Technol Assess Health Care*;**5**:391–400.

Ebrahim S (1995). Clinical and public health perspectives and applications of health-related quality of life measurement. *Soc Sci Med*;**41**:1383–94.

Editorial (1995). Quality of life and clinical trials. *Lancet*;**346**:1–2.

Eliasziw M, Donner A (1987). A cost–function approach to the design of reliability studies. *Stat Med*;**6**:647–55.

Epstein AM (1990). The outcomes movement – will it get us where we want to go? *N Engl J Med*;**323**:266–70.

Erickson P, Taeuber RC, Scott J (1995). Operational aspects of quality-of-life assessment: choosing the right instrument: review article. *PharmacoEconomics*;**7**:39–48.

European Research Group on Health Outcomes Measures (1996). Choosing a health outcomes measurement instrument. *Quality of Life Newsletter*;6–7.

Fallowfield LJ (1993). Quality of life measurement in breast cancer. *J R Soc Med*;**86**:10–12.

Fallowfield LJ (1996). Quality of quality-of-life data. *Lancet*;**348**:421–2.

Farquhar M (1994). The quality of life in older people. In: Quality of Life in Health Care (Albrecht G, Fitzpatrick R, editors). Connecticut: JAI Press. p. 139–58.

Farquhar M (1995). Definitions of quality of life: a taxonomy. *J Adv Nurs*;**22**:502–8.

Fava GA (1990). Methodological and conceptual issues in research on quality of life. *Psychother Psychosom*;**54**:70–6.

Fayers PM, Hand D (1997). Factor analysis, causal indicators and quality of life. *Qual Life Res*;**6**:139–50.

Fayers PM, Jones DR (1983). Measuring and analysing quality of life in cancer clinical trials: a review. *Stat Med*;**2**:429–46.

Fayers PM, Machin D (1998). Factor analysis. In: Quality of life assessment in clinical trials (Staquet M, Hays R, Fayers P, editors). Oxford: Oxford University Press. p. 191–226.

Fayers PM, Hopwood P, Harvey A, Girling DJ, Machin D, Stephens R (1997). Quality of life assessment in clinical trials – guidelines and a checklist for protocol writers: the U.K. Medical Research Council experience. MRC Cancer Trials Office. *Eur J Cancer*;**33**:20–8.

Feeny DH, Torrance GW (1989). Incorporating utility-based quality-of-life assessment measures in clinical trials. Two examples. *Med Care*;**27**:S190–S204.

Feeny DH, Furlong W, Boyle M, Torrance GW (1995). Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics*;**7**:490–502.

Feinstein AR (1987). The theory and evaluation on sensibility. In: Clinimetrics (Feinstein AR, editor). New Haven: Yale University Press. p. 141–66.

Feinstein AR (1992). Benefits and obstacles for development of health status assessment measures in clinical settings. *Med Care*;**30**:MS50–6.

Feldstein ML (1991). Quality-of-life-adjusted survival for comparing cancer treatments. A commentary on TWiST and Q-TWiST. *Cancer*;**67**:851–4.

Fitzpatrick R (1993). The measurement of health status and quality of life in rheumatological disorders. In: Psychological aspects of rheumatic disease (Newman S, Shipley M, editors). London: Bailliére Tindall. p. 297–317.

Fitzpatrick R (1994). Applications of health status measures. In: Measuring health and medical outcomes (Jenkinson C, editor). London: University College London Press. p. 27–41.

Fitzpatrick R (1996). Alternative approaches to the assessment of health-related quality of life. In: In pursuit of the quality of life (Offer A, editor). Oxford: Clarendon Press. p. 140–62.

Fitzpatrick R, Albrecht G (1994). The plausibility of quality-of-life measures in different domains of health care. In: Concepts and measurement of quality of life in health care (Nordenfelt L, editor). Dordrecht: Kluwer Academic Publishers. p. 201–27.

Fitzpatrick R, Newman S, Lamb R, Shipley M (1989). A comparison of measures of health status in rheumatoid arthritis. *Br J Rheumatol*;**28**:201–6.

Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A (1991). The social dimension of health status measures in rheumatoid arthritis. *Int Disabil Stud*;**13**:34–7.

Fitzpatrick R, Fletcher AE, Gore SM, Jones D, Spiegelhalter DJ, Cox DR (1992). Quality of life measures in health care. I: applications and issues in assessment. *BMJ*;**305**:1074–7.

Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A (1993a). A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol*;**20**:429–36.

Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A (1993b). Transition questions to assess outcomes in rheumatoid arthritis. *Br J Rheumatol*;**32**:807–11.

Fletcher AE (1988). Measurement of quality of life in clinical trials of therapy. *Recent Results Cancer Res*;**111**:216–30.

Fletcher AE, Hunt BM, Bulpitt CJ (1987). Evaluation of quality of life in clinical trials of cardiovascular disease. *J Chronic Dis*;**40**:557–69.

Fletcher AE, Gore SM, Jones D, Fitzpatrick R, Spiegelhalter DJ, Cox DR (1992). Quality of life measures in health care. II: design, analysis and interpretation. *BMJ*;**305**:1145–8.

Fowlie M, Berkeley J (1987). Quality of life – a review of the literature. *Fam Pract*;**4**:226–34.

Frankel SJ (1991). The epidemiology of indications [editorial]. *J Epidemiol Community Health*;**45**:257–9.

Fries JF (1983). Toward an understanding of patient outcome measurement. *Arthritis Rheum*;**26**:697–704.

Froberg DG, Kane RL (1989a). Methodology for measuring health-state preferences – IV: progress and a research agenda. *J Clin Epidemiol*;**42**:675–85.

Froberg DG, Kane RL (1989b). Methodology for measuring health-state preferences – I: measurement strategies. *J Clin Epidemiol*;**42**:345–54.

Froberg DG, Kane RL (1989c). Methodology for measuring health-state preferences – II: scaling methods. *J Clin Epidemiol*;**42**:459–71.

Froberg DG, Kane RL (1989d). Methodology for measuring health-state preferences – III: population and context effects. *J Clin Epidemiol*;**42**:585–92.

Gafni A (1994). The standard gamble method: what is being measured and how it is interpreted. *Health Serv Res*;**29**:207–24.

Gafni A, Birch S (1993). Searching for a common currency: critical appraisal of the scientific basis underlying European harmonization of the measurement of health related quality of life (EuroQol). *Health Policy*;**23**:219–28.

Ganiats TG, Palinkas LA, Kaplan RM (1992). Comparison of Quality of Well-Being scale and Functional Status Index in patients with atrial fibrillation. *Med Care*;**30**:958–64.

Ganz PA, Moinpour CM, Cella DF, Fetting JH (1992). Quality-of-life assessment in cancer clinical trials: a status report [editorial]. *J Natl Cancer Inst*;**84**:994–5.

Ganz PA (1994). Long-range effect of clinical trial interventions on quality of life. *Cancer*;**74**:2620–4.

Gardiner PV, Sykes HR, Hassey GA, Walker DJ (1993). An evaluation of the Health Assessment Questionnaire in long-term longitudinal follow-up of disability in rheumatoid arthritis. *Br J Rheumatol*;**32**:724–8.

Garratt A, Ruta D, Abdalla M, Russell I (1994). SF-36 health survey questionnaire: II responsiveness to changes in health status in four common clinical conditions. *Quality in Health Care*;**3**:186–92.

Gill TM, Feinstein AR (1994). A critical appraisal of the quality of quality-of-life measurements. *JAMA*;**272**:619–26.

Gold M, Patrick DL, Torrance GW, *et al* (1996). Identifying and valuing outcomes. In: Cost-effectiveness in health and medicine (Gold M, Siegel J, Russell L, Weinstein M, editors). New York: Oxford University Press. p. 82–134.

Gotay CC, Korn EL, McCabe MS, Moore TD, Cheson BD (1992). Quality-of-life assessment in cancer treatment protocols: research issues in protocol development. *J Natl Cancer Inst*;**84**:575–9.

Gough IR, Furnival CM, Schilder L, Grove W (1983). Assessment of the quality of life of patients with advanced cancer. *Eur J Cancer Clin Oncol*;**19**:1161–5.

Gower NH, Rudd RM, Ruiz de Elvira MC, Spiro SG, James LE, Harper PG, *et al* (1995). Assessment of 'quality of life' using a daily diary card in a randomised trial of chemotherapy in small-cell lung cancer. *Ann Oncol*;**6**:575–80.

Greenfield S, Nelson EC (1992). Recent developments and future issues in the use of health status assessment measures in clinical settings. *Med Care*;**30**:MS23–41.

Groenvold M, Bjorner JB, Klee MC, Kreiner S (1995). Test for item bias in a quality of life questionnaire. *J Clin Epidemiol*;**48**:805–16.

Guyatt GH (1995). A taxonomy of health status instruments. *J Rheumatol*;**22**:1188–90.

Guyatt GH, Cook DJ (1994). Health status, quality of life and the individual. *JAMA*;**272**:630–31.

Guyatt GH, Bombardier C, Tugwell P (1986). Measuring disease-specific quality of life in clinical trials. *Can Med Assoc J*;**134**:889–95.

Guyatt GH, Townsend M, Berman LB, Keller JL (1987a). A comparison of Likert and visual analogue scales for measuring change in function. *J Chronic Dis*;**40**:1129–33.

Guyatt GH, Walter S, Norman GR (1987b). Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*;**40**:171–8.

Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A (1989a). Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol*;**42**:403–8.

Guyatt GH, Veldhuyzen van Zanten SJ, Feeny DH, Patrick DL (1989b). Measuring quality of life in clinical trials: a taxonomy and review. *Can Med Assoc J*;**140**:1441–8.

Guyatt GH, Feeny DH, Patrick DL (1991). Issues in quality-of-life measurement in clinical trials. *Control Clin Trials*;**12**:81S–90S.

Guyatt GH, Kirshner B, Jaeschke R (1992a). Measuring health status: what are the necessary measurement properties? *J Clin Epidemiol*;**45**:1341–5.

Guyatt GH, Kirshner B, Jaeschke R (1992b). A methodologic framework for health status measures: clarity or oversimplification? *J Clin Epidemiol*;**45**:1353–5.

Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith LE, McIlroy W, *et al* (1993a). Measuring quality of life in the frail elderly. *J Clin Epidemiol*;**46**:1433–4.

Guyatt GH, Feeny DH, Patrick DL (1993b). Measuring health-related quality of life. *Ann Intern Med*;**118**:622–9.

Hadorn DC, Uebersax J (1995). Large-scale health outcomes evaluation: how should quality of life be measured? Part I – Calibration of a brief questionnaire and a search for preference subgroups. *J Clin Epidemiol*;**48**:607–18.

Hadorn DC, Hays RD, Uebersax J, Hauber T (1992). Improving task comprehension in the measurement of health state preferences. A trial of informational cartoon figures and a paired-comparison task. *J Clin Epidemiol*;**45**:233–43.

Hadorn DC, Sorensen J, Holte J (1995). Large-scale health outcomes evaluation: how should quality of life be measured? Part II – Questionnaire validation in a cohort of patients with advanced cancer. *J Clin Epidemiol*;**48**:619–29.

Haig TH, Scott DA, Wickett LI (1986). The rational zero point for an illness index with ratio properties. *Med Care*;**24**:113–24.

Haley SM, McHorney CA, Ware J (1994). Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol*;**47**:671–84.

Hall JA, Epstein AM, McNeil BJ (1989). Multidimensionality of health status in an elderly population. Construct validity of a measurement battery. *Med Care*;**27**:S168–77.

Hayes V, Morris J, Wolfe C, Morgan M (1995). The SF-36 health survey questionnaire: is it suitable for use with older adults? *Age Ageing*;**24**:120–5.

Hays RD, Hadorn DC (1992). Responsiveness to change: an aspect of validity, not a separate dimension. *Qual Life Res*;**1**:73–5.

Hays RD, Anderson RB, Revicki DA (1993). Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res*;**2**:441–9.

Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelly FD, Shannon W (1996). A new short form individual quality of life measure (SEIQoL-DW): application in a cohort of individuals with HIV/AIDS. *BMJ*;**313**:29–33.

Hill S, Harries U, Popay J (1996). Is the short form 36 (SF-36) suitable for routine health outcomes assessment in health care for older people? Evidence from preliminary work in community based health services in England. *J Epidemiol Community Health*;**50**:94–8.

Hobart JC, Lamping DL, Thompson AJ (1996). Evaluating neurological outcome measures: the bare essentials [editorial]. *J Neurol Neurosurg Psychiatry*;**60**:127–30.

Hopwood P (1992). Progress, problems and priorities in quality of life research. *Eur J Cancer*;**28A**:1748–52.

Hopwood P, Stephens RJ, Machin D (1994). Approaches to the analysis of quality of life data: experiences gained from a medical research council lung cancer working party palliative chemotherapy trial. *Qual Life Res*;**3**:339–52.

Hughes C, Hwang B, Kim JH, Eisenman LT, Killian DJ (1995). Quality of life in applied research: a review and analysis of empirical measures. *Am J Ment Retard*;**99**:623–41.

Hunt SM (1988). Measuring health in clinical care and clinical trials. In: Measuring health: a practical approach (Teeling-Smith G, editor). Chichester: John Wiley & Sons. p. 7–21.

Hunt SM (1998). Cross-cultural issues in the use of quality of life measures in randomised controlled trials. In: Quality of life assessment in clinical trials (Staquet M, Hays R, Fayers P, editors). Oxford: Oxford University Press. p. 51–68.

Hunt SM, McEwen J, McKenna SP (1985). Measuring health status: a new tool for clinicians and epidemiologists. *J R Coll Gen Pract*;**35**:185–8.

Hurny C, Bernhard J, Coates A, Peterson HF, Castiglione Gertsch M, Gelber RD, *et al* (1996). Responsiveness of a single-item indicator versus a multi-item scale. *Med Care*;**34**:234–48.

Hurst J, Mooney G (1983). Implicit values in administrative decisions. In: Health indicators (Culyer A, editor). Oxford: Martin Roberston. p. 173–85.

Idler EL, Angel RJ (1990). Self-rated health and mortality in the NHANES – I Epidemiologic Follow-up Study. *Am J Public Health*;**80**:446–52.

Jaeschke R, Singer J, Guyatt GH (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*;**10**:407–15.

Jaeschke R, Singer J, Guyatt GH (1990). A comparison of seven-point and visual analogue scales. Data from a randomized trial. *Control Clin Trials*;**11**:43–51.

Jaeschke R, Guyatt GH, Keller JL, Singer J (1991). Interpreting changes in quality-of-life score in n of 1 randomized trials. *Control Clin Trials*;**12**:226S–233S.

Jaeschke R, Guyatt GH, Cook DJ (1992). Quality of life instruments in the evaluation of new drugs. *PharmacoEconomics*;**1**:84–94.

Jenkins CD (1992). Assessment of outcomes of health intervention. *Soc Sci Med*;**35**:367–75.

Jenkinson C (1991). Why are we weighting? A critical examination of the use of item weights in a health status measure. *Soc Sci Med*;**32**:1413–16.

Jenkinson C (1994). Weighting for ill health: the Nottingham Health Profile. In: Measuring health and medical outcomes (Jenkinson C, editor). London: University College London Press. p. 77–87.

Jenkinson C (1995). Evaluating the efficacy of medical treatment: possibilities and limitations. *Soc Sci Med*;**41**:1395–401.

Jenkinson C, Ziebland S, Fitzpatrick R, Mowat A (1993). Hospitalisation and its influence upon results from health status questionnaires. *Int J Health Sciences*;**4**:13–18.

Jenkinson C, Peto V, Coulter A (1994). Measuring change over time: a comparison of results from a global single item of health status and the multi-dimensional SF-36 health status survey questionnaire in patients presenting with menorrhagia. *Qual Life Res*;**3**:317–21.

Jenkinson C, Carroll D, Egerton M, Frankland T, McQuay H, Nagle C (1995a). Comparison of the sensitivity to change of long and short form pain measures. *Qual Life Res*;**4**:353–7.

Jenkinson C, Lawrence KC, McWhinnie D, Gordon J (1995b). Sensitivity to change of health status measures in a randomized controlled trial: comparison of the COOP charts and the SF-36. *Qual Life Res*;**4**:47–52.

Jenkinson C, Layte R, Wright L, Coulter A (1996). The U.K. SF-36: an analysis and interpretation manual. A guide to health status measurement with particular reference to the Short Form 36 Health Survey, Health Services Research Unit, Oxford.

Jette AM (1980). Health status indicators: their utility in chronic-disease evaluation research. *J Chronic Dis*;**33**:567–79.

Johnson JR (1993). Quality of life assessment. Key issues in the 1990's (Walker SR, Rosser RM, editors). Dordrecht: Kluwer Academic Publishers. p. 393–400.

Jones PW, Quirk FH, Baveystock CM, Littlejohns P (1992). A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am Rev Respir Dis*;**145**:1321–7.

Joyce CR (1994). Health status and quality of life: which matters to the patient? *J Cardiovasc Pharmacol*;**23** Suppl 3:S26–33.

Julious SA, George S, Campbell MJ (1995). Sample sizes for studies using the short form 36 (SF-36). *J Epidemiol Community Health*;**49**:642–4.

Juniper EF, Guyatt GH, Willan A, Griffith LE (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol*;**47**:81–7.

Juniper EF, Guyatt GH, Streiner DL, King DR (1997). Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol*;**50**:233–8.

Kaasa S (1992). Measurement of quality of life in clinical trials. *Oncology*;**49**:289–94.

Kahneman D, Varey C (1991). Notes on the psychology of utility. In: Interpersonal comparisons of well-being (Elster J, Roemer J, editors). Cambridge: Cambridge University Press. p. 127–59.

Kaplan RM (1993). Application of a general health policy model in the American health care crisis. *J R Soc Med*;**86**:277–81.

Kaplan RM, Anderson JP, Wu AW, Mathews WC, Kozin F, Orenstein D (1989). The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis and arthritis. *Med Care*;**27**:S27–43.

Kaplan RM, Coons SJ, Anderson JP (1992). Quality of life and policy analysis in arthritis. *Arthritis Care Res*;**5**:173–83.

Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH (1992). Comparative measurement sensitivity of short and longer health status instruments. *Med Care*;**30**:917–25.

Kazis LE, Anderson JJ, Meenan RF (1989). Effect sizes for interpreting changes in health status. *Med Care*;**27**:S178–89.

Kazis LE, Callahan LF, Meenan RF, Pincus T (1990). Health status reports in the care of patients with rheumatoid arthritis. *J Clin Epidemiol*;**43**:1243–53.

Kelly H, Russell EM, Stewart S, McEwen J (1996). Needs assessment: taking stock. *Health Bulletin*;**54**:115–18.

Keoghane SR, Lawrence KC, Jenkinson C, Doll HA, Chappel DB, Cranston DW (1996). The Oxford Laser Prostate Trial: sensitivity to change of three measures of outcome. *Urology*;**47**:43–7.

Kessler RC, Mroczek DK (1995). Measuring the effects of medical interventions. *Med Care*;**33**:AS109–19.

Kind P, Carr-Hill R (1987). The Nottingham Health Profile: a useful tool for epidemiologists? *Soc Sci Med*;**25**:905–10.

Kind P, Gudex C, Dolan P, Williams A (1994). Quality of life in health care (Albrecht G, Fitzpatrick R, editors). Greenwich CT: JAI Press. p. 219–53.

King MT, Dobson AJ, Harnett PR (1996). A comparison of two quality–of–life questionnaires for cancer clinical trials: the functional living index – cancer (FLIC) and the quality of life questionnaire core module (QLQ-C30). *J Clin Epidemiol*;**49**:21–9.

Kirshner B (1991). Methodological standards for assessing therapeutic equivalence. *J Clin Epidemiol*;**44**:839–49.

Kirshner B, Guyatt GH (1985). A methodological framework for assessing health indices. *J Chronic Dis*;**38**:27–36.

Kline P (1986). A handbook of test construction. London: Methuen.

Korner Bitensky N, Wood Dauphinee S, Siemiatycki J, Shapiro S, Becker R (1994). Health-related information post-discharge: telephone versus face-to-face interviewing. *Arch Phys Med Rehabil*;**75**:1287–96.

Krause NM, Jay GM (1994). What do global self-rated health items measure? *Med Care*;**32**:930–42.

Kuyken W, Orley J, Hudelson P, Sartorius N (1994). Quality of life assessment across cultures. *International Journal of Mental Health*;**23**:5–27.

Lansky D, Butler JB, Waller FT (1992). Using health status measures in the hospital setting: from acute care to 'outcomes management'. *Med Care*;**30**:MS57–73.

Leavey R, Wilkin D (1988). A comparison of two survey measures of health status. *Soc Sci Med*;**27**:269–75.

Leplege A, Verdier A (1995). The adaptation of health status measures: methodological aspects of the translation procedure. In: The international assessment of health-related quality of life: theory, translation, measurement and analysis (Shumaker S, Berzon R, editors). Oxford: Rapid Communications of Oxford. p. 93–101.

Liang MH (1995). Evaluating measurement responsiveness. *J Rheumatol*;**22**:1191–2.

Liang MH, Cullen KE, Larson M (1982). In search of a more perfect mousetrap (health status or quality of life instrument). *J Rheumatol*;**9**:775–9.

Liang MH, Larson MG, Cullen KE, Schwartz JA (1985). Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum*;**28**:542–7.

Liang MH, Fossel AH, Larson MG (1990). Comparisons of five health status instruments for orthopedic evaluation. *Med Care*;**28**:632–42.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF (1982). The measurement of patients' values in medicine. *Med Decis Making*;**2**:449–62.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF (1984). Describing health states. Methodologic issues in obtaining values for health states. *Med Care*;**22**:543–52.

Llewellyn-Thomas H, Sutherland HJ, Thiel EC (1993). Do patients' evaluations of a future health state change when they actually enter that state? *Med Care*;**31**:1002–12.

Lomas J, Pickard L, Mohide A (1987). Patient versus clinician item generation for quality-of-life measures. The case of language-disabled adults. *Med Care*;**25**:764–9.

Long A, Dixon P (1996). Monitoring outcomes in routine practice: defining appropriate measurement criteria. *J Evaluation in Clin Practice*;**2**:71–8.

Lovatt B (1992). An overview of quality of life assessments and outcome measures. *Br J Med Econ*;**4**:1–7.

Lundberg O, Manderbacka K (1996). Assessing reliability of a measure of self-rated health. *Scand J Soc Med*;**24**:218–24.

Lydick E, Epstein RS (1993). Interpretation of quality of life changes. *Qual Life Res*;**2**:221–6.

Lyons RA, Perry HM, Littlepage BN (1994). Evidence for the validity of the Short-Form 36 Questionnaire (SF-36) in an elderly population. *Age Ageing*;**23**:182–4.

MacKenzie CR, Charlson M (1986). Standards for the use of ordinal scales in clinical trials. *BMJ Clin Res Ed*;**292**:40–3.

MacKenzie CR, Charlson M, DiGioia D, Kelley K (1986a). A patient-specific measure of change in maximal function. *Arch Intern Med*;**146**:1325–9.

MacKenzie CR, Charlson M, DiGioia D, Kelley K (1986b). Can the Sickness Impact Profile measure change? An example of scale assessment. *J Chronic Dis*;**39**:429–38.

Maguire P, Selby P (1989). Assessing quality of life in cancer patients. *Br J Cancer*;**60**:437–40.

Mancuso CA, Charlson ME (1995). Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Med Care*;**33**:AS77–88.

Mauskopf JA, Austin R, Dix LP, Berzon RA (1995). Estimating the value of a generic quality-of-life measure. *Med Care*;**33**:AS195–202.

McColl E, Steen IN, Meadows KA, Hutchinson A, Eccles MP, Hewison J, *et al* (1995). Developing outcome measures for ambulatory care – an application to asthma and diabetes. *Soc Sci Med*;**41**:1339–48.

McDaniel RW, Bach CA (1994). Quality of life: a concept analysis. Rehabilitation. *Nursing Research*;**3**:18–22.

McDowell I, Jenkinson C (1996). Development standards for health measures. *J Hlth Serv Res Policy*;**1**:238–46.

McDowell, I. and Newell, C.(1996). Measuring health: a guide to rating scales and questionnaires. 2nd edn. New York: Oxford University Press.

McNair D, Lorr M, Dropleman L (1992). EDITS Manual for the Profile of Mood States (POMS) (abstract). San Diego CA: EDITS/Educational and Industrial Testing Service.

McNeil BJ, Pauker SG, Sox HC Jr., Tversky A (1982). On the elicitation of preferences for alternative therapies. *N Engl J Med*;**306**:1259–62.

Medical Research Council (1995). Guidelines for the collection of quality of life data. For those administering quality of life questionnaires in the clinical setting (abstract).

Meenan RF (1982). The AIMS approach to health status measurement: conceptual background and measurement properties. *J Rheumatol*;**9**:785–8.

Meenan RF, Pincus T (1987). The status of patient status measures. *J Rheumatol*;**14**:411–14.

Meenan RF, Gertman PM, Mason JH (1980). Measuring health status in arthritis. The arthritis impact measurement scales. *Arthritis Rheum*;**23**:146–52.

Meenan RF, Anderson JJ, Kazis LE, Egger MJ, Altz Smith M, Samuelson CO Jr., *et al* (1984). Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum*;**27**:1344–52.

Melzack R (1975). The McGill Pain Questionnaire: major properties and scoring methods. *Pain*;**1**:277–99.

Moinpour CM, Feigl P, Metch B, Hayden KA, Meyskens FL Jr., Crowley J (1989). Quality of life end points in cancer clinical trials: review and recommendations. *J Natl Cancer Inst*;**81**:485–95.

Mor V, Guadagnoli E (1988). Quality of life measurement: a psychometric tower of Babel. *J Clin Epidemiol*;**41**:1055–8.

Morrow GR, Lindke J, Black P (1992). Measurement of quality of life in patients: psychometric analyses of the Functional Living Index – Cancer (FLIC). *Qual Life Res*;**1**:287–96.

Mossey JM, Shapiro E (1982). Self-rated health: a predictor of mortality among the elderly. *Am J Public Health*;**72**:800–8.

Mosteller F, Ware J, Levine S (1989). Finale panel: comments on the conference on Advances in Health Status Assessment. *Med Care*;**27**:S282–93.

Mulley AG (1989). Assessing patients' utilities. Can the ends justify the means? *Med Care*;**27**:S269–81.

Najman JM, Levine S (1981). Evaluating the impact of medical care and technologies on the quality of life: a review and critique. *Soc Sci Med*;**15**:107–15.

Nayfield SG, Ganz PA, Moinpour CM, Cella DF, Hailey BJ (1992). Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials. *Qual Life Res*;**1**:203–10.

Nelson EC, Berwick DM (1989). The measurement of health status in clinical practice. *Med Care*;**27**:S77–90.

Nelson EC, Landgraf JM, Hays RD, Wasson JH, Kirk JW (1990). The functional status of patients. How can it be measured in physicians' offices? *Med Care*;**28**:1111–26.

Neugebauer E, Troidl H, WoodDauphinee S, Eypasch E, Bullinger M (1991). *Theoretical Surgery*;**6**:123–37.

Nord E (1991). The validity of a visual analogue scale in determining social utility weights for health states. *Int J Health Planning and Management*;**6**:234–42.

Nord E (1992). Methods for quality adjustment of life years. *Soc Sci Med*;**34**:559–69.

Nord E (1993). Toward quality assurance in QALY calculations. *Int J Technol Assess Health Care*;**9**:37–45.

Nord E, Richardson J, Macarounas-Kirchmann K (1993). Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling instruments using Norwegian and Australian surveys. *Int J Technol Assess Health Care*;**9**:463–78.

Nord E, Richardson J, Street A, Kuhse H, Singer P (1995). Maximizing health benefits vs egalitarianism: an Australian survey of health issues. *Soc Sci Med*;**41**:1429–37.

Nordenfelt L (1994). Concepts and measurement of quality of life in health care. Dordrecht: Kluwer Academic Publishers.

Nunnally J, Bernstein JC (1994). Psychometric theory. 3rd edn. New York: McGraw-Hill.

O'Boyle CA (1992). Assessment of quality of life in surgery. *Br J Surg*;**79**:395–8.

O'Boyle CA (1995). Making subjectivity scientific. *Lancet*;**345**:602.

O'Boyle CA, McGee H, Hickey AM, O'Malley K, Joyce CR (1992). Individual quality of life in patients undergoing hip replacement. *Lancet*;**339**:1088–91.

O'Brien J, Francis A (1988). The use of next-of-kin to estimate pain in cancer patients. *Pain*;**35**:171–8.

O'Hanlon M, Fox-Rushby J, Buxton MJ (1994). A qualitative and quantitative comparison of the EuroQol and Time Trade-off techniques. *Int J Health Sciences*;**5**:85–97.

O'Neill C, Normand C, Cupples M, McKnight A (1996). A comparison of three measures of perceived distress: results from a study of angina patients in general practice in Northern Ireland. *J Epidemiol Community Health*;**50**:202–6.

Olsson G, Lubsen J, van Es GA, Rehnqvist N (1986). Quality of life after myocardial infarction: effect of long term metoprolol on mortality and morbidity. *BMJ Clin Res Ed*;**292**:1491–3.

Oppenheim A (1992). Questionnaire design, interviewing and attitude measurement. London: Pinter Publishers.

Orley J, Kuyken W (1994). Quality of life assessment: international perspectives. Berlin: Springer Verlag.

Osoba D (1992). The Quality of Life Committee of the Clinical Trials Group of the National Cancer Institute of Canada: organization and functions. *Qual Life Res*;**1**:211–18.

Patrick DL (1976). Constructing social metrics for health status indexes. *Int J Health Serv*;**6**:443–53.

Patrick DL (1992). Health-related quality of life in pharmaceutical evaluation: forging progress and avoiding pitfalls. *PharmacoEconomics*;**1**:76–8.

Patrick DL, Deyo RA (1989). Generic and disease-specific measures in assessing health status and quality of life. *Med Care*;**27**:S217–32.

Patrick DL, Erickson P (1993a). Assessing health-related quality of life for clinical decision-making. In: Quality of life assessment. Key issues in the 1990's (Walker SR, Rosser RM, editors). Dordrecht: Kluwer Academic Publishers. p. 11–63.

Patrick DL, Erickson P (1993b). Health status and health policy. Oxford: Oxford University Press.

Patrick DL, Peach H (1989). Disablement in the community. Oxford: Oxford University Press.

Patrick DL, Sittampalam Y, Somerville SM, Carter WB, Bergner M (1985). A cross-cultural comparison of health status values. *Am J Public Health*;**75**:1402–7.

Peck JR, Smith TW, Ward JR, Milano R (1989). Disability and depression in rheumatoid arthritis. A multi-trait, multi-method investigation. *Arthritis Rheum*;**32**:1100–6.

Peloso PM, Wright JG, Bombardier C (1995). A critical appraisal of toxicity indexes in rheumatology. *J Rheumatol*;**22**:989–94.

Petrou S, Malek M, Davey PG (1993). The reliability of cost-utility estimates in cost-per-QALY league tables. *PharmacoEconomics*;**3**:345–53.

Pincus T, Callahan LF, Bradley LA, Vaughn WK, Wolfe F (1986). Elevated MMPI scores for hypochondriasis, depression and hysteria in patients with rheumatoid arthritis reflect disease rather than psychological status. *Arthritis Rheum*;**29**:1456–66.

Pocock SJ (1991). A perspective on the role of quality-of-life assessment in clinical trials. *Control Clin Trials*;**12**:257S–265S.

Read JL(1993). The new era of quality of life assessment. In: Quality of life assessment. Key issues in the 1990's (Walker SR, Rosser RM, editors). Dordrecht: Kluwer Academic Publishers. p. 3–10.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC (1984). Preferences for health outcomes. Comparison of assessment methods. *Med Decis Making*;**4**:315–29.

Read JL, Quinn RJ, Hoefer MA (1987). Measuring overall health: an evaluation of three important approaches. *J Chronic Dis*;**40** Suppl 1:7S–26S.

Remington M, Tyrer PJ, Newson Smith J, Cicchetti DV (1979). Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychol Med*;**9**:765–70.

Revicki DA, Kaplan RM (1993). Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Qual Life Res*;**2**:477–87.

Revicki DA, Wu AW, Murray MI (1995). Change in clinical status, health status and health utility outcomes in HIV-infected patients. *Med Care*;**33**:AS173–82.

Richardson J (1992). Cost-utility analyses in health care. In: Researching health care (Daley J, McDonald I, Willis E, editors). London: Routledge. p. 21–44.

Richardson J (1994). Cost utility analysis: what should be measured? *Soc Sci Med*;**39**:7–21.

Robine JM, Ritchie K (1991). Healthy life expectancy: evaluation of global indicator of change in population health. *BMJ*;**302**:457–60.

Rogerson RJ (1995). Environmental and health-related quality of life: conceptual and methodological similarities. *Soc Sci Med*;**41**:1373–82.

Rorabeck CH, Bourne RB, Laupacis A, Feeny D, Wong C, Tugwell P, *et al* (1994). A double-blind study of 250 cases comparing cemented with cementless total hip arthroplasty. Cost-effectiveness and its impact on health-related quality of life. *Clin Orthop*;156–64.

Rosenberg R (1995). Health-related quality of life between naturalism and hermeneutics. *Soc Sci Med*;**41**:1411–15.

Rosser R, Sintonen H (1993). The EuroQol quality of life project. In: Quality of life assessment. Key issues in the 1990's (Walker SR, Rosser RM, editors). Dordrecht: Kluwer Academic Publishers. p. 197–99.

Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ (1991). The validity of proxy-generated scores as measures of patient health status. *Med Care*;**29**:115–24.

Rubenstein LV, Calkins DR, Young RT, Cleary PD, Fink A, Kosecoff J, *et al* (1989). Improving patient function: a randomized trial of functional disability screening. *Ann Intern Med*;**111**:836–42.

Ruta D, Garratt A (1994). Health status to quality of life measurement. In: Measuring health and medical outcomes (Jenkinson C, editor). London: University College London Press. p. 138–59.

Rutten van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S (1995a). Methodological issues of patient utility measurement. Experience from two clinical trials. *Med Care*;**33**:922–37.

Rutten van Molken MP, Custers F, van Doorslaer EK, Jansen CC, Heurman L, Maesen FP, *et al* (1995b). Comparison of performance of four instruments in evaluating the effects of salmeterol on asthma quality of life. *Eur Respir J*;**8**:888–98.

Sackett DL, Torrance GW (1978). The utility of different health states as perceived by the general public. *J Chronic Dis*;**31**:697–704.

Sadura A, Pater JL, Osoba D, Levine MN, Palmer M, Bennett K (1992). Quality-of-life assessment: patient compliance with questionnaire completion. *J Natl Cancer Inst*;**84**:1023–6.

Schipper H, Clinch J (1988). Assessment of treatment in cancer. In: Measuring health: a practical approach (Teeling-Smith G, editor). Chichester: John Wiley & Sons. p. 109–39.

Schipper H, Clinch J, Olweny C (1996). Quality of life studies: definitions and conceptual issues. In: Quality of life and pharmacoeconomics in clinical trials (Spilker B, editor). Philadelphia: Lippincott-Raven Publishers. p.11–23.

Schumacher M, Olschewski M, Schulgen G (1991). Assessment of quality of life in clinical trials. *Stat Med*;**10**:1915–30.

Schwarz N, Strack F (1991). Subjective well-being (Strack F, Argyle M, Schwarz N, editors). Oxford: Pergamon Press. p. 27–48.

Scientific Advisory Committee of the Medical Outcomes Trust (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin*;**3**:I–IV. (Abstract)

Scrivens E, Cunningham D, Charlton J, Holland W (1985). Measuring the impact of health interventions: a review of available instruments. *Effective Health Care*;**2**:247–160.

Segovia J, Bartlett RF, Edwards AC (1989). An empirical analysis of the dimensions of health status measures. *Soc Sci Med*;**29**:761–8.

Selby P (1993). Quality of life assessment. Key issues in the 1990's (Walker SR, Rosser RM, editors). Dordrecht: Kluwer Academic Publishers. p. 235–67.

Selby P, Robertson B (1987). Measurement of quality of life in patients with cancer. *Cancer Surv*;**6**:521–43.

Sen A (1970). Collective choice and social welfare. Edinburgh: Oliver Boyd.

Sensky T, Catalan J (1992). Asking patients about their treatment [editorial]. *BMJ*;**305**:1109–10.

Shumaker S, Berzon RA (1995). The international assessment of health-related quality of life: theory, translation, measurement and analysis. Oxford: Rapid Communications of Oxford.

Siegrist J, Junge A (1989). Conceptual and methodological problems in research on the quality of life in clinical medicine. *Soc Sci Med*;**29**:463–8.

Skeel RT (1989). Quality of life assessment in cancer clinical trials – it's time to catch up [editorial]. *J Natl Cancer Inst*;**81**:472–3.

Slevin ML, Plant H, Lynch D, Drinkwater J, Gregory WM (1988). Who should measure quality of life, the doctor or the patient? *Br J Cancer*;**57**:109–12.

Slevin ML, Stubbs L, Plant HJ, Wilson P, Gregory WM, Armes PJ, Downer SM (1990). Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses and general public [see comments]. *BMJ*;**300**:1458–60.

Smith R, Dobson M (1993). Measuring utility values for QALYs: two methodological issues. *Health Econ*;**2**:349–55.

Sneeuw KC, Aaronson NK, Sprangers MA, Detmar SB, Wever LD, Schornagel JH (1997). Value of caregiver ratings in evaluating the quality of life of patients with cancer. *J Clin Oncology*;**15**:1206–17.

Spector WD, Katz S, Murphy JB, Fulton JP (1987). The hierarchical relationship between activities of daily living and instrumental activities of daily living. *J Chronic Dis*;**40**:481–9.

Spiegel JS, Leake B, Spiegel TM, Paulus HE, Kane RL, Ward NB, *et al* (1988). What are we measuring? An examination of self-reported functional status measures. *Arthritis Rheum*;**31**:721–8.

Spiegelhalter DJ, Gore SM, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR (1992). Quality of life measures in health care. III: resource allocation. *BMJ*;**305**:1205–9.

Spilker B (1992). Standardisation of quality of life trials. An industry perspective. *PharmacoEconomics*;**1**:73–5.

Spilker B (1996). Quality of life and pharmacoeconomics in clinical trials. 2nd edn. Philadelphia: Lippincott-Raven Publishers.

Spitzer WO, Dobson AJ, Hall JA, Chesterman E, Levi J, Shepherd R, *et al* (1981). Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. *J Chronic Dis*;**34**:585–97.

Spitzer WO (1987). State of science 1986: quality of life and functional status as target variables for research. *J Chronic Dis*;**40**:465–71.

Sprangers MA, Aaronson NK (1992). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. *J Clin Epidemiol*;**45**:743–60.

Sprangers MA, Cull A, Bjordal K, Groenvold M, Aaronson NK (1993). The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. *Qual Life Res*;**2**:287–95.

Sprangers MA, Groenvold M, Arraras JI, Franklin J, te Velde A, Muller M, *et al* (1996). The European Organization for Research and Treatment of Cancer breast cancer-specific quality-of-life questionnaire module: first results from a three-country field study. *J Clin Oncology*;**14**:2756–68.

Stalfelt AM (1994). Quality of life during induction treatment of acute myeloid leukaemia. A comparison of three intensive chemotherapy regimens using three instruments for quality of life assessment. *Acta Oncol*;**33**:477–85.

Staquet M, Berzon R, Osoba D, Machin D (1996). Guidelines for reporting results of quality of life assessments in clinical trials. *Qual Life Res*;**5**:496–502.

Stewart AL (1992). Conceptual and methodologic issues in defining quality of life: state of the art. *Progress in Cardiovascular Nursing*;**7**:3–11.

Street RL, Gold WR, McDowell T (1994). Using health status surveys in medical consultations. *Med Care*;**32**:732–44.

Streiner DL, Norman GR (1995). Health measurement scales: a practical guide to their development and use. second edition. 2nd edn. Oxford: Oxford University Press.

Stucki G, Liang MH, Fossel AH, Katz JN (1995). Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol*;**48**:1369–78.

Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH (1996). Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol*;**49**:711–17.

Sullivan LM, Dukes KA, Harris L, Dittus RS, Greenfield S, Kaplan SH (1995). A comparison of various methods of collecting self-reported health outcomes data among low-income and minority patients. *Med Care*;**33**:AS183–94.

Sullivan M, Ahlmen M, Bjelle A (1990). Health status assessment in rheumatoid arthritis. I. Further work on the validity of the Sickness Impact Profile. *J Rheumatol*;**17**:439–47.

Sutherland HJ, Dunn V, Boyd NF (1983). Measurement of values for states of health with linear analog scales. *Med Decis Making*;**3**:477–87.

Sutherland HJ, Lockwood GA, Boyd NF (1990). Ratings of the importance of quality of life variables: therapeutic implications for patients with metastatic breast cancer. *J Clin Epidemiol*;**43**:661–6.

Sutherland HJ, Till JE (1993). Quality of life assessments and levels of decision making: differentiating objectives. *Qual Life Res*;**2**:297–303.

Tarlov AR, Ware J, Greenfield S, Nelson EC, Perrin E, Zubkoff M (1989). The Medical Outcomes Study. An application of methods for monitoring the results of medical care. *JAMA*;**262**:925–30.

Taylor KM, Macdonald KG, Bezjak A, Ng P, DePetrillo AD (1996). Physicians' perspective on quality of life: an exploratory study of oncologists. *Qual Life Res*;**5**:5–14.

Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA (1996). Are we making the most of the Stanford Health Assessment Questionnaire? *Br J Rheumatol*;**35**:574–8.

Testa MA, Anderson RB, Nackley JF, Hollenberg NK (1993). Quality of life and antihypertensive therapy in men. A comparison of captopril with enalapril. The Quality-of-Life Hypertension Study Group. *N Engl J Med*;**328**:907–13.

Testa MA, Simonson DC (1996). Assessment of quality-of-life outcomes. *N Engl J Med*;**334**:835–40.

The EuroQol Group (1992). EuroQol – a reply and reminder. *Health Policy*;**20**:329–32.

Till JE, Sutherland HJ, Meslin EM (1992). Is there a role for preference assessments in research on quality of life in oncology? *Qual Life Res*;**1**:31–40.

Till JE, Osoba D, Pater JL, Young JR (1994). Research on health-related quality of life: dissemination into practical applications. *Qual Life Res*;**3**:279–83.

Torrance GW (1973). Health index and utility models: some thorny issues. *Health Serv Res*;**8**:12–14.

Torrance GW (1986). Measurement of health state utilities for economic appraisal: a review. *J Health Econ*;**5**:1–30.

Torrance GW (1987). Utility approach to measuring health-related quality of life. *J Chronic Dis*;**40**:593–603.

Torrance GW (1995). An interview on utility measurement [interview by Bernie O'Brien]. *J Rheumatol*;**22**:1200–2.

Torrance GW, Furlong W, Feeny DH, Boyle M (1995). Multi-attribute preference functions: Health Utilities Index. *PharmacoEconomics*;**7**:503–20.

Tugwell P, Bombardier C (1982). A methodologic framework for developing and selecting endpoints in clinical trials. *J Rheumatol*;**9**:758–62.

Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B (1987). The MACTAR Patient Preference Disability Questionnaire – an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheumatol*;**14**:446–51.

Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Bennett K, *et al* (1990). Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. *Arch Intern Med*;**150**:59–62.

Tugwell P, Bombardier C, Bell M, Bennett K, Bensen W, Grace E, *et al* (1991). Current quality-of-life research challenges in arthritis relevant to the issue of clinical significance. *Control Clin Trials*;**12**:217S–225S.

Tuley MR, Mulrow CD, McMahan CA (1991). Estimating and testing an index of responsiveness and the relationship of the index to power. *J Clin Epidemiol*;**44**:417–21.

van Alphen A, Halfens R, Hasman A, Imbos T (1994). Likert or Rasch? Nothing is more applicable than good theory. *J Adv Nurs*;**20**:196–201.

van Knippenberg FC, de Haes JC (1988). Measuring the quality of life of cancer patients: psychometric properties of instruments. *J Clin Epidemiol*;**41**:1043–53.

van Praag B (1993). The relativity of the welfare concept. In: The quality of life (Nussbaum M, Sen A, editors). Oxford: Clarendon Press. p. 362–85.

Veldhuyzen van Zanten SJ (1991). Quality of life as outcome measures in randomized clinical trials. An overview of three general medical journals. *Control Clin Trials*;**12**:234S–242S.

Ventegodt S (1996). Measuring the quality of life. Copenhagen: Forskningcentrets Forlag.

Visser MC, Fletcher AE, Parr G, Simpson A, Bulpitt CJ (1994). A comparison of three quality of life instruments in subjects with angina pectoris: the Sickness Impact Profile, the Nottingham Health Profile and the Quality of Well Being Scale. *J Clin Epidemiol*;**47**:157–63.

von Neumann J, Morgernstern O (1953). The theory of games and economic behaviour. 3rd edn. Princeton: Princeton University Press.

Wade DT (1988). Measurement in rehabilitation. *Age Ageing*;**17**:289–92.

Wade DT (1992). Measurement in neurological rehabilitation. Oxford: Oxford University Press.

Ware J (1984). Methodological considerations in the selection of health status assessment procedures. In: Assessment of quality of life in clinical trials of cardiovascular therapies (Wenger NK, Mattson ME, Furberg CD, Elinson J, editors). New York: LeJacq Publishing Inc. p. 87–111.

Ware J (1987). Standards for validating health measures: definition and content. *J Chronic Dis*;**40**:473–80.

Ware J (1995). The status of health assessment 1994. *Ann Rev Public Health*;**16**:327–54.

Ware J, Sherbourne CD (1992). The MOS 36-item short-form health survey (SF–36). I. Conceptual framework and item selection. *Med Care*;**30**:473–83.

Ware J, Brook RH, Davies AR, Lohr KN (1981). Choosing measures of health status for individuals in general populations. *Am J Public Health*;**71**:620–5.

Ware J, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A (1995). Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. *Med Care*;**33**:AS264–79.

Ware J, Bayliss MS, Rogers WH, Kosinski M, Tarlov AR (1996a). Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study. *JAMA*;**276**:1039–47.

Ware J, Kosinski M, Keller SD (1996b). A 12-Item Short Form Health Survey. Construction of scales and preliminary tests of reliability and validity. *Med Care*;**34**:220–33.

Weinberger M, Nagle B, Hanlon JT, Samsa GP, Schmader K, Landsman PB, *et al* (1994). Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *J Am Geriatr Soc*;**42**:1295–9.

Weinberger M, Oddone EZ, Samsa GP, Landsman PB (1996). Are health-related quality-of-life measures affected by the mode of administration? *J Clin Epidemiol*;**49**:135–40.

Wells KB, Stewart AL, Hays RD, Burnam MA, Rogers WH, Daniels M, *et al* (1989). The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *JAMA*;**262**:914–19.

Weymark JA (1991). Reconsideration of the Harsanyi-Sen debate on utilitarianism. In: Interpersonal comparisons of well-being (Elster J, Roemer J, editors). Cambridge: Cambridge University Press. p. 225–320.

WHO (1947). Constitution. WHO Chronicle 1–29.

WHO (1980). International classification of impairments. Disabilities and handicaps. (Abstract)

WHOQOL Group (1993). Study protocol for the World Health Organization project to develop a quality of life assessment instrument (WHOQOL). *Qual Life Res*;**2**:153–9.

WHOQOL Group (1998). The World Health Organization quality of life assessment (WHOQOL): development and general psychometric properties. *Soc Sci Med*;**46**:1569–86.

Wiklund I, Karlberg J (1991). Evaluation of quality of life in clinical trials. Selecting quality-of-life measures. *Control Clin Trials*;**12**:204S–216S.

Wiklund I, Dimenas E, Wahl M (1990). Factors of importance when evaluating quality of life in clinical trials. *Control Clin Trials*;**11**:169–79.

Wilkin D, Hallam L, Doggett M (1993). Measures of need and outcome for primary health care. 2nd edn. Oxford: Oxford University Press.

Williams A (1988). Do we really need to measure the quality of life? [editorial]. *Br J Hosp Med*;**39**:181.

Williams JI, Naylor CD (1992). How should health status measures be assessed? Cautionary notes on procrustean frameworks. *J Clin Epidemiol*;**45**:1347–51.

Williams R (1983). Disability as a health indicator. In: Health indicators (Culyer A, editor). Oxford: Martin Robertson. p. 150–64.

Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF (1988). The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis. *J Rheumatol*;**15**:1480–8.

Wolfe F, Pincus T (1991). Standard self-report questionnaires in routine clinical and research practice – an opportunity for patients and rheumatologists [editorial]. *J Rheumatol*;**18**:643–6.

Wright JG, Feinstein AR (1992). A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol*;**45**:1201–18.

Wright JG, Young NL (1997). A comparison of different indices of responsiveness. *J Clin Epidemiol*;**50**:239–46.

Yancik R, Yate JW (1986). Quality-of-life assessment of cancer patients: conceptual and methodologic challenges and constraints. *Cancer Bulletin*;**38**:217–22.

Young JR, Chamberlain MA (1987). The contribution of the Stanford Health Assessment Questionnaire in rheumatology clinics. *Clinical Rehabilitation*;**1**:97–100.

Ziebland S, Fitzpatrick R, Jenkinson C (1993). Tacit models of disability underlying health status instruments. *Soc Sci Med*;**37**:69–75.

Zwinderman AH (1990). The measurement of change of quality of life in clinical trials. *Stat Med*;**9**:931–42.

Appendix 1

Method of the review

The aim of the literature review is to give a comprehensive report of the range of issues and views regarding methods of evaluating patient-based outcome measures. The review is based on a structured and extensive search of the literature. It was not, however, the purpose of the review to calculate or survey the total number of papers published on the methodology of evaluating patient-based outcome measures, nor to report the frequency with which particular views were expressed.

Intellectual mapping of the topic

The first step in the structured review of the literature was to focus the broad remit of the project by intellectual mapping of the topic, the aim being to establish central and surrounding issues and specify inclusion and exclusion criteria for the subsequent literature search. This was done by project members reviewing an in-house collection of journal articles. Overall, 94 publications were identified, of which 41 articles were used as a base set because of their emphasis on methodology of evaluating patient-based outcome measures. This joint exercise enabled the group to initiate inclusion and exclusion criteria for obtaining relevant articles, with adjustments made during collaborative project meetings. *Box 5* shows the inclusion and exclusion criteria for the literature search.

BOX 5 Inclusion and exclusion criteria for selecting articles

- *Include articles that focus on:*
  Reviews of methods of evaluating patient-based outcome measures
  Psychometric evaluation of patient-based outcome measures, i.e. responsiveness, reliability, validity, acceptability
  Practical feasibility: response rates, time to complete
  Principles of selection of patient-based outcome measures
  Patient-based outcome measures used in clinical trials
  Utility methodology
  Comparative studies of patient-based outcome measures
  Validation publications of prominent patient-based outcome measures, with specific evaluation and methodological sections
- *Exclude articles comprising only these issues:*
  Routine use of patient-based outcome measures in particular conditions/diseases
  Translation or cross-culture studies
  Clinician-based outcome measures
  Economic theory
  Validation studies of questionnaires and interviews in general

NB: If an article contained information that covered both inclusion and exclusion criteria it was included.

Main literature review

The chosen strategy for the main literature review comprised the following steps:

1. Retrospective searching
2. Handsearching of relevant journals
3. Searching of in-house database (Pro-cite) at Brunel University
4. Qualitative analysis of articles retrieved in steps one to three
5. Electronic search of databases.

It was decided to conduct the electronic search after the retrospective and hand searching for a number of reasons. Firstly, the initial in-house collection of articles provided a reference point of publications that referred to relevant articles. Secondly, the heterogeneity of the terminology used in this field required an extensive list of search terms to make the electronic search as sensitive as possible. The first stage of the literature review provided a full and comprehensive range of search terms that were then used to establish the electronic search strategy. However, a sensitive electronic search ran the risk of lacking specificity. Thus through having established a base of reviewed literature, the researcher had an up-to-date knowledge of the issues and was in a favourable position to be selective of crucial and, importantly, new publications.

Almost all relevant articles were identified by the process shown in the flow chart in *Figure 1*. The exception was those identified during hand-searching of relevant journals and the Pro-cite

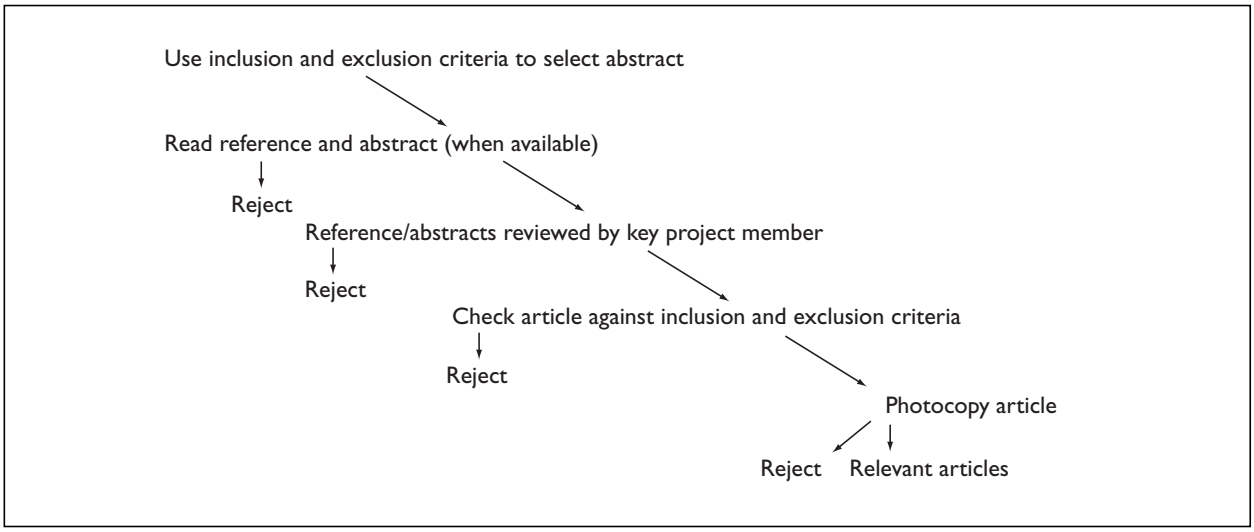


FIGURE 1 *Strategy for obtaining relevant articles*

search as they were done on site and did not offer the opportunity for further review by the key project member.

Retrospective searching and handsearching (steps 1–3)

This first stage of the literature review used the 41 'base set' articles for the retrospective search as they provided a targeted set of publications. Handsearches were conducted in journals that focused on methodology of patient-based outcome measures. The journals included *Quality of Life Research* (1992–1996), *Medical Care* (1992–1996), *Medical Decision Making* (1994–1996), *Controlled Clinical Trials* (1990–1996), *Journal of Clinical Epidemiology* (1990–1996). Additional journals were suggested for handsearching during a collaborative project meeting for their economic perspective and focus on methodology, these included *Health Economics* (1992–1995), *Health Policy* (1991–1996) and *Pharmacoeconomics* (1992–1996 and supplements).

Additionally, the Health Economics Research Group at Brunel University made available their in-house database (Pro-cite) containing approximately 9000 articles. The search term 'quality of life' was used, connected with 'utilit*', 'preference*', 'psychometri*' or 'clinical trial', to retrieve articles.

Results from the first stage of the literature review (steps 1–3)

The main components of the first stage of the literature review consisted of retrospective and hand searching and identified 153 relevant articles. The break down of this number is provide in *Table 4*.

TABLE 4 *Number of articles identified at the first stage*

Method	No. of abstracts reviewed	No. reviewed by key member	No. skim read	No. of articles photo-copied
Retrospective search	210	190 (53 rejected)	137 (70 rejected)	67
Handsearch	N/A	N/A	All	59
Pro-cite	91	N/A	27	7
Serendipitous	N/A	N/A	N/A	20
Total				**153**

Qualitative analysis (step 4)

A first draft version of part two of the report ('How to select a patient-based outcome measure'), written by the lead member of the collaborative group, provided a platform for the fourth phase of the process. This stage provided statements on the different desirable properties of outcome measures as well as indications of how such criteria may be defined and judged. It also provided definitions, uncovered issues and opinions and provided supporting evidence and references for the report.

This stage involved reading the articles retrieved in the first round of the literature search and conducting a qualitative analysis. Qualitative analysis involves indexing textual information to ensure

that nothing relevant was lost to subsequent examination. Time restrictions curtailed the total number of articles reviewed to 209, with a decision taken one month into reading, to time order the remaining publications, giving precedence to articles published after 1990. However, all 247 articles were skim read and important articles published in the 1980s and 1970s were transcribed. The qualitative analysis involved transcribing key points and summary statements from the 209 articles into files under the following topics: general issues and concepts, selection criteria, validity, reliability, responsiveness, acceptability, feasibility, utility, comparison of instruments, numerical properties and weights.

An additional benefit of this stage was the identification of an extensive list of possible search terms for the next phase of the literature search. Possible search terms were sought from the full text of articles and not just keywords.

Electronic searching (step 5)

The electronic literature search was used to achieve two objectives. It would validate the first phase of the literature review strategy, by cross checking how many of the 223 articles (24 of the 247 references were either books, book chapters or unpublished) previously obtained, appeared in the results of the electronic search. The inspection revealed that 58% (130/223) of articles were represented in the electronic literature search. This low figure is consistent with Chalmers *et al* (1992), who found that MEDLINE searching only retrieved half of the relevant studies, with those missed actually contained within MEDLINE but inaccurately indexed or described by the author or the coding procedure. More importantly, the electronic search provided a substantial number of abstracts to review in order to identify any publications that provided new dimensions or additional perspectives to issues already uncovered in the first stage of the literature review.

The electronic search did not initially have any date restrictions and went back as far as the databases would allow. However, the actual review of abstracts was limited to 1991–1996 in order to capture only up to date information. The electronic literature search was limited to English introducing a selective bias. An attempt was made to reduce the total number of records retrieved by only searching in title and keywords but this was found to be too narrow and risked missing many references.

The electronic search was carried out in MEDLINE (1966–1996), CINAHL (Cumulative Index of Nursing and Allied Health, 1982–1996), PsychLIT (1974–1996), Sociofile (1974–1996), Econlit (1969–1996), all of which were on the University of Oxford's electronic reference library (ERL) and accessed using Winspirs/Silverplatter software via the University network. Additionally, the EMBASE database (1990–1996) was accessed using the Bath Information & Data Services (BIDS). As the BIDS and the ERL databases are assessed via two different pathways it was not possible to perform a combined search.

The electronic search strategy combined the term 'patient-based outcome*' and its synonyms with related methodological terms to retrieve only publications that looked at methods of evaluating patient-based outcome measures. The search terms used after refinement of an original set are shown in *Box 6*. The terms in search one were combined using the 'or' connector and the same done for search two. The results of the two searches were then combined using the 'and' connector.

BOX 6 Electronic search terms

Search one: retrieval of all records using the terms

Patient-based outcome*	
Health status	Subjective health status
Health status indicator*	Health status assessment
Quality of life	Disability scale
Health-related quality of life	Performance status
Functional status	

Search two: retrieval of all records using the terms

Methodol*	Effect size
Psychometric*	Sensitivity to change
Validity	Reproducibility
Reliability	Acceptability
Responsiveness	Utility measure*

Combine search one and two with 'and' connector
** = truncation symbol*

The original list of search terms was run in the MEDLINE database and refined by eliminating terms that retrieved a high number of false-positives. This was done by reviewing a sample of the records retrieved from individual search terms and estimating the number of false hits. The exclusion of terms was then discussed and verified at project meetings. A summary of the excluded terms is shown in *Table 5*.

TABLE 5 *Excluded search terms*

Search term	Records retrieved	Reason for exclusion refined
Synonyms of patient-based outcomes		
Outcome measure	1976	Records retrieved included 'primary outcome measure' and 'main outcome measures' that are included in most abstracts.
Outcome research	249	
Questionnaire	63,152	Would pick up references that referred to questionnaires outside the context of QoL.
Synonyms related to methodology		
Method	12,445,500	All terms too broad.
Sensitivity	175,416	
Selection	58,330	

All combinations of the search terms were retrieved without the use of additional hyphenated terms, as the above search terms were the same as the relevant MeSH headings and the search was conducted in free text. For instance quality of life also retrieved articles with the terms quality-of-life, quality-of life as shown in the example below:

- MEDLINE EXPRESS (R) 1/96-8/96
- TI: Assessment of **quality-of-life** outcomes.
- AU: Testa-MA; Simonson-DC
- SO: N-Engl-J-Med. 1996 Mar 28; 334(13): 835–840
- MeSH: Data-Collection-methods; **Health-Status-Indicators;** Models,-Theoretical; **Reproducibility**-of-Results; Sensitivity-and-Specificity
- MeSH: *Health-Services-Research-methods; *Outcome-Assessment-Health-Care; *Quality-of-Life

An example of the problems incurred in choosing appropriate search terms is illustrated by the citation 'Deyo *et al.*, The significance of treatment effects: the clinical perspective. *Med Care* 1995;**33**: AS286–91'. In such cases, if the example did not include an abstract or appropriate keywords, it was unlikely to be retrieved.

Results of the second stage of the literature review (step 5)

The total number of records retrieved from the electronic search across all databases and encompassing a full range of years, as allowed by the databases, was 3813 records. This figure was then limited to publications between 1991 and 1996, resulting in the retrieval of 2613 records. This figure is slightly lower than the sum of all the database as listed in *Table 5* as the duplicates records were eliminated. An additional search was conducted in BIDS/EMBASE (1990–1996) using the same search terms and retrieved a total of 2935 records. As previously explained, the results of the BIDS search could not be incorporated with the other databases allowing the opportunity to eliminate duplications, so they are given separately.

All records (title, abstract and keywords) from each individual database was downloaded into Microsoft Word for Windows version 6.0 and the abstracts between 1991 and 1996 reviewed in order to add fresh publications that would provide new arguments/dimensions to the qualitative analysis phase. Relevant articles were selected by the process described in *Figure 1*. A total of 48 relevant articles was obtained as a result of the electronic search. The breakdown of this number is provide in *Table 6*.

Drafting of the review

The review has been drafted over four stages. A preparatory draft of section three 'How to select a patient-based outcome measure' provided a framework for the qualitative analysis (step 4). The information generated from the qualitative analysis was then incorporated into a second draft. The third version of the review was produced in light of the articles obtained from the electronic literature search, additional in-house references (82) and other articles obtained by word of mouth (14). A total of 391 references was used to produce the third draft of the review that was sent out for consultation. The final copy of the review incorporates comments from the ten external expert reviewers (listed in acknowledgements) and their suggested references.

Process of consulting experts

An initial list of 25 experts was compiled by the project members. Quota sampling was then used to reduce the number to 10, ensuring a coverage of relevant disciplines and a mix of clinicians, methodologists and trialists. The final list of 10 included individuals with expertise in economics, psychology, sociology, statistics, clinical medicine,

TABLE 6 *Number of articles identified at the second stage of the literature review*

Database	No. of abstracts	No. of abstracts reviewed	No. of articles skim read	No. of articles photocopied by collaborative member
MEDLINE	1367	89	22	21
CINAHL	851	15	5	4
PsychLIT	323	26	8	7
Sociofile	119	6	2	2
Econlit	26	0	0	0
Total	**2686**	**136**	**28**	**34**
Additional search				
BIDS (EMBASE)	2935	63	16	14
Final total	**5621**	**199**	**43**	**48**

health services research and clinical trials. The experts were sent the document accompanied by a feedback form with both unstructured and structured sections in order to obtain unbiased as well as standardised information. They were also encouraged to provide detailed comments throughout the manuscript.

The ten reviewers' comments were assessed independently by all four members of the collaborating group members and action points abstracted for discussion at a group meeting. Comments from reviewers and from the collaborating group were therefore as far as possible taken account of in the further draft of this document.

Health Technology Assessment panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Chair: Professor John Farndon, University of Bristol †

Professor Senga Bond, University of Newcastle-upon-Tyne †

Professor Ian Cameron, Southeast Thames Regional Health Authority

Ms Lynne Clemence, Mid-Kent Health Care Trust †

Professor Francis Creed, University of Manchester †

Professor Cam Donaldson, University of Aberdeen

Mr John Dunning, Papworth Hospital, Cambridge †

Professor Richard Ellis, St James's University Hospital, Leeds

Mr Leonard Fenwick, Freeman Group of Hospitals, Newcastle-upon-Tyne †

Professor David Field, Leicester Royal Infirmary †

Ms Grace Gibbs, West Middlesex University Hospital NHS Trust †

Dr Neville Goodman, Southmead Hospital Services Trust, Bristol †

Professor Mark P Haggard, MRC †

Mr Ian Hammond, Bedford & Shires Health & Care NHS Trust

Professor Adrian Harris, Churchill Hospital, Oxford

Professor Robert Hawkins, University of Bristol †

Dr Gwyneth Lewis, Department of Health †

Dr Chris McCall, General Practitioner, Dorset †

Professor Alan McGregor, St Thomas's Hospital, London

Mrs Wilma MacPherson, St Thomas's & Guy's Hospitals, London

Professor Jon Nicholl, University of Sheffield †

Professor John Norman, University of Southampton

Dr John Pounsford, Frenchay Hospital, Bristol †

Professor Gordon Stirrat, St Michael's Hospital, Bristol

Professor Michael Sheppard, Queen Elizabeth Hospital, Birmingham †

Dr William Tarnow-Mordi, University of Dundee

Professor Kenneth Taylor, Hammersmith Hospital, London

Diagnostics and Imaging Panel

Chair: Professor Mike Smith, University of Leeds †

Professor Michael Maisey, Guy's & St Thomas's Hospitals, London *

Professor Andrew Adam, UMDS, London †

Dr Pat Cooke, RDRD, Trent Regional Health Authority

Ms Julia Davison, St Bartholomew's Hospital, London †

Professor Adrian Dixon, University of Cambridge †

Mr Steve Ebdon-Jackson, Department of Health †

Professor MA Ferguson-Smith, University of Cambridge †

Dr Mansel Hacney, University of Manchester

Professor Sean Hilton, St George's Hospital Medical School, London

Mr John Hutton, MEDTAP International Inc., London

Professor Donald Jeffries, St Bartholomew's Hospital, London †

Dr Andrew Moore, Editor, *Bandolier* †

Professor Chris Price, London Hospital Medical School †

Dr Ian Reynolds, Nottingham Health Authority

Professor Colin Roberts, University of Wales College of Medicine

Miss Annette Sergeant, Chase Farm Hospital, Enfield

Professor John Stuart, University of Birmingham

Dr Ala Szczepura, University of Warwick †

Mr Stephen Thornton, Cambridge & Huntingdon Health Commission

Dr Gillian Vivian, Royal Cornwall Hospitals Trust †

Dr Jo Walsworth-Bell, South Staffordshire Health Authority †

Dr Greg Warner, General Practitioner, Hampshire †

Methodology Panel

Chair: Professor Martin Buxton, Brunel University †

Professor Anthony Culyer, University of York *

Dr Doug Altman, Institute of Health Sciences, Oxford †

Professor Michael Baum, Royal Marsden Hospital

Professor Nick Black, London School of Hygiene & Tropical Medicine †

Professor Ann Bowling, University College London Medical School †

Dr Rory Collins, University of Oxford

Professor George Davey-Smith, University of Bristol

Dr Vikki Entwistle, University of Aberdeen †

Professor Ray Fitzpatrick, University of Oxford †

Professor Stephen Frankel, University of Bristol

Dr Stephen Harrison, University of Leeds

Mr John Henderson, Department of Health †

Mr Philip Hewitson, Leeds FHSA

Professor Richard Lilford, Regional Director, R&D, West Midlands †

Mr Nick Mays, King's Fund, London †

Professor Ian Russell, University of York †

Professor David Sackett, Centre for Evidence Based Medicine, Oxford †

Dr Maurice Slevin, St Bartholomew's Hospital, London

Dr David Spiegelhalter, Institute of Public Health, Cambridge †

Professor Charles Warlow, Western General Hospital, Edinburgh †

* Previous Chair
† Current members

continued

continued

Pharmaceutical Panel

Chair: Professor Tom Walley, University of Liverpool †

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford †

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Tim Elliott, Department of Health †

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Dr Felicity Gabbay, Transcrip Ltd †

Dr Alistair Gray, Health Economics Research Unit, University of Oxford †

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Professor Trevor Jones, ABPI, London †

Ms Sally Knight, Lister Hospital, Stevenage †

Dr Andrew Mortimore, Southampton & SW Hants Health Authority †

Mr Nigel Offen, Essex Rivers Healthcare, Colchester †

Dr John Posnett, University of York

Mrs Marianne Rigge, The College of Health, London †

Mr Simon Robbins, Camden & Islington Health Authority, London †

Dr Frances Rotblat, Medicines Control Agency †

Mrs Katrina Simister, Liverpool Health Authority †

Dr Ross Taylor, University of Aberdeen †

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield †

Population Screening Panel

Chair: Professor Sir John Grimley Evans, Radcliffe Infirmary, Oxford †

Dr Sheila Adam, Department of Health*

Ms Stella Burnside, Altnagelvin Hospitals Trust, Londonderry †

Dr Carol Dezateux, Institute of Child Health, London †

Dr Anne Dixon Brown, NHS Executive, Anglia & Oxford †

Professor Dian Donnai, St Mary's Hospital, Manchester †

Dr Tom Fahey, University of Bristol †

Mrs Gillian Fletcher, National Childbirth Trust †

Professor George Freeman, Charing Cross & Westminster Medical School, London

Dr Mike Gill, Brent & Harrow Health Authority †

Dr JA Muir Gray, RDRD, Anglia & Oxford RO †

Dr Anne Ludbrook, University of Aberdeen †

Professor Alexander Markham, St James's University Hospital, Leeds †

Professor Theresa Marteau, UMDS, London

Dr Ann McPherson, General Practitioner, Oxford †

Professor Catherine Peckham, Institute of Child Health, London

Dr Connie Smith, Parkside NHS Trust, London

Dr Sarah Stewart-Brown, University of Oxford †

Ms Polly Toynbee, Journalist †

Professor Nick Wald, University of London †

Professor Ciaran Woodman, Centre for Cancer Epidemiology, Manchester

Primary and Community Care Panel

Chair: Dr John Tripp, Royal Devon & Exeter Healthcare NHS Trust †

Professor Angela Coulter, King's Fund, London *

Professor Martin Roland, University of Manchester *

Dr Simon Allison, University of Nottingham

Mr Kevin Barton, East London & City Health Authority †

Professor John Bond, University of Newcastle-upon-Tyne †

Ms Judith Brodie, Age Concern, London †

Dr Nicky Cullum, University of York †

Professor Shah Ebrahim, Royal Free Hospital, London

Mr Andrew Farmer, Institute of Health Sciences, Oxford †

Ms Cathy Gritzner, The King's Fund †

Professor Andrew Haines, RDRD, North Thames Regional Health Authority

Dr Nicholas Hicks, Oxfordshire Health Authority †

Professor Richard Hobbs, University of Birmingham †

Professor Allen Hutchinson, University of Sheffield †

Mr Edward Jones, Rochdale FHSA

Professor Roger Jones, UMDS, London

Mr Lionel Joyce, Chief Executive, Newcastle City Health NHS Trust

Professor Martin Knapp, London School of Economics & Political Science

Dr Phillip Leech, Department of Health †

Professor Karen Luker, University of Liverpool

Professor David Mant, NHS Executive South & West †

Dr Fiona Moss, Thames Postgraduate Medical and Dental Education †

Professor Dianne Newham, King's College London

Professor Gillian Parker, University of Leicester †

Dr Robert Peveler, University of Southampton †

Dr Mary Renfrew, University of Oxford

Ms Hilary Scott, Tower Hamlets Healthcare NHS Trust, London †

* Previous Chair
† Current members

Health Technology Assessment 1998; Vol. 2: No. 14

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

ISSN 1366-5278