Quality-of-life measurements in the evaluation of treatment: proposed guidelines

ASTRID FLETCHER

Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

Introduction

In recent years a dramatic increase has occurred in the number of studies in which claims have been made for treatments, based on their benefits, or otherwise, on quality of life. Unfamiliarity with the methodological issues involved, in particular, interpretation of the importance of treatment-related changes in quality of life, make a full appraisal of such studies inaccessible to many health-care workers. Moreover, because many of these studies are sponsored by the pharmaceutical industry, there is reasonable scepticism that 'quality of life' is yet another marketing ploy.

If quality-of-life measurement is to make a serious contribution to the evaluation of treatments, scientifically acceptable studies are required. Guidelines are needed to set standards and define the criteria by which quality-of-life studies should be assessed. Guidelines would assist regulatory authorities, purchasers, journal editors, individual clinicians, and health workers; enhance the quality of studies; and lead to a better informed research community. They would also be used to assess promotional materials from the pharmaceutical industry.

The adoption of statistical guidelines represented a major step forward in improving scientific research and educating health professionals. Sufficient methodological research and experience now exist to establish a similar minimum set of guidelines for quality-of-life studies. This article is intended to suggest some basic criteria from which definitive guidelines could be drawn (Table 1).

A standard minimal definition of quality of life and its components

A need exists to standardise the terminology used in studies and to define a minimum set of concepts and dimensions of quality of life in order to justify a claim to have measured quality of life. At present, a bewildering array of terms are used to describe quality of life measurements. 'Quality of life', 'health related quality of life', 'health status', 'functional assessment' and even 'needs assessment' are terms frequently used to describe the same dimensions, and even the same

instruments. Almost identical sets of scales have been described as measures of 'need and outcome for primary health care' [1], measures of 'quality of life' [2, 3] and measures of 'health' [4]. The terms, to some extent, reflect differences both in the conceptualisation of scales, and in the uses to which they may be put. The simple term 'quality of life' has been widely adopted. The term represents an acknowledgement that health care evaluation is concerned with wider issues of relevance to the patient in addition to clinical outcomes.

Good agreement exists in the literature that the essential set of dimensions of quality of life includes physical, psychological and social functioning expressed through the subjective experience of individuals [5–8] (Table 2). Under each of these broad dimensions will be found a range of subdimensions. Claims to have measured quality of life, either in journal reports, regulatory applications, or product

Table 1 Guidelines for assessing quality of life studies to evaluate treatments

Has the study included the essential dimensions?
What was the rationale for choice of subdimensions?
How were the instruments for the study chosen?
Summary of the measurement properties of the instrument
Description of any alterations made to standard instruments
Description of any extra validation work
What was the study design?
Sample size calculations
Methods of collection of quality of life data
Training of interviewers
Quality assurance
How were the instruments scored?
Is the presentation of the raw data adequate?
At all time periods measured
Mean, standard deviation, median, range
Scatter plots
What statistical analyses have been carried out?
Are the analyses appropriate for the distributions?
Have the analyses included all randomised subjects?
What is the effect on the analyses of excluding withdrawals?
Have multiple testing issues been considered?
Is the interpretation fair?
Comparison with other studies
Presentation of effect sizes

Table 2 Essential dimensions of quality of life and examples of subdimensions*Social function*

Participation in social events
 Contact with friends or relatives
 Quality of family and social relationships
 Marital satisfaction
 Performance of usual social role (e.g. work, parent)

Physical function

Aspects of everyday living (e.g. self-care, shopping, housework, home repairs)
 Physical aspects of leisure-time pursuits or work
 Physical symptoms (e.g. pain, fatigue)

Psychological function

Mood (e.g. depression, anxiety)
 Cognition (e.g. memory, concentration)
 Overall feeling of well being

promotion, must include the essential dimensions. The rationale for choice of subdimensions should be clearly described and justified. At present, the term 'quality of life' is used indiscriminately to include proxy measures, such as side effects, or clinical parameters. A survey of published clinical trials in reputable cardiology and oncology journals found that more than 40% claiming to have measured quality of life had used such measures as frequency of hospitalization, and an additional 24% used single item descriptors of well being [9].

A further issue is whose viewpoint should be considered, and usually it is that of the patient. There may be other situations where the perspective of another close family member such as a spouse or carer may be relevant, for example if there are sexual difficulties or serious behavioural problems. We need to distinguish these circumstances from the situation where a proxy respondent has to be used if the patient, for reasons of ill health or mental competence, is unable to respond. Additionally in some trials, where the burden of the patient's condition also falls on family members, we may also be interested in measuring their quality of life.

Criteria for instrument selection*Measurement properties of instruments*

To assess the results of a study, it is important to establish the credentials of the instruments used. These include a summary of the validity, reliability, and responsiveness of the instrument from other studies plus any extra development work, including pilot studies, conducted within the present study.

Negative results may reflect the use of a poor instrument, for example, because the instrument was inappropriate for the study dimensions (low validity), produced inconsistent answers in the stable situation (low reliability), or was unable to measure changes that occurred (low responsiveness).

Reviews are available for many instruments both generally [2–4, 10], in specific disease states, such as cancer [11], in specific populations, such as the elderly

[12, 13], and in the context of primary care [1]. Clearinghouses (US Clearinghouse of Health Indexes, UK Clearing House for Information on the Assessment of Health Outcomes) and on-line computer programmes, such as OLGA [14], provide an invaluable service in updating the literature on instrument properties. Methodological issues in assessing the measurement properties of instruments are described in detail in most reviews or specialist texts [15].

Validity The criteria for establishing validity are not clear-cut but are judged from evidence on the performance of the instrument in a range of studies. Content validity is an assessment of the relevance and validity of items included in a questionnaire. It includes the testing of a scale at the development stage for comprehension and absence of ambiguity (face validity). Criterion validity is the strength of the relationship between the new instrument and some other criterion measuring the same characteristic, ideally an accepted 'gold standard'. For example, criterion validity of a new scale to measure depression would be measured from the correlation of the new scale with a standard, such as the Beck Depression Inventory, administered at the same time [16].

Construct validity is evidence that the instrument is broadly behaving as expected and shows similar (convergent validity) or opposite (divergent validity) relationships with other appropriate measurements. For example, in UK populations, SF36 scores were lower in subjects who had recently consulted their general practitioners or reported longstanding illness and differed in the expected direction by gender, age, and socioeconomic status [17]. Discriminative validity is the ability of the instrument to distinguish groups of people by their responses to the questionnaire and some other criteria, such as disease severity. For example, the Symptom Rating Test (SRT), a self rating scale of psychological distress discriminated between depressed patients and normal individuals, agreed with the results of psychiatrists' ratings using the Hamilton Depression Rating Scale [18], and correlated with the concentration of benzodiazepine measured in the blood of anxious patients [19].

Simple statements that the instruments used in a study are valid are uninformative and may be misleading. Evidence for validity of instruments may be less than optimal but should be summarized.

Reliability Reliability, sometimes also called repeatability or reproducibility, is a measure of the consistency of the performance of an instrument when used by different observers or over different time periods. Reliability should be reported using accepted statistical methods, such as Kappa, reliability coefficient, and test-retest agreement [15, 20]. Pearson correlation coefficients have commonly been used as evidence of reliability of instruments but are measures of association, not agreement [20]. Reliability between and within interviewers in a particular study should be reported.

Responsiveness The most important property of an instrument used as an outcome measure is responsive-

ness, the ability to detect changes that occur as the result of an intervention. Instrument responsiveness is optimally established in placebo controlled trials, with treatments of known efficacy, or from trials in which treatment related changes in other measurements have also occurred. Formal methods for assessing responsiveness should be reported, if available, such as effect size (see below), and correlation of instrument changes with other measures [21]. When no data on responsiveness are available, other information may be used to suggest the likely performance of instruments in detecting changes in health status. For example, instruments may have 'floor' or 'ceiling' effects, that is, most patients score at the top or the bottom of the scale and are therefore unable to measure improvements or deteriorations. Presentation of the distribution of baseline data from the same study will also allow an estimate of the likely responsiveness of the instrument.

Type of instrument

In addition to the measurement properties of instruments, researchers should report other factors influencing their choice of instruments. A large number of instruments are currently available for measuring a range of quality-of-life dimensions. These may be a single general instrument (for example, the Nottingham Health Profile) [5], instruments developed specifically for certain diseases (for example, the EORTC QL Questionnaire for cancer patients) [22], or instruments to measure certain attributes (for example, the Profile of Mood States) [23]. Advantages and disadvantages of using a single general instrument, compared with disease-specific instruments, have been discussed [24, 25]. General instruments may not cover the appropriate areas of interest in sufficient detail and their validation may be inappropriate for a particular study. Both of these factors are likely to reduce responsiveness. Conversely the use of a general instrument allows for unexpected effects by taking a broad approach, and allows for comparisons to be made across different studies and patient groups. A general recommendation is to use a general standard instrument supplemented by a disease-specific instrument [6–8, 22, 25, 26]. Extra, and possibly, less well validated questions specific for the study may also be required.

For some conditions there are a large number of potential instruments and researchers planning to use a quality-of-life instrument in a clinical trial may be overwhelmed by the apparent choice. In fact many instruments do not fulfil the criteria previously described and the choice is much narrower. Ideally, there should be a limited number of instruments to be used as standards although there is as yet no consensus on which ones. A general guide would be to use instruments that have been used in other trials in the same condition and which have been shown to be responsive to treatment.

Tampering with instruments is not recommended; apart from issues of validation, comparability and interpretation in the context of other studies are compromised. It is generally acceptable to use subscales

rather than the full instrument if more appropriate for a particular study, but pulling out individual questions to create a new scale would require extensive additional validation. Researchers should report any alterations they have made to existing instruments and subsequent validation studies.

Many instruments have been developed in the United States and are used in other countries and language groups. Studies should report the methods of questionnaire translation and establishment of content validity, to ensure that appropriate and relevant items are included. Instruments should not be used in individuals or groups whose culture is significantly different from the country in which the questionnaire was developed, unless considerable evidence of new validation exists.

Study design

The usual standards for assessment of the study design must be reported including details of randomisation, inclusion and exclusion criteria, and treatment schedules. Sample size calculations and practical aspects of collection of quality-of-life data deserve particular attention.

Sample size

Sample size calculations must specify the power of the study to detect a stated treatment difference in a particular quality-of-life dimension at a given level of significance. This involves deciding in advance the most important aspect of quality of life for the patients in the study that is expected to be modified by treatment, and the size of the difference that it would be undesirable to miss and would be meaningful to patients. The rationale for choosing the expected treatment difference should be described.

Effect size is a parameter that provides a measure of the relative importance of changes observed in a quality-of-life dimension as a result of treatment [27]. In general effect sizes are calculated as the difference between pre- and post-treatment means divided by the standard deviation between patients of the pre-treatment means. An effect size of 0.2 indicates a small effect, 0.5 a moderate effect and 0.8 or greater a large effect. Effect sizes can be used for sample size calculations and should be based on treatment differences. The power of the study to detect effects in other dimensions for the estimated sample size should also be given.

Collection of quality-of-life data

Practical details of the collection of quality-of-life data in the study should be reported in order to ascertain possible biases or deficiencies. These include the method of administration of the questionnaire (self-administered, interviewer administered), place of completion, and the timing and number of assessments both during the trial and during the run-in or baseline period. If interviewers are used, further information should be given on training details, and interobserver

and intraobserver variability (if examined). Strategies to reduce interviewer bias in open or single-blind trials should be described. Studies should also report whether attempts were made to collect data from patients who withdrew from the trial.

Methods of quality control of data should be described, including checks on the authenticity and completeness of the quality-of-life interviews and systematic differences between interviewers in patient responses.

Scoring of quality-of-life instruments

Methods for scoring quality-of-life instruments in the study should be reported. Scoring manuals are available for most instruments and should be used to facilitate cross study comparisons and interpretation. More problematic is the use of weighting schemes in which the weights have been derived in populations or patient groups not represented in the current study. The sensitivity of the analyses to changes in these weights should be explored [26].

Reporting and presentation of data

Quality-of-life data should include all time periods during which data were collected, as well as adequate information to characterize the distribution of data using typical measures, such as mean, median, standard deviation, and range [25]. Scatter plots of individual baseline and treatment scores and histograms of score distributions at baseline and on treatment aid interpretation and could be presented for the key outcome variable. Results of complex analyses, such as regression models, or using measures of effect size may *additionally* be presented, but presentation of the raw data is mandatory. Key clinical data must also be reported alongside the quality-of-life results even if there are additional manuscripts reporting clinical data in more detail. This allows for a full interpretation of the trial results. Costs data, when available, should also be presented.

Analyses

Methods of analyses should be stated in full and be appropriate to the data distribution. Intention-to-treat analyses may be problematic if data are unavailable for patients who withdrew from the trial but are preferred to per protocol analyses, because treatment withdrawals may reflect a poor quality of life on the randomised treatment. In one study that claimed benefits for an antihypertensive agent, primarily on the results from patients who completed the 4 months of the trial, 40% of patients randomised to that drug had withdrawn principally for side effects [28]. Intention-to-treat analyses revealed no between-drug differences.

Consideration should be given to multiple testing issues in studies with many dimensions and significance tests reported only for predefined key variables identified in the planning of the study [25]. Confidence intervals should be presented for between-treatment differences. Aggregation issues are also problematic. Total scores may conceal differences in directions of effect in subdimensions. Global significance tests may be useful in dimensions with similar attributes.

Interpretation

Fair and unbiased interpretation of trial results is always essential. Two particular problems in interpreting results of quality-of-life trials deserve special attention: score improvements irrespective of treatment, and the relative importance of the observed changes. Improvement in patient scores is a well documented feature of quality-of-life trials, and has been observed in patients receiving placebo treatment and in patients in the run-in phase prior to receiving either active or placebo treatment [29]. Interpretation of trial results should be limited to relative effects between treatments, not absolute effects.

Interpretation of the relative importance of changes can be aided by such measures as effect size and confidence intervals and can be compared with results in similar patient groups in other studies. Table 3 shows effect sizes and 95% confidence intervals for between drug differences in a measure of psychological well being in three studies of antihypertensive drugs [30–32]. Small to moderate effect sizes were observed for the differences between captopril and methyldopa (0.3) and between captopril and propranolol (0.2) [30] whereas studies comparing more recent classes of drugs with angiotensin converting enzyme (ACE) inhibitors found negligible differences, with effect sizes ranging from 0.01 to 0.03 [31, 32]. The effect sizes accorded with clinical experience of these drugs.

Quality-of-life results need to be considered alongside the clinical results. This is particularly important when the benefits of either are marginal, or are in the opposite direction, for example small improvements in survival but poorer quality of life. Combining quality-of-life data with other clinical outcomes, particularly survival, in a single model is problematic. The QALY (Quality Adjusted Life Year) is an aggregation of survival and quality of life in a single index and has been proposed in combination with cost data as the method for resource decisions. Although the principle of considering both quality of life and survival is generally accepted concerns have been expressed about QALYs [33, 34]. A major problem with the QALY approach is aggregating survival and quality of life over patients. One QALY represents a range of scenarios. The treatment with the lowest cost per QALY may conceal outcomes unacceptable to patients or physicians. As a simple example consider the annual costs, survival and quality of life of two treatments (A and B) averaged over 100 patients (Table 4). The lowest cost per QALY is given by Treatment A although it has a 40%

Table 3 Differences in the Psychological General Well Being Index between antihypertensive drugs (95% confidence intervals) and effect sizes

<i>Antihypertensive drugs</i>	<i>Difference (95% CI)</i>	<i>Effect size</i>
Croog <i>et al.</i> [30]		
Captopril—Methyldopa	3.8 (not given)	0.3
Captopril—Propranolol	2.5 (not given)	0.2
Fletcher <i>et al.</i> [31]		
Cilazapril—Atenolol	0.7 (−1.9, 3.4)	0.04
Cilazapril—Nifedipine	0.5 (−2.2, 3.1)	0.03
Omvik <i>et al.</i> [32]		
Enalapril—Amlodipine	0.2 (−2.0, 2.4)	0.01

Table 4 Hypothetical example of derivation of cost per QALY for two treatments A and B

<i>Parameter</i>	<i>Treatment A</i>	<i>Treatment B</i>
Number of patients	100	100
Annual cost of treatment	£150	£200
Total cost	£15 000	£20 000
Outcome	40 die (QOL 0) 60 survive (QOL 0.5)	100 survive (40 QOL 0.3; 60 QOL 0.4)
QALYs	$(40 \times 0) + (60 \times 0.5) = 30$	$(40 \times 0.3) + (60 \times 0.4) = 36$
Cost per QALY	£500	£555

Reproduced from reference 33 with permission.

probability of death. In Treatment B all patients survive but with a reduced quality of life. It is important therefore that such facts are made explicit by presentation of all the relevant data.

This problem is compounded when QALYs are compared across different diseases with very different outcomes such as hip replacement (no survival consequences, increased quality of life), and haemodialysis (major survival consequences, reduced quality of life).

Conclusions

Although many methodological questions about quality-of-life measurement remain unanswered, sufficient agreement exists on important issues to define required standards. At present a considerable range in the quality of quality-of-life studies is seen which, with the greater dissemination of methodological reviews and experience of studies, is unacceptable.

References

- Wilkin D, Hallam L, Doggett M. *Measures of need and outcome for primary health care*. Oxford: Oxford University Press, 1992.
- Fallowfield L. *The quality of life*. London: Souvenir, 1990.
- Bowling A. *Measuring health: a review of quality of life measurement scales*. Milton Keynes: Open University Press, 1991.
- McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. Oxford: Oxford University Press, 1987.
- Hunt SM, McEwan J, McKenna SP. *Measuring health status*. London: Croom Helm, 1986.
- Guyatt G, Feeny D, Patrick D. Issues in quality of life measurements in clinical trials. *Controlled clin Trials* 1991; **12**: 81S–90S.
- Patrick DL, Erickson P. Assessing health related quality of life for clinical decision making. Ch 2, pp 11–63. In *Quality of life assessment: key issues in the 1990s*, eds Walker SR, Rosser RM. London: Kluwer Academic Publishers, 1993.
- Wiklund I, Dimenas E, Wahl M. Factors of importance when evaluating quality of life in clinical trials. *Controlled clin Trials*. 1990; **11**: 169–179.
- Schumacher M, Olschewski M, Schulgen G. Assessment of quality of life in clinical trials. *Stats Med* 1991; **10**: 1915–1930.
- Spilker B (ed) *Quality of life assessments in clinical trials*. New York: Raven Press, 1990.
- Maguire P, Selby P. Assessing the quality of life in cancer patients. *Br J Cancer* 1989; **60**: 437–440.
- Kane RA, Kane RL. *Assessing the elderly: a practical guide to measurement*. Lexington: Lexington Books, 1981.
- George LK, Bearon LB. *Quality of life in older persons: meaning and measurement*. New York: Human Sciences Press, 1980.
- Erickson P, Scott J. The On-Line Guide to Quality-of-Life Assessment (OLGA): resource for selecting quality of life assessments. Ch 13, pp 221–232. In *Quality of life assessment: key issues in the 1990s*, eds Walker SR,

- Rosser RM. London: Kluwer Academic Publishers, 1993.
- 15 Streiner DL, Norman GR. *Health Measurement Scales. A practical guide to their development and use*. Oxford: Oxford Medical Publications, 1992.
 - 16 Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. *Arch gen Psychiatry* 1961; **4**: 561–571.
 - 17 Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF36 health survey questionnaire: new outcome measure for primary care. *Br med J* 1992; **305**: 160–164.
 - 18 Fava GA, Kellner R, Munari F, Pavan L. The Hamilton Depression Rating Scale in normals and depressives. *Acta Psychiatr Scand* 1982; **66**: 26–32.
 - 19 Robin A, Curry SH, Whelpton R. Clinical and biochemical comparison of clorazepate and diazepam. *Psychol Med* 1974; **4**: 388–392.
 - 20 Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–310.
 - 21 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. *Controlled clin Trials* 1991; **12**: 142S–158S.
 - 22 Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality of life assessment in cancer clinical trials. *Rec Results Cancer Res* 1988; **111**: 231–249.
 - 23 McNair DM, Lorr M, Doppleman LF. *Manual for the Profile of Mood states*. San Diego: San Diego Educational and Industrial Testing Service, 1971.
 - 24 Fitzpatrick R, Fletcher A, Jones D, Gore S, Spiegelhalter D, Cox D. Quality of life measures in health care. I: Applications and issues in assessment. *Br med J* 1992; **305**: 1074–1077.
 - 25 Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis and interpretation. *Br med J* 1992; **305**: 1145–1148.
 - 26 Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality of life assessment: can we keep it simple? *J Roy Stat Soc* 1992; **155**: 353–393.
 - 27 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989; **27**: S178–S189.
 - 28 Testa MA, Hollenberg NK, Anderson RB, Williams GH. Assessment of quality of life by patient and spouse during antihypertensive therapy with atenolol and nifedipine gastrointestinal therapeutic system. *Am J Hypertension* 1991; **4**: 363–373.
 - 29 Fletcher A, McLoone P, Bulpitt CJ. Quality of life on angina therapy: a randomised controlled trial of transdermal glyceryl trinitrate against placebo. *Lancet* 1988; **ii**: 4–7.
 - 30 Croog SH, Levine S, Testa MA, Brown B, Bulpitt CJ, Jenkins CD, Klerman GL, Williams GH. The effects of antihypertensive therapy on the quality of life. *New Engl J Med* 1986; **314**: 1657–1664.
 - 31 Fletcher AE, Bulpitt CJ, Chase D, Collins W, Furberg CD, Goggin TK, Hewett AJ, Neiss AM. Quality of life on three antihypertensive treatments: cilazapril, atenolol, nifedipine. *Hypertension* 1992; **19**: 499–507.
 - 32 Omvik P, Thaulow E, Herland OB, Eide I, Midha R, Turner RR. Double blind parallel comparative study on quality of life during treatment with amlodipine or enalapril in mild or moderate hypertensive patients: a multicentre study. *J Hypertension* 1993; **11**: 103–113.
 - 33 Fletcher AE, Bulpitt CJ. Pharmacoeconomic evaluation of risk factors for cardiovascular disease. An epidemiological perspective. *Pharmacoeconomics* 1992; **1**: 33–44.
 - 34 Spiegelhalter D, Gore SM, Fitzpatrick R, Fletcher AE, Jones DR, Cox DR. Quality of life measures in health care. III: Resource allocation. *Br med J* 1992; **305**: 1205–1209.

(Received 22 August 1994,
accepted 16 November 1994)