Quality of life measures in health care. II: Design, analysis, and interpretation

Astrid Fletcher, Sheila Gore, David Jones, Ray Fitzpatrick, David Spiegelhalter, David Cox

The design, analysis, and interpretation of studies using measures of quality of life vary according to the context of use. In this paper we are primarily concerned with quality of life measures in clinical trials but our comments are relevant in other contexts.

Design

Apart from the usual considerations of good study design, particular issues in studies measuring quality of life are the choice of dimensions and the selection of instruments to measure these dimensions. There are also several practical considerations.

CHOOSING DIMENSIONS

The choice of dimensions is influenced by the severity and nature of the disease, the expected benefits and adverse effects of treatment, and considerations such as the length of the study, the availability of suitable instruments, and the environment in which the measurements will take place. For example, most patients with severe heart failure are elderly, retired, and physically inactive, and benefits of treatment are likely to be improvements in physical and social functioning at an already restricted level. But in trials in hypertensive patients dimensions need to be chosen that reflect potential adverse effects of treatment such as poorer work performance, problems with sexual function, and deleterious effects on mood.

SELECTION OF INSTRUMENTS

Several reviews of quality of life instruments have been published.[1-4] The first and most important issue when selecting an instrument is how well it will perform in the required situation (box 1). This can be assessed from the instrument's psychometric properties, which we discussed in greater detail in the first paper of this series.[5] In brief, validity and reliability are necessary for all contexts; the importance of other psychometric properties varies with context—for example, sensitivity and specificity are important for screening and responsiveness for clinical trials.

Another consideration is whether to use a generic or a disease specific instrument. Generic instruments cover a broad range of quality of life dimensions in a single instrument. Well known examples include the Nottingham health profile,[6] the sickness impact profile (box 2)[7] and the MOS short form general health survey.[8] Generic instruments have advantages and disadvantages. On the one hand including many health related dimensions removes the need to select dimensions for a particular study and allows for the detection of unexpected effects, on the other hand a broad approach may reduce responsiveness to effects of health care. A further benefit of generic instruments is to facilitate comparisons among different disease groups.

Another type of generic instrument is a health index in which an individual is described by a single score from a continuum, usually from 0 (death) to 1 (perfect health). A single summary score limits the clinical usefulness of health indexes, unless the score can be broken down into its components to allow identification of the areas in which change has taken place. In contrast, health profiles provide separate scores for different dimensions which, when appropriate, can also be presented as a single aggregate score across the dimensions.

Disease specific instruments have several theoretical advantages. They reduce patient burden and increase acceptability by including only relevant dimensions. This may increase responsiveness. Disadvantages are the lack of comparability of results with those from other disease groups and the possibility of missing effects in dimensions that are not included. The arthritis impact measurement scale is an example of a disease specific instrument (box 3).[9] Even when a disease specific instrument is used it may not be entirely appropriate to the particular study, perhaps because treatments with novel effects are being used. In this case additional items and questions may need to be added, although these may not have been validated. Such items should be used for exploration and preliminary psychometric testing should be included in the study.

One approach which represents a halfway house between the generic and disease specific approach is to select relevant dimension specific instruments. For example, several instruments are available for measuring psychological wellbeing: profile of mood states (box 4),[10] psychological general wellbeing index,[11] symptom rating test.[12] This approach has been used in trials of hypertension.[13-15] The same dimension specific instruments have been used in several of these trials, so allowing direct comparison of results.

A common recommendation is to include both disease specific and generic measures in a study. Ideally a few instruments should be established as standards for use in most studies, although there is as yet no consensus on which should be selected. Elsewhere we have suggested some ways to reduce the questionnaire burden on patients without loss of information in the study group.[16]

Some instruments are intended for specific populations. In particular, numerous instruments have been developed for research in elderly subjects, although many are unlikely to be useful as outcome measures.[17-19]

Researchers should be wary of using an instrument

Box 1: Factors influencing selection of instruments

- Good measurement properties:
  - Validity
  - Reliability
  - Responsiveness
- Type of instrument:
  - Generic
  - Disease specific
  - Dimension specific
  - Items specific to study
- Method of administration:
  - Self administered
  - Interviewer administered
- Cultural setting

Division of Geriatric Medicine, Department of Medicine, Royal Postgraduate Medical School, London W12 0HS
Astrid Fletcher, *senior lecturer in epidemiology*

MRC Biostatistics Unit, Cambridge CB2 2SR
Sheila Gore, *senior statistician*
David Spiegelhalter, *senior statistician*

Department of Epidemiology and Community Health, University of Leicester, Leicester
David Jones, *professor of medical statistics*

Department of Public Health and Primary Care, University of Oxford, Nuffield College, Oxford OX1 1NF
Ray Fitzpatrick, *university lecturer in medical sociology*

Nuffield College, Oxford OX1 1NF
David Cox, *warden*

Correspondence to:
Dr Fletcher, Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, London WC1E 7HT.

BMJ 1992;305:1145-8

in a cultural setting different from that in which it was developed. Apart from face or content validity, other problems include the validity of the translations and the relative importance of items in the instrument. Instruments developed in North America usually need some language modification for use in the United Kingdom. It is also possible that items of particular relevance to a group have been excluded. Mumford *et al* described the development of the Bradford somatic inventory as a screening questionnaire for psychiatric morbidity in British Asians after studies suggested that conventional screening methods were not useful because Asians somatise feelings of mental distress.[20]

PRACTICAL CONSIDERATIONS

Practical considerations include administration of the instrument and standardisation of data collection. Self administered questionnaires exclude patients who cannot read or write (for educational, cultural, or health reasons) or who may be nervous about completing a questionnaire. Questionnaires administered by interviewers avoid these problems but require extra resources for staff and for training to minimise inter-observer and intraobserver variability. The conditions under which questionnaires are completed are important; privacy and quietness should be available and confidentiality assured.

The timing of data collection also needs careful consideration. Early measurements are particularly important when studying conditions with a poor survival or large loss to follow up (for example, from side effects) while longer term measurements are also required in chronic diseases where treatment may be lifelong or alterations in lifestyle may be slow to occur. Timing of assessments should also take account of treatment cycles (as in chemotherapy) and when maximum response to treatment is expected.

Box 2: Structure of sickness impact profile—a generic instrument

Dimensions	*Number of items*	*Example*
Ambulation	12	I walk shorter distances or often stop for a rest
Mobility	10	I stay at home most of the time
Self care	23	I only get dressed with someone's help
Home management	10	I do not do heavy work around the house
Social interaction	20	I go out to visit people less often
Communication	9	I have trouble writing or typing
Emotional behaviour	9	I laugh or cry suddenly
Alertness	10	I do not keep my attention on any activity for long
Eating	9	I just pick or nibble at my food
Work	9	I am not getting as much work done as usual
Sleep and rest	7	I spend much of the day lying down to rest
Recreation	8	I spend shorter periods of time on my hobbies and recreation

Box 3: Arthritis impact measurement scale:

Dimension	*Number of items*	*Example*
Mobility	4	Do you have to stay indoors most or all of the day because of your health?
Physical activity	5	Do you have trouble bending, lifting, or stooping because of your health?
Dexterity	5	Can you easily tie a pair of shoes?
Household activity	7	If you had a kitchen could you prepare your own meals?
Social activity	4	During the past month about how often have you had friends or relatives to your home?
Activities of daily living	4	How much help do you need in getting dressed?
Pain	4	During the past month how often have you had severe pain from your arthritis?
Depression	6	During the past month how often have you been in low or very low spirits?
Anxiety	6	During the past month how much of the time have you felt tense or "high strung"?

Box 4: Example of a dimension specific instrument: the profile of mood states

Subscale	*Number of items*	*Example*
Anxiety	9	On edge
Depression	15	Gloomy
Fatigue	7	Listless
Vigour	8	Lively
Confusion	7	Muddled
Hostility	12	Bitter

The instrument uses a five point rating scale to describe feeling in the previous week

Analysis and reporting

The scoring of instruments, the multidimensionality of the data, and the relation of quality of life to patient withdrawal (including survival) deserve special attention when analysing quality of life data. We have discussed these issues in more detail elsewhere and given additional methods of analysis and components of variance.[16]

Clarity in reporting quality of life measures is particularly important as many readers will be unfamiliar with this type of data. Below we make some general recommendations (box 5).

Box 5: Steps to ensure maximum informativeness of quality of life analyses

Scoring
- Use conventional methods
- Keep weightings simple or avoid if possible
- Analyse sensitivity

Multidimensional issues
- Specify key variables before starting study
- Analyse each dimension separately
- Test for treatment-dimension interactions

Withdrawal of patients
- All subjects should complete quality of life assessments at withdrawal
- Analyse quality of life and survival separately

Clinically important effects

SCORING OF INSTRUMENTS

Scoring manuals are available for many instruments, and in general these should be used to ensure consistency across studies. It is also useful to explore the effects on the results of varying the scoring. Such sensitivity analyses provide reassurance about the stability of treatment effects, which is particularly appropriate when instruments use complex weighting schemes. Weighting schemes may not offer any advantage over simple scoring methods[21] as interpretation is more difficult and it is uncertain whether weights derived in one context are appropriate for the patients in another. Simple uniform scoring schemes are preferable in sensitivity analyses and when new questionnaires or new items are used. Thus an item with five responses on a scale from "none" to "extremely" can be scored from 0 to 4.

Aggregated dimension scores derived from individual items should also be simply expressed—for example, as the percentage out of the maximum

achievable score. This allows direct interpretation even when readers are unfamiliar with the instrument. If percentages are used it can be shown that a patient had achieved 60% of the possible score at the beginning of a trial and, say, 30% after treatment.

MULTIDIMENSIONAL ISSUES

The fact that quality of life is measured across several dimensions and is not a single outcome measure raises two particular issues in analysis. Firstly, there is the usual concern that significant effects are found simply because many variables are being tested. For example, a study using the sickness impact profile (12 dimensions) and the profile of mood states (six subscales) will have up to a 60% chance that at least one of the 18 possible test results will be significant at the 5% level. This problem is best resolved by focusing on a limited number of hypotheses that are specified in advance and estimating effects in other variables. The quality of life dimensions to be analysed should be those previously selected as the key outcomes in the study.

The second issue is whether dimensions should be combined in a summary score. This depends on whether such a score is useful and interpretable and whether there are any interactions between treatment and dimensions. For example, in antihypertensive trials some ß blockers might reduce anxiety but increase depression. Combining anxiety and depression in a global score would remove a treatment effect. It must also be remembered that many studies have low power to detect interactions, and too much reliance should not be placed on a negative interaction test.

WITHDRAWAL OF PATIENTS

Analysis by randomised group irrespective of subsequent changes (intention to treat analysis) is the method recommended for analysis of clinical trials. In quality of life trials withdrawal of patients raises particular problems. If data are unavailable on withdrawn patients the results will apply only to those who continued treatment and will ignore possible adverse effects that have caused patients to withdraw or improvements in health that have led to non-attendance. Patients who do not provide quality of life data, for whatever reason, may be different from those who do. Every effort should be made to ensure that patients complete questionnaires at withdrawal and are followed up. The strategy of reporting both a per protocol (patients who completed the trial and conformed to the protocol) and intention to treat analysis is therefore particularly appropriate in quality of life trials.

Further information on the relation between quality of life and subsequent outcome, such as death or loss to follow up, can be obtained by presenting quality of life scores at the last assessment. In a heart transplantation study, Nottingham health profile scores, broken down by status at the next follow up, suggested that patients who died or were lost to follow up had relatively poor scores at their last visit, while those who missed a visit tended to have better scores.[16]

The reporting and analysis of quality of life combined with survival presents problems. Several models have been proposed for combining quality and length of life, and technical aspects of these are considered elsewhere.[16] Some specific aspects relating to quality adjusted survival will be discussed in the last paper of this series. In general, it is preferable to report data on quality and length of life separately so that any conflict—for example, a treatment that prolongs life but increases adverse effects—is apparent.

PRESENTATION OF DATA

Presentation of results of quality of life assessments in journals is inevitably limited by space but should convey as much information about the distribution of the data as possible (box 6). Estimates of the size of treatment and other effects should be reported—for example, by confidence intervals—while results of testing should be given for only a few prespecified hypotheses.

Box 6: Presentation of results

By treatment
- Mean (SD) scores for patients
- Percentage of maximum possible score at baseline and at different assessment points
- Scatter diagrams

Treatment differences
- Means with 95% confidence intervals

Profiles
- Randomly selected typical profiles

Outcome
- Number of withdrawals
- Number of survivors

Interpretation

In studies with traditional outcomes there is usually a consensus on what constitutes a meaningful clinical effect. As yet there is no similar direct interpretation of quality of life scores, partly because of the limited experience of these measures in everyday clinical practice and clinical trials. Below we outline some general and statistical approaches to interpretation.

GENERAL APPROACHES

Comparison of treatment effects across studies with similar patient groups and instruments gives information on the relative effects of different treatments. For example, a trial in hypertension found adverse effects in psychological wellbeing in patients treated with methyldopa and propranolol compared with those treated with an angiotensin converting enzyme inhibitor.[14] In a later trial of an angiotensin converting enzyme inhibitor and newer drugs, which used the same instrument, the results were interpreted using the original study as the yardstick.[15] The 95% confidence intervals excluded effects as large as that found in the first study and it was concluded that the original benefits ascribed to the angiotensin converting enzyme inhibitor resulted from the comparison with methyldopa and propranolol rather than the advantage of the inhibitor itself. What neither trial could answer was whether or not all treatments had deleterious effects on quality of life. Absolute effects could be ascertained only by comparison with placebo, which is not feasible except in trials of mild hypertension.

Use of "population norms" to interpret effects of treatment is attractive but has pitfalls and limitations. Hunt *et al* showed that Nottingham health profile scores from community samples vary by age, gender, and socioeconomic group.[6] Comparisons between patients and "normal" subjects would need to take account of these variables, and there are probably many unidentified factors influencing scores. Moreover, most diseases will not be cured and the aim of treatment is to achieve small but important benefits in terms of symptoms, function, and prognosis.

STATISTICAL APPROACHES

Effect size may be a useful parameter for comparing scores in different studies or deciding on the relative importance of a treatment effect within a study.[22-24] A general method of calculating effect size is to divide the difference between pretreatment and posttreatment means by the standard deviation of the pretreatment mean. An alternative approach is to divide differences

in mean changes between two groups by the pooled standard deviation at baseline. A preferable measure of effect size takes account of the fact that scores vary naturally in a stable situation—for example, the within patient standard deviation obtained from two or more assessments before baseline. In this method the difference between pretreatment and posttreatment means for a treatment group is divided by this within patient standard deviation. For skewed distributions effect sizes may be expressed as the median over the interquartile range. Effect sizes can build up a picture of the meaning of changes in quality of life measures—for example, from comparison with clinical measures, in placebo controlled trials, or in trials with treatments of known efficacy. Effect sizes can act as reference values or benchmarks against which newer drugs can be assessed. Within a study the relation of treatment effects to either the within or between patient standard deviation can be used as an indicator of effects likely to be noticed by patients. For an individual patient a treatment effect of one to two times the within patient standard deviation would probably be important. When considering average treatment effects for a randomised group it is likely that effects which are larger than, say, a third of the standard deviation of the between patient scores in that treatment group would be detected by patients.

Conclusions

The potential importance of quality of life measures in health care has resulted in considerable work addressing the issues of measurement and interpretation. The results of these studies should lead to recommendations for standard instruments, both disease specific and generic; identification of areas where new instruments are needed; and in clinical trials at least, guidelines for reporting and measuring the effects of treatment.

In general, we recommend using a validated standard instrument, supplemented by dimensions specific to the study. The instrument may be disease specific or generic depending on context. Steps should be taken to ensure that the questionnaire is completed under optimum conditions. Analysis and reporting should be simple and informative and allow comparisons with other studies.

Quality of life measures have an important contribution in evaluating health care. We have given some guidance on selection and analysis of such measures. In the next paper we consider the controversial use of quality of life measures in resource allocation.

1 McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. Oxford: Oxford University Press, 1987.
2 Fallowfield L. *The quality of life*. London: Souvenir, 1990.
3 Bowling A. *Measuring health: a review of quality of life measurement scales*. Milton Keynes: Oxford University Press, 1991.
4 Wilkin D, Hallow L, Doggett MA. Measures of need and outcome for primary health care. Oxford: Oxford Medical Press, 1992.
5 Fitzpatrick R, Fletcher AE, Gore SM, Jones DR, Spiegelhalter DJ, Cox DR. Quality of life measures in health care. 1. Applications and issues in assessment. *BMJ* 1992;305:1074-7
6 Hunt SM, McEwen J, McKenna SP. *Measuring health status*. Beckenham: Croom Helm, 1986.
7 Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measurement. *Med Care* 1981;19:787-805.
8 Stewart A, Hays R, Ware J. The MOS short form general health survey: reliability and validity in a patient population. *Med Care* 1988;26:724-35.
9 Meenan R, Gertman P, Mason J, Dunaif R. The arthritis impact measurement scales: further investigation of a health status instrument. *Arth Rheum* 1982;25:1048-53.
10 McNair DM, Lorr M, Doppleman LF. *Manual for the profile of mood states*. San Diego: San Diego Educational and Industrial Testing Service, 1971.
11 Dupuy HJ. The psychological general well-being (PGWB) index. In: Wenger NK, Mattson ME, Furberg CD, Eliason J, eds. *Assessment of quality of life in clinical trials of cardiovascular therapy*. New York: Le Jacq, 1984:170-83.
12 Kellner R, Sheffield BF. A self-rating scale of distress. *Psychol Med* 1973;3: 88-100.
13 Bulpitt CJ, Fletcher AE. Measurements of quality of life in hypertension: a practical approach. *Br J Clin Pharmacol* 1990;30:353-64.
14 Croog SH, Levine S, Testa MA, Brown B, Bulpitt CJ, Jenkins CD, *et al*. The effects of antihypertensive therapy on the quality of life. *N Engl J Med* 1986;314:1657-64.
15 Fletcher AE, Bulpitt CJ, Chase D, Collins WCJ, Furberg CD, Goggin TK, *et al*. Quality of life on three antihypertensive treatments: cilazapril, atenolol, nifedipine. *Hypertension* 1992;19:499-507.
16 Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality of life assessment: can we keep it simple? *Journal of the Royal Statistical Society* 1992;155:353-93.
17 George LK, Bearon LB. *Quality of life in older persons: meaning and measurement*. New York: Human Sciences Press, 1980.
18 Kane RA, Kane RL. Assessing the elderly: a practical guide to measurement. Lexington: Lexington Books, 1981.
19 Fletcher AE, Dickinson E, Philp I. Review: audit measures: quality of life instruments for everyday use with elderly patients. *Age Ageing* 1992;21: 142-50.
20 Mumford DB, Bavington JT, Bhatnagar KS, Hussain Y, Mirza S, Naraghi MM. The Bradford somatic inventory: a multiethnic inventory of somatic symptoms reported by anxious and depressed patients in Britain and the Indo-Pakistan subcontinent. *Br J Psychiatry* 1991;158:379-86.
21 Jenkinson C, Ziebland S, Fitzpatrick R, Mowat A, Mowat A. Sensitivity to change of weighted and unweighted versions of two health status measures. *Int J Health Sci* 1991;2:189-94.
22 Guyatt G, Walter S, Norman G. Measuring change over time. Assessing the usefulness of evaluative instruments. *J Chron Dis* 1987;40:171-8.
23 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27(suppl):S178-89.
24 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clin Trials* 1991;12(suppl):142S-58S.

(Accepted 9 September 1992)

ONE HUNDRED YEARS AGO

THE FEAR OF THE LORD AND OF THE LANDLORD.

The low death-rate of London this week and last reflects in no small degree the special activity of the local authorities and of the citizens of London in adopting ordinary sanitary precautions and vigorously enforcing the provisions of the Public Health Acts. We are anticipating a similarly happy result from the unwonted activity of local sanitary boards all over the country, due to the fear of cholera. But why should this be only a spasmodic and temporary activity? Mr. Ernest Hart, in his address at Toynbee Hall on Saturday night, addressed to the local authorities and inhabitants of the East End of London, observed that if, as Whitfield preached, "Cleanliness is next to godliness," there was reason to think that the local authorities were treading with quickened footsteps in the path of godliness; but there was also reason to fear that if they were actuated in their proceedings by the fear of the Lord, they were also much influenced by the "fear of the landlord." The filthy overcrowded slums of the East and West, North and South of London, of which the proceedings this week at St. Giles's afford one out of many examples, bear testimony to this. It is so throughout the country. We give examples, taken at random, this week, from a few out of a hundred or more cases brought under our notice, in which well-known plague spots and foci of disease have been left undisturbed till the last week or two. So long as medical officers of health have no security of tenure, but are appointed at small annual salaries, dependent upon the goodwill of bodies often largely made up of persons directly or indirectly interested in bad house property, this dangerous state of things will continue. If the fear of cholera is to leave permanent traces, it should lead to a remodelling of much of our sanitary administrative system. (*BMJ* 1892;ii:651.)