Quality-of-life Assessment: Can We Keep It Simple?

By D. R. COX and R. FITZPATRICK,
Nuffield College, Oxford, UK
A. E. FLETCHER,
Royal Postgraduate Medical School, London, UK
S. M. GORE and D. J. SPIEGELHALTER and D. R. JONES†
Medical Research Council Biostatistics Unit, Cambridge, UK
University of Leicester, UK

[Read before The Royal Statistical Society on Wednesday, March 11th, 1992,
the President, Professor T. M. F. Smith, in the Chair]

SUMMARY

The importance of general statistical principles of study design and analysis to quality-of-life assessment in clinical trials is emphasized. Basic methods are reviewed briefly, with reference to three examples. Careful use of standard tools supplemented with context-specific scales is recommended. Problems of weighting and aggregation are discussed; the use of simple weighting schemes supplemented by sensitivity analysis is suggested. Some technical issues are explored, including factorial question structure, components of variance to distinguish mean treatment and patient-specific treatment effects and informative loss to follow-up. Simplicity of design, analysis and presentation are stressed.

Keywords: ANALYSIS OF VARIANCE; CLINICAL TRIALS; FACTORIAL DESIGNS; INFORMATIVE CENSORING; QUALITY OF LIFE; SENSITIVITY ANALYSIS; SURVIVAL ANALYSIS; WEIGHTING

1. INTRODUCTION

In the last decade interest has increased in quality-of-life (QOL) measures in four broad health contexts: measuring the health of populations, assessing the benefit of alternative uses of resources, comparing two or more interventions in a clinical trial and making a decision on treatment for an individual patient (Katz, 1987; Spitzer, 1987; Aaronson and Beckmann, 1987; Walker and Rosser, 1988; McDowell and Newell, 1987). Each context requires an assessment of the impact of ill health on aspects of the everyday life of the individual.

Simple measures of health status—such as the Karnofsky scale in cancer (Karnofsky and Burchenal, 1949) and the American Rheumatism Association (ARA) scale (Steinbrocker et al., 1949)—began to appear within clinical medicine in the 1940s in response to inadequacies of traditional measures of mortality and morbidity as descriptions of outcome in many chronic diseases. In the USA in the late 1960s and early 1970s, concerns about the effectiveness and economic costs of health services led the National Center for Health Services Research to support the development of measures of health status. Similar considerations in the UK have motivated the recent emphasis on an audit of clinical practice (Secretaries of State for Health, 1989).

Increasing and sometimes indiscriminate use of QOL measures, including quality-adjusted life years (QALYs), has provoked concern about these methods in the four
contexts above, especially when important consequences, such as treatment decisions or resource allocation, depend on them (Smith, 1987; Loomes and McKenzie, 1989; Carr-Hill, 1989; Fletcher, 1991). Instruments developed in one context are sometimes applied in studies with different objectives. QOL assessment in clinical trials comparing treatments differs from the resource allocation and individual treatment decision contexts. In making individual treatment decisions, for example, any QOL assessment has often been collapsed to a single scale to allow a direct comparison between options. Clinical trial results, in contrast, are best left multidimensional and reported so as to make it straightforward for future users to weigh up the alternatives with their own, currently unspecified, values.

In this paper we comment on issues of weighting and validation, make some general comments about trial design, analysis and reporting and address some more technical points such as the use of factorial designs in questionnaires, analysis of components of variance to distinguish mean treatment and patient-specific treatment effects, and the analysis of QOL in conjunction with length of survival, with particular reference to problems of informative loss to follow-up. We concentrate primarily on applications to clinical trials, although the discussion may be more widely applicable.

The current practice in the use of standard instruments in trials is reviewed briefly in the next section. Three specific examples are then introduced, to be used in Section 4 to illustrate design issues. Sections 5–7 deal with aggregation, analysis and reporting issues respectively, which, although placed in the order in which they are met in a given study, are nevertheless interdependent. Practical recommendations in each section are intended not as ‘tablets of stone’ but to emphasize a commonsense approach to design and analysis issues. In our conclusions in Section 8, we recommend the careful use of standard tools supplemented with simple context-specific scales used without elaborate weighting schemes. Analysis of variance can be used to investigate heterogeneity both among response measures and between individuals. Finally, the importance of clinical interpretability is emphasized.

2. BRIEF APPRAISAL OF USE AND PROPERTIES OF STANDARD INSTRUMENTS

Aaronson (1989) reviewed some recent examples of the use of QOL measures in clinical trials. Such measures are not universally accepted as legitimate. Although QOL could be considered as a global holistic judgment, a more reductionist approach is common. The nebulous concept is broken into components or dimensions; these dimensions may be further decomposed into a number of questions or items, which in turn may be answered on a scale. In describing or comparing groups we may want to recompose these detailed responses back into dimensions or even into a single global assessment. This process of recomposition involves the use of weighting schemes; see Section 5.

Most clinical trials have made use of standard instruments that have been to some extent evaluated by psychometric criteria. We describe these procedures only briefly, without making recommendations on the use of specific instruments. Several systematic reviews of some standard scales are available (Maguire and Selby, 1989; Spilker, 1990; Fallowfield, 1990; Bowling, 1991). Other reviews of QOL assessment in trials concentrate on specific disease or patient groups, such as cancer (see, for example, Fayers and Jones (1983)), cardiovascular disease (Fletcher, 1988) and the elderly (Kane and Kane, 1981).
2.1. Use of Standard Instruments

Many 'standard' instruments are currently available although the extent of experience of their use in clinical trials varies widely. The first and probably most widely used scale, the Karnofsky index (Karnofsky and Burchenal, 1949), measures the physical performance of patients with cancer on a scale of 0–100. Other multidimensional instruments were developed for general use on patient populations: two of the best known are the sickness impact profile (SIP) which comprises 136 questions scoring on 12 dimensions (Bergner et al., 1981) and the Nottingham health profile (NHP) which asks 38 questions on six dimensions: pain, physical mobility, energy, emotions, sleep and social contacts, and seven more general questions (Hunt et al., 1986). Many such instruments employ weighted items; the questions concerning sleep in the NHP are shown in Table 1, with the weights contributing to an overall score on that dimension.

2.2. Evaluation of Instruments

The criteria for evaluation of instruments have been the concern mainly of psychometricians (van Knippenberg and de Haes, 1988). Briefly, scales should be

(a) valid,
(b) reliable and
(c) responsive or sensitive.

2.2.1. Validity of instruments

Criterion validity involves assessing an instrument against an accepted absolute standard, which is unavailable for QOL instruments since they measure phenomena which are experiential and subjective. Other methods of evaluating a QOL instrument entail varying degrees of formality.

An assessment of face (or content) validity of an instrument involves checking whether items in an instrument appear to cover its intended topics clearly and unambiguously. It has been suggested that face validity may be maximized by including individuals of diverse backgrounds (patients, doctors, nurses, social scientists, etc.) among the assessors (Aaronson, 1989).

The construct validity of an instrument may be assessed by a more formal inspection of the overall pattern of relationships between the instrument and other measures. Thus instruments such as the NHP and the SIP have been assessed by comparing responses in groups with apparently different health statuses. There should also be
some agreement between the instrument and other measures of overlapping constructs, such as correlations with various biochemical, radiological or clinical measures of severity of disease. For example, in a study of anxious patients, higher levels of distress measured by the symptom rating test (Kellner and Sheffield, 1973) were associated with lower concentrations of benzodiazepam in the blood (Robin et al., 1974). At its most formal level this may be approached by the multitrait-multimethod analysis developed in psychometrics (Campbell and Fiske, 1959) or by factor analytic methods (Brown et al., 1984; Mason et al., 1988).

2.2.2. Reliability

The reliability of an instrument is a measure of its ability to yield the same results on repeated trials under the same conditions. Although reliability is generally considered easier than validity to test using standard methods (Nunnally, 1978; Fleiss, 1975), inappropriate methods, such as the use of correlation coefficients as measures of agreement, should be avoided (Bland and Altman, 1986; Chinn, 1990). A basic approach to the assessment of the reliability of an instrument is to examine its internal reliability at a single administration (Cronbach, 1951). Split-test reliability is examined by dividing instruments into equivalent halves for all dimensions and inspecting the degree of agreement between halves. The alternative approach of test–retest measures (Chinn and Burney, 1987) may, however, prove to be practically difficult in the context of a clinical trial. Further, if patients undergo significant clinical changes in health between the test and retest, random error and true change may be impossible to distinguish.

Reliability between interviewers or raters should also be assessed if instruments are not self-administered. High interrater reliability reported for some instruments may be the result of intensive training or special expertise not necessarily available in subsequent usage.

2.2.3. Responsiveness

Any QOL instrument used in a clinical trial should be able to detect clinically significant changes over time (Kirshner and Guyatt, 1985). Such responsiveness has been studied far less than has validity. Responsiveness of QOL instruments may be examined through associations with other changes in health status (Fitzpatrick et al., 1987) or in physiological measures (Meenan et al., 1984), or the sensitivity and specificity of instruments compared against an established criterion of change if it exists, e.g. by examination of receiver operator curves (Mackenzie et al., 1986). However, there are difficulties in deciding the criterion: should assessments by the patient, the doctor, a consensus of the two or physiological measures be used as the standard against which to judge QOL instruments (Deyo and Centor, 1986; Fitzpatrick et al., 1989)? Discrimination of treatment responses from placebo or trial participation effects (e.g. through learning effects) may be problematical, since patients will report improvements due to both.

Finally, the degree of generalizability of an instrument is important. When using an instrument in a new context, its measurement properties should be tested before the main study.
3. THREE EXAMPLES

QOL assessments arise in a range of applications; reference will be made in subsequent sections to some of the issues and the examples introduced here:

(a) a life threatening disease in children in which QOL forms a secondary outcome measure;
(b) treatment to reduce the risk from an asymptomatic condition in which delayed potential benefits are traded off against immediate but possibly long-term side-effects;
(c) conditions of chronic disability with no implications for survival.

3.1. Randomized Control Trial of Treatments for Advanced Neuroblastoma

Advanced forms of neuroblastoma (stages III and IV), a childhood solid tumour, are aggressive cancers (Evans et al., 1971), with generally very poor prognosis. They are usually treated with a cocktail of cytotoxic drugs; severe side-effects are associated with these treatment regimens. A multicentre trial of a randomized comparison of a single high dose of melphalan (requiring barrier nursing techniques) versus no further treatment following the induction therapy is reported elsewhere (Pinkerton et al., 1988). The length of survival was the primary outcome measure in this trial.

3.2. Trials of Anti-hypertensive Therapy

Hypertension is a risk factor for cardiovascular disease. Although very high levels of blood pressure may be associated with headache and dizziness, most patients are asymptomatic. The benefits from the drug treatment of hypertension, primarily in reducing strokes, have been well established, but controversy remains over the risk–benefit ratio for mild hypertension where treatment benefits may be offset by adverse effects. In general, the main issue is which treatment produces minimal interference with a patient's QOL.

3.3. Evaluation of Treatments for Arthritis and Rheumatism

Rheumatoid arthritis (RA) is a chronic disease which is rarely fatal; its main health impacts are pain, stiffness and functional limitations. Several laboratory, radiological and clinical measures are used to assess the course of the disease, but there is by no means complete agreement on which measure or combination of measures best represents the short-term or long-term outcome. Perhaps as a result, health status measures have received particular attention in rheumatology.

4. DESIGN ISSUES IN USE OF QUALITY-OF-LIFE MEASURES IN CLINICAL TRIALS

4.1. Standard Design Considerations

QOL measurements are particularly susceptible to systematic errors associated with observer effects and the conditions under which the measurement is made. Consequently, standard precautions for avoiding bias should be adopted whenever feasible. These precautions include randomization, blindness of the person administering the questionnaire and standardization of recording procedures.
4.2. What to Measure?: Choice of Quality-of-life Dimensions

What to measure in a clinical trial depends on the nature of the disease, the expected benefits and adverse effects of treatment and the length of the observation period. For example, in severe illness, patients are unlikely to be working or leading physically active lives and questions on these dimensions would not be included. An evaluation of QOL in a clinical trial can usefully be preceded by a descriptive study in the same patient population group, both to identify the QOL dimensions impaired by the disease and to suggest those on which treatment might have an effect.

For a few health problems, there are disease-specific instruments which concentrate on particularly relevant aspects of QOL. In many studies, however, the range of possible adverse effects cannot reliably be specified in advance and a wide battery of QOL scales or wider ranging generic scales may be needed (Croog et al., 1986). This may lead to patients being overburdened with inappropriate questions, and to an approach that is insensitive to small or specific but important changes.

4.2.1. Neuroblastoma trial

The QOL questionnaire for the neuroblastoma trial addressed both physical and psychological aspects of the child’s QOL. Topics included functional status (restriction of physical activity), symptoms (including pain), side-effects (nausea and vomiting, loss of appetite, difficulties in hearing), worry about side-effects (such as hair loss) and about future treatment, and overall assessment of enjoyment of life. An inclusion of additional items on social functioning of the child, and on the impact of the disease and treatment on the family and friends of the child, would have been desirable.

In some trials patients are allowed to select their own QOL objectives at the beginning of the trial, e.g. by selecting from the QOL instrument items that they consider to be of particular concern. This involves a loss of standardized information and of the ability to generalize, although it may provide a more accurate indicator of responsiveness (Tugwell et al., 1987). The ‘N=1’ trial design performed on an individual patient to decide their future treatment (Guyatt et al., 1986) represents an extreme example.

4.3. How to Measure?: Choice of Instruments

A key choice is that between the use of a standard instrument and an ad hoc or specially developed questionnaire. Standard instruments should ideally have a history of successful use, but they may not be entirely appropriate to the intended application; in particular, they may lack sensitivity compared with an especially developed instrument. Their validity and reliability should not be assumed without question (Hutchinson et al., 1979; Maguire and Selby, 1989).

Our general recommendation is to use a standard, validated core instrument, with customized additions for the particular application. The European Organization for Research on Treatment of Cancer’s modular questionnaire (Aaronson et al., 1988), for example, provides a suitable starting point for many cancer and chronic disease trials.

Standardization of technique is widely accepted for measurements such as blood pressure but is even more necessary in the QOL context because of the potentially greater variety and impact of subjective measurement errors.

The issue of the burden on the patient is also crucial; lengthy questionnaires are
an imposition on sick patients, and even on healthy ones. If it is felt that supplementary questions are appropriate, general principles of questionnaire design should be followed (Sudman and Bradburn, 1982). Wording should avoid ambiguity or bias, and attention should be given to a pleasing layout and ease of completion to encourage a good response, and to clear coding specification to allow rapid and accurate data processing. It is also vital to pilot both the questionnaires and their processing.

4.3.1. Rheumatoid arthritis studies
Several RA-specific instruments have been developed, with the health assessment questionnaire (HAQ) (Fries et al., 1982) and the arthritis impact measurement scales (Meenan et al., 1982) the most widely used. However, generic health status instruments have also been quite widely used as outcome measures in RA. The SIP has been compared with the earlier and simpler four-point ARA functional scale and found to have higher test–retest reliability and to distinguish clinically meaningful differences within a single ARA class (Deyo and Inui, 1984). This is important because the kinds of change produced by interventions are generally too small to be detected by the gross categories of the ARA functional scale.

In terms of construct validity, these instruments correlate with measures of severity of disease such as grip strength, indices of joint damage and laboratory tests (Fitzpatrick et al., 1989), although there are limitations of correlation analyses in this context. Health status instruments have detected changes from the use of non-steroidal anti-inflammatory drugs within quite short time periods such as 4–12 weeks, and that changes occur on scales such as pain and mobility rather than on scales such as social activities that might be expected to respond over longer time periods (Anderson et al., 1989; Parr et al., 1988). In patients experiencing less dramatic clinical changes, a shorter instrument (the HAQ) was as responsive to change as the SIP (Fitzpatrick et al., 1989) and an abbreviated instrument with selected items was more sensitive than the full length version of the SIP (Deyo and Centor, 1986).

4.3.2. Construction of specific scale
There are several ways of recording the required information:

(a) as a binary response, e.g. condition present or absent;
(b) as a response on a $k$-point ordinal scale representing increasing (or decreasing) severity, often $k = 3, 5, 7$;
(c) on a visual analogue scale, by marking a point on a line, often 10 cm in length, on which positions to the left or right represent increasing or decreasing severity.

For methods (b) and (c) it is desirable that, so far as is feasible, scale points should be anchored by clear verbal descriptions, preferably between categories (rather than at their centres) to facilitate assignment of borderline cases.

An advantage of the visual analogue and ordinal scales with larger values of $k$ is the protection against loss of information resulting from a bunching of responses into one or two cells. The simplicity of a scale with a limited number of points, however, will often be very appealing in implementation (McCormack et al., 1988). Firm empirical evidence of the superiority of visual analogue scales over categorical scales is difficult to find (Remington et al., 1979).
In terms of reliability, there is evidence (Lissitz and Green, 1975) that there is little to be gained by using scales comprising more than five points, but no clear advantage to the use of even or odd numbers of points (Remington et al., 1979). When typical responses are distributed around the centre of the scale, theoretical studies of grouping based on a latent normally distributed response give efficiencies of 80% and 90% for optimally placed groupings of three and five cells respectively (Cox, 1957). These results further support the widely held, if ill-documented, belief that there is no gain in going beyond 5–7 categories in such cases.

Visual analogue scales involve more labour in analysis and will be less familiar to many patients. Their main advantage lies not in the apparent gain in precision of measurement but in the possibility that they offer for appropriate post hoc specification of groupings for analysis. Loss of information due to inappropriate prior choice of categories may thus be avoided.

The discussion is different in detail if responses cluster around one end of the scale, as they may, for example, if the absence of a side-effect to which the question relates is common; the broad conclusions are similar, however. For simple methods for the analysis of visual analogue scale data of this kind see Solomon (1989).

4.3.3. Asking long series of questions: technical solution

There are some design issues that are more specific to answering a long series of questions. Here we point out some possibilities. Suppose that there are $k$ dimensions, each comprising a number of questions. The questions in each dimension are divided into two equal sections, any absolutely key question being placed in both sections. If the questionnaire is administered just once to each patient, the patient receives one section of each dimension. If, say, each patient receives the questionnaire twice, they receive one section of a particular dimension the first time and the complementary set the second time. There are now various designs that can be used associated with the fractional replication and confounding patterns in the $2^k$ factorial system, regarding the possible questions to be asked of an individual as forming the $2^k$ possible treatments in the associated factorial scheme.

There are many variants on this theme. With two questioning times per patient and a modest value of $k$ we could divide the patients into homogeneous sets of $2^{k-1}$ for allocation to a treatment and its complement. For a larger value of $k$ and one application per patient we could use a suitable fraction of the $2^k$ system as the basis.

4.4. Who and Whom to Measure?

The essence of the QOL approach is the expression of a subjective viewpoint and therefore the main respondent should always be the patient. Inability of a patient, e.g. some mentally disabled, demented elderly or terminally ill patients, to respond adequately may necessitate proxy assessment by a relative or professional. The use of standard health questionnaires such as behavioural rating scales in such circumstances involves changing from a subjective individual assessment of QOL to a social valuation assuming, for example, that deviant behaviour reflects a poor QOL.

4.4.1. Neuroblastoma trial

As the median age of children in the trial was approximately 3 years at diagnosis,
proxy assessment and reporting of QOL was generally necessary. The emphasis was on parental assessments, although a few (simple!) questions were included in the clinician’s follow-up assessment forms. The agreement between parental and clinician assessments was, where evaluable, found to be rather poor. An attempt to separate parental perception of the child’s QOL from a parental report of the child’s perception of the child’s QOL was made via both the form and the wording of questions.

4.5. Where and When to Measure?

The response to a QOL assessment may be changed by the physical context in which it takes place and the respondent’s perception of the degree to which it is under the control of, say, the treating clinician. In some (but not all) societies it may be acceptable as well as convenient to administer QOL surveys by telephone, but it remains possible that this will lead to loss of sensitivity (Tandon et al., 1989).

When to measure QOL is largely dictated by the objectives of the trial. General statistical principles apply: the need for base-line observations, a final assessment of QOL in patients who withdraw from the follow-up, avoidance of unnecessary assessments to the detriment of doctor or patient compliance and targeting assessments

(a) to distinguish early from late treatment effects,
(b) to reflect the pattern of treatment administration, e.g. cycles of chemotherapy, and
(c) to concentrate measurements when maximum treatment response is expected.

The assessment should be specific to some explicit time period (e.g. the last few days or last month). Some QOL changes may not be apparent if the follow-up period is too short since patients may take time to modify a life style adapted to a chronic disability.

5. ISSUES IN WEIGHTING AND AGGREGATION

In Section 2 we noted that the concept of QOL is generally decomposed into dimensions and items, but that even in clinical trials there is often a need for some recombination. This requires sets of weights, implicitly quantifying different states of health. In this section we raise some of the difficulties of obtaining and using weights, particularly when attempting to combine QOL and length of life in a single response measure.

5.1. How May Weights be Assigned?

Two main approaches have been used to assign weights. First, data analytic techniques such as factor analysis are intended to identify distinct ‘constructs’ that underlie the responses given by individuals to a battery of questions, and in so doing can also provide weights for questions within identified factors (Olschewski and Schumacher, 1990). Alternatively, there have been attempts to scale the states according to implicit or explicit personal valuations. Sometimes arbitrary values have been assigned to ordered sets of health states: for example, the Karnofsky index places the states at equal intervals (100, 90, 80, . . .), whereas Fanshel (1972) adopts an equally arbitrary power law (1, 1 − 2\(^{-8}\), 1 − 2\(^{-7}\), . . .).
The effect of the differential weights used in some standard instruments can be severe. For example, in the quality of well-being (Kaplan and Anderson, 1988), some individual symptoms are weighted much more highly than items on the three subscales describing physical and social activity and mobility. The weight associated with confinement to a wheelchair is 0.06, while that for cough or wheezing is 0.257, which could lead to extraordinary conclusions. In most questionnaires, however, the range of weights is not so extreme and the scores are aggregated from several items.

Torrance (1987) describes some techniques for eliciting personal valuations for states of health. The most basic is direct rating on a scale from 0 to 100, while the most sophisticated requires a simulated gamble in which a utility, in the full decision theoretic sense, is assessed. Other techniques include the time trade-off method, in which the relative values of alternative health states A and B are implicitly derived by varying the periods of time spent in these states until the subject judges them equivalent. Ratio scaling similarly derives implicit valuations by requiring the subject to weigh the equivalence of differing numbers of people being in those states. Equivalence of weights obtained by different methods in a given study is far from guaranteed (Froberg and Kane, 1989).

5.2. Critique of Weighting Methods

Consider, for example, the four states: no angina, moderate angina, severe angina and death. Suppose being symptom free were given the value 1 and being dead given value 0. Where should the other two states lie on this scale? One possibility is direct judgment of the value with no operational definition: Torrance (1987) suggests 'utilities' of 0.7 and 0.5 respectively for moderate and severe angina. Ignoring for the moment the interpersonal variation that may be encountered, we may consider how such values might be used in a clinical trial on patients with severe angina, where 'partial relief' indicates a change from severe to moderate angina.

The stated values would imply that treatments A and B (see Table 2) were equivalent, since each would yield an increase in total utility of 0.2 per patient. This 'patient-equivalent' interpretation is used as the basis of an assessment procedure by Rosser and Kind (1978). For A and B to be considered equivalent requires the assumption that between-patient variation in response is unimportant.

If serial measurements are made on the patients, we might declare treatment C also equivalent to treatments A and B. Time trade-off has been extensively used as an assessment technique by Torrance and others and forms the basis of the use of QALYs as an outcome measure combining both quality and duration of life.

<table>
<thead>
<tr>
<th>TABLE 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potentially equivalent (hypothetical) treatments in severe angina</td>
</tr>
<tr>
<td>Treatment A gave — partial relief for all patients</td>
</tr>
<tr>
<td>Treatment B gave — no relief to 60% of patients but total relief to 40% of patients</td>
</tr>
<tr>
<td>Treatment C gave — all patients total relief for 40% of the time, and no relief for the rest of the time</td>
</tr>
<tr>
<td>Treatment D gave — no pain relief, but extended life expectancy by 9.6 months</td>
</tr>
<tr>
<td>Treatment E gave — 50% of patients no relief and 1 year reduction in life expectancy, while remaining 50% experience full pain relief and life expectancy increased by 3.6 months</td>
</tr>
<tr>
<td>Treatment F gave — total relief and no change in life expectancy to 70%, but 30% die immediately</td>
</tr>
</tbody>
</table>
Suppose further that all the patients had a life expectancy of 2 years, giving each patient a mean prospect of 1 QALY (since they have severe angina with value 0.5). Then we would also consider treatments D, E and F equivalent to treatments A, B and C, since each on average provides an additional 0.4 QALYs per patient.

In practice, however, many people may not consider treatments A–F as equally desirable, even though a particular pair may be considered equivalent. This is not simply a consequence of people assigning different valuations, although this is undoubtedly also a problem. The primary issue is that there is no strong reason why values assigned on the basis of one type of aggregation, say over patients or over time, should be valid when used differently. Also, any use of clinical trial results to treat individual patients equates aggregation over patients with aggregation over personal uncertainty, in which true ‘utilities’ are relevant. Thus, for example, a future patient would need to feel that certain partial relief was equivalent to a 60% chance of no relief and a 40% chance of total relief.

The use of QALYs in resource allocation can lead to even more extreme equivalences, since different numbers of patients are balanced. Hence equally expensive interventions on the patients with severe angina described above would be considered equivalent (each giving 10 additional QALYs) if, for instance, they

(a) gave one person 20 years of extra life,
(b) gave 20 people 1 year of extra life,
(c) gave 2000 people 3.65 days of extra life each,
(d) gave 10 people total pain relief (with 2-year life expectancy) or
(e) gave 20 people total pain relief (life expectancy 2 years), but cut 1 year off life expectancy of 20 others without relief of symptoms.

It seems unlikely that any valuations could make these investments equally desirable. We conclude that it is generally inappropriate to aggregate QOL and length of life in reporting clinical trial results; some alternative approaches are discussed in Section 6.4. Moreover, since no valuations can be expected to be appropriate under the set of aggregations used in clinical trials, it is pointless to conduct elaborate experiments to derive sophisticated weighting schemes. Instead, the crucial need is for simple scales that appear reasonable (i.e. have face validity) and for sensitivity analysis of results to different valuations. Fortunately, the accepted robustness of linear scoring schemes should aid this approach (Dawes, 1979).

5.3. What Pragmatic Advice on Weighting Can be Offered?

The scoring of different items on a k-point scale can to a limited extent be regarded as a technical statistical problem in that one can examine the circumstances under which elaborate systems of weighting give heightened sensitivity over straight integer scores. More elaborate weights might be based on

(a) discriminant analysis,
(b) normal scores, for responses which are approximately symmetrically distributed around a central value, or
(c) exponential scores for responses concentrated around one end, representing the absence of a particular symptom.
In most cases, sensible design demands that the end points of the scale are not extravagantly different from neighbouring values, and provided that this is ensured simple integer scoring is likely to be enough for many purposes.

Given the general lack of a pre-eminent weighting scheme, a prudent course includes checks on the robustness of conclusions to alternative, but still arbitrary, choices of weighting schemes. We suggest that reliance on an explicit weighting scheme to yield a global QOL index should be avoided where possible. An exception may arise when a limited number of states is involved and a sensitivity analysis can be explicitly presented (Gore, 1988; Glasziou et al., 1990). Instead, we recommend the use of simple weighting schemes within dimensions, followed by the application of the analytical techniques outlined in the next section to investigate patterns of variation, together with sensitivity analyses. This approach can be extended to allow a presentation of the profile of weights representing the boundaries between preferences for the alternative treatments under comparison (Glasziou et al., 1990). More detailed recommendations for simple approaches to the analysis of QOL data are made in the following section.

6. ANALYSIS

6.1. General Approach

Special aspects of the analysis of QOL data in a clinical trial setting stem partly from the essentially multidimensional character of the concept and partly from the substantial components of variation not only between patients within treatment groups but also across time within patients. There are often the further complications of patient withdrawal and a need to consider QOL data alongside information on survival. Initially, we ignore problems associated with censoring; see, however, Section 6.4.

For instrument development, formal psychometric multivariate methods may be helpful, but for interpretation of particular trials we favour a much simpler approach. Essential features of complex serial data, for example, may be highlighted by use of the straightforward methods described by Matthews et al. (1990).

For an instrument with a fairly small number of clearly identified dimensions, each with a number of items, we suggest that cross-sectional and summary profiles through time be dealt with as follows:

(a) a simple, largely uniform, scoring scheme is adopted for the possible answers to a particular item, e.g. scores 0, 1, 2, . . . ;

(b) for a particular patient and dimension, the unweighted average over item answers is taken and expressed in standardized form, e.g. as a percentage of the maximum achievable dimension score;

(c) for an overall summary, the unweighted average of (b) over the separate dimensions is taken.

In the absence of specific a priori considerations, the separate dimension means, dimension-by-treatment means and the overall sum of the treatment means are the key values for interpretation. There are two broad strategies to the combination across dimensions. First, it may on general grounds be quite possible that the treatment effects under study are qualitatively different in the different dimensions; for example, a drug may improve anxiety but increase depression. If a summary score is required then the separate dimensions should be merged only subject to the absence of
treatment-by-dimension qualitative interaction. If it is necessary to test the significance of departures from a global null hypothesis of treatment equivalence, the multivariate character of the analysis should be recognized, e.g. by a Bonferroni-type adjustment to the most significant $p$-value (Miller, 1981).

Secondly, if it is likely that the treatments affect all dimensions similarly, and especially in a small trial, it may be sensible to put more emphasis on an overall mean across dimensions, checking, however, for treatment-by-dimension interaction, despite the likely lack of power of such checks. Such comparisons need, however, to be supplemented by checks for anomalous items and, usually more importantly, for interaction of treatment and prognostic features. Although it may be useful to look for patient characteristics at entry that affect the preferred choice of treatment, care is necessary to avoid the familiar trap of ‘data dredging’ and subgroup analysis.

For some items, e.g. the presence of and severity of particular adverse reactions, it may be wise to deal separately with

(a) the proportion of patients for whom the effect is present at all and
(b) some measure of its average importance when present, perhaps even distinguishing frequency from severity.

As already noted, the multitude of issues open to study in an investigation of typical complexity raises familiar and major issues of multiplicity. The restriction of analyses to a few key issues nominated a priori avoids the difficulties of anarchic search for ‘significant’ effects, yet may lead to important matters being overlooked. There is no simple approach without snags; we do not see the formal techniques of multiple comparisons as particularly helpful or relevant. A structured application of Bonferroni-type adjustments in the analysis of a limited number of key issues nominated in advance, with other effects thrown up in further analysis being regarded as primarily hypothesis generating, at least offers simplicity of approach.

### 6.2. Additional Role of Components of Variance

To consider analysis in a little more detail, suppose for simplicity that there is a single medically homogeneous group of patients and that on each patient a single score is considered. For the moment it is immaterial whether the score is a dimension total or an overall total, and whether or not it is weighted. Then in principle we can determine components of variance, $\sigma^2_p$ between patients, $\sigma^2_o$ between observers and $\sigma^2_{t,po}$ between times within patients within observers over a short time span. More than three variance components are involved in more complex situations.

From these variance components, correlation coefficients can be determined if required. Of these probably the most relevant is $\sigma^2_{t,po}/(\sigma^2_p + \sigma^2_{t,po})$, the inferred correlation between repeat measurements by the same observer on the same patient over a short time span. This is indeed a useful dimensionless summary statistic for a particular study but is a poor measure for comparisons because it is strongly dependent on the true variability of the patient group involved. Thus for broad comparisons and planning it is important to record separately the ‘error’ component of variance, which is likely to be relatively stable between different studies, and the component of variance between patients, which is likely to be relatively study specific.

It is important to distinguish between average treatment effects over populations of patients and the treatment effect encountered by an individual patient, particularly...
whenever a treatment-by-patient interaction is appreciable in a comparative trial. For two treatments, one crude measure is the proportion \( \pi \) of individuals showing a long run difference favouring treatment A over treatment B. If there is no treatment-by-patient interaction, \( \pi \) is either 0 or 1. All patients are then assumed to show the same treatment effect or more generally any changes in treatment effect are assumed to be captured via suitable concomitant variables. The estimation of \( \pi \) is difficult and requires assumptions that are often largely uncheckable.

The simplest case is of a two-period two-treatment crossover design. Here the second period results can be corrected for any overall period effect and the proportion of patients for whom the second treatment observation exceeds the first treatment observation found. Alternatively, if \( \hat{\tau} \) denotes an estimated treatment difference and \( \hat{\sigma}_p^2 \) the estimated variance between times within treatment groups, the normal theory estimate \( \Phi(\hat{\tau}/\hat{\sigma}_p) \) can be used. This will tend to be nearer to \( \frac{1}{2} \) than it should be because a part of \( \sigma_p^2 \) will be ‘noise’. In particular there will be a contribution from variation between times within patients within treatments. This suggests that we should correct \( \sigma_p^2 \) by subtracting an estimate \( \hat{\sigma}_e^2 \) of the noise component and therefore use the estimate \( \Phi(\hat{\tau}/\sqrt{(\hat{\sigma}_p^2 - \hat{\sigma}_e^2)}) \). There is also the design consideration that data should be collected on some if not all patients to permit the estimation of \( \sigma_p^2 \).

From a study that is not of crossover form the estimation of \( \pi \) is even more speculative. We can estimate a component of variance between patients within treatments, preferably eliminating a variation between times within patients. There are now two extreme possibilities. One is that when two different treatments are used this source of variability is sampled independently, when essentially the same arguments as for crossover designs are available. The other is that the same random component is involved for both treatments when \( \pi \) is either 0 or 1. The real situation will be intermediate but \( \pi \) cannot be explored without some information about the correlation coefficient between the two associated random components or some equivalent quantity. The only simple way of estimating this would be via crossover data obtained under broadly comparable conditions.

To estimate the proportion \( \pi \), the most satisfactory parametric route is now via a balanced analysis of variance which includes the following degrees of freedom, mean squares and expected mean squares:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>treatments (( T ))</td>
<td>patient (( P ))</td>
<td>( d_{\text{tp}} )</td>
<td>( \text{MS}_{\text{tp}} )</td>
</tr>
<tr>
<td>( T \times P )</td>
<td></td>
<td>( \sigma_p^2 + r\sigma_{\text{tp}}^2 )</td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>( d_e )</td>
<td>( \text{MS}_e )</td>
<td>( \sigma_e^2 )</td>
</tr>
</tbody>
</table>

If \( \tau \) is the overall population difference of treatment means and \( \hat{\tau} \) is the sample mean difference, the proportion of individuals with a positive true difference is \( \Phi(\xi) \), say, where \( \xi \) is estimated by \( \hat{\xi}/\sqrt{(\text{MS}_{\text{tp}} - \text{MS}_e)/r} \).

If \( \sigma_e^2 \) is negligible compared with \( \sigma_{\text{tp}}^2 \), confidence limits for \( \xi \) can be obtained via the non-central \( t \)-distribution. More generally, the best way to obtain confidence limits for \( \xi \) is to reduce by sufficiency to the consideration of the three independent random variables \( \hat{\tau}, \text{MS}_{\text{tp}} \) and \( \text{MS}_e \), to write down their likelihood, to reparameterize in terms of \( \xi, \sigma_{\text{tp}} \) and \( \sigma_e \) and thence to derive the profile likelihood for \( \xi \). With a large amount of high quality data it would be possible in principle to obtain a nonparametric
estimate. In practice the main difficulty is likely to be in separating the real treatment-by-individual interaction from noise.

6.3. Complex Weighting and Sensitivity Analysis

We have argued in favour of simple weighting schemes and methods of analysis. The arguments against the use of formal multivariate methods and associated complex weightings are essentially that an interpretation is made less direct and, most importantly, that the basis for determining weights, however well defined on formal statistical grounds, is in no way guaranteed to be meaningful either clinically or to individual patients. Further, ‘standard’ weights determined from a previous study may not be appropriate for the investigation under analysis, yet the use of non-standard weights destroys comparability between investigations. The simple procedures that we have suggested make it relatively easy to assess the effect of changes of emphasis, especially between dimensions.

If, nevertheless, ‘complex’ weights are used, sensitivity analysis is highly desirable. For this, one systematic approach is as follows. If weights \( w_1, \ldots, w_m \) are initially used for \( m \) items or dimensions, choose reasonable practical perturbations \( d_1, \ldots, d_m \). If \( m \) is reasonably small, the primary analysis can be repeated with all \( 2^m \) possible systems of weights \( w_1(1 \pm d_1), \ldots, w_m(1 \pm d_m) \).

Otherwise, a fractional replicate should be examined to allow estimation of at least the main effects of the separate perturbations. A refinement of this would recognize that the system has the structure of a mixture experiment in that equal proportional changes in all weights would have no effect on the conclusions drawn.

6.4. Combination with Other Types of Data

For a full interpretation, it may be important to combine QOL data with other kinds of information including objective measures such as the results of exercise tests, as well as physiological and biochemical measurements. Here substantive research hypotheses (Wermuth and Lauritzen, 1990), often concerning conditional independences, may arise on subject-matter grounds. A particularly important case arises when QOL data have to be combined with survival data.

6.4.1. Quality of life and survival

Patients with incomplete follow-up raise particular problems for the analysis and reporting of serial QOL data. If QOL can be summarized as a short list of health states, then it may be possible to adopt the techniques of multistate survival analysis with death as an ‘absorbing state’; transition intensities between states may, for example, be structured as a proportional hazard model including a covariate representing treatment (Kay, 1982; Andersen, 1988). However, this requires extensive data, which are rarely available, and quite strong assumptions about the nature of the dependences.

Other approaches have assumed that QOL has been placed on a single dimension between 0 and 1, and so a patient’s progress can be plotted as a profile (‘Carlen’s vitagram’) which reduces to 0 at death. The standard QALY approach takes the area under this profile as the summary response measure (Olschewski and Schumacher, 1990). Glasziou et al. (1990) emphasize that this assumes that the value of a health
state is independent of time and previous states of health, and that QOL is additive over time, thus relating to the problems of aggregation raised in Section 5.

If only a limited number of states are considered, sensitivity analyses can be used to guide the choice of treatment for a particular patient with their own personal values (Simes, 1986; Hilden, 1987; Gore, 1988). Since an individual may be averse to risks in the immediate future, the assumption of additivity over time can be dropped by 'discounting' time, making a year in the distant future less valuable. Glasziou et al. (1990) give plots that show how the choice of treatment in breast cancer may depend on the valuations given to health states and to the discount rate.

A major problem arises when the follow-ups are censored—what QOL should be assigned to the patients' unknown futures? Certain logical bounds could be assumed (Gelber et al., 1989), or it may be naively assumed that techniques of censored survival analysis could be adopted, simply by using quality-adjusted survival times as the response measure. Unfortunately, this may introduce substantial bias since the censoring mechanism is now informative. For example, in two groups with identical censoring mechanisms, the group with poorer QOL is accumulating QALYs slowly, and hence will tend to be censored earlier on the QALY scale, thus systematically underestimating the hazard in that group. This phenomenon is illustrated in detail by Glasziou et al. (1990).

If patients always progress monotonically through a limited sequence of states, then, even with censored data, estimates of the mean time spent in each state are obtainable and hence treatments may be compared. This approach has been illustrated (Goldhirsch et al., 1989; Glasziou et al., 1990) in a study of operable breast cancer, in which patients progress through stages of treatment toxicity (TOX), remission (TWIST—time without symptoms or toxicity) and relapse (REL) to death. For a particular treatment the survival curves representing the time to reach each stage, restricted to an upper time limit, can be plotted as shown in Fig. 1 for a limit of 7 years post-operative follow-up. The shaded areas then estimate the mean time in each state. Zero weight may, for example, by given to TOX and REL, and only TWIST used as a response measure (Gelber et al., 1989), or sensitivity analyses provided for a range of weights, with confidence intervals being given for the division between preferred treatments (Goldhirsch et al., 1989; Glasziou et al., 1990).

![Fig. 1. Restricted survival by stage, from Glasziou et al. (1990) (reproduced by permission of John Wiley and Sons, Ltd)](image-url)
This approach, although quite attractive, has clear limitations. More generally, it is preferable to avoid the explicit combination of QOL and length of life and instead to present them as multiple end points of a trial. These end points will have to be formally or informally combined for making treatment decisions, but responsibility for this is left to the clinician and the patient. This places a special onus on clarity of presentation of results.

In some simple situations it may be possible to provide an approximate description of the distribution of QOL versus time, conditional on survival time. Another possibility is to provide the survival curve, but juxtaposed with a description of the QOL of those patients still in the risk set. The analysis can then take the form of epoch-specific comparison of QOL. Thus treatments are compared with respect to survival and with respect to QOL of survivors: if one treatment is superior on both counts then reporting is straightforward. If, however, there is conflict then the relationship between QOL and survival should be investigated. The existence of conflict between measures of QOL and length of life exactly characterizes circumstances in which they should not be combined into a single measure.

We illustrate some problems associated with serial QOL assessment in the presence of censoring by death by reference to data on patients in the year following a heart transplant at Papworth Hospital. Here we consider data from the NHP (Hunt et al., 1986) collected at 3, 6, 12, 18 and 24 months post transplant; previous analyses have illustrated pre- to post-transplant improvements (Caine et al., 1990). Problems that arise include sporadic missing observations, censoring of follow-up at April 1990 and loss to follow-up due to either death or a persistent refusal to return questionnaires. Approximately 7% of the cohort were lost between each follow-up, apart from those censored by time. Our discussion is preliminary and informal.

Table 3 displays summary statistics for the sleep section of the NHP, details of which were given in Section 2.1. Such a table gives no idea of the progress of individual patients and does not indicate the extent of selective withdrawal of the more severe patients. However, it is apparent that a reduction in sleep scores would be attributable to an increasing proportion with zero sleep scores (i.e. no sleep problems) rather than a reduced severity among those who experience sleep problems. A fuller version of the display in Table 3 would give a histogram of the ordered scores at each follow-up. Profiles of individual patients chosen at random from the cohort could be superimposed on the population graph to illustrate the variation over time in individual patients.

The issue of selective withdrawal may be investigated by subdivision of responders at any point into those who do and those who do not withdraw before the next follow-up. Table 4 shows a breakdown of the total NHP score (obtained by adding the scores over the six main sections) according to status at next follow-up, to give a simple illustration of an approach to exploration of the data.

Without formal analysis, this suggests that those about to die or to be lost to follow-up tend to have worse current scores, whereas those who miss their next follow-up tend to have somewhat better current scores. This accords with the idea that current health status may be predictive of future clinical outcome, and that those experiencing fewer problems may not be so diligent in returning questionnaires.
TABLE 3
QOL following heart transplant: summary measures of NHP sections—proportion zero and mean and standard deviation of non-zero scores

<table>
<thead>
<tr>
<th>Sleep score measurement time</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>18 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion zero</td>
<td>0.50</td>
<td>0.52</td>
<td>0.59</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>(total)</td>
<td>(151)</td>
<td>(165)</td>
<td>(135)</td>
<td>(128)</td>
<td>(109)</td>
</tr>
<tr>
<td>Mean (standard deviation) of non-zeros</td>
<td>26.5</td>
<td>30.6</td>
<td>31.2</td>
<td>30.4</td>
<td>26.2</td>
</tr>
<tr>
<td></td>
<td>(19.8)</td>
<td>(21.0)</td>
<td>(20.2)</td>
<td>(20.0)</td>
<td>(20.8)</td>
</tr>
</tbody>
</table>

TABLE 4
Mean (and standard deviation) of total NHP score and numbers at risk in the heart transplant study, broken down by status at next follow-up

<table>
<thead>
<tr>
<th>Results after the following times:</th>
<th>3 months</th>
<th>6 months</th>
<th>12 months</th>
<th>18 months</th>
<th>24 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>52.2 (6.1)</td>
<td>52.0 (6.0)</td>
<td>38.4 (4.8)</td>
<td>38.6 (5.5)</td>
<td>36.3 (6.2)</td>
</tr>
<tr>
<td>151</td>
<td>165</td>
<td>134</td>
<td>128</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Status at next follow-up</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recorded</td>
<td>53.7 (6.5)</td>
<td>40.2 (5.1)</td>
<td>31.8 (4.6)</td>
<td>37.6 (6.6)</td>
<td>36.3 (6.2)</td>
</tr>
<tr>
<td>127</td>
<td>115</td>
<td>102</td>
<td>94</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>19.8 (10.8)</td>
<td>29.2 (12.8)</td>
<td>44.5 (18.6)</td>
<td>27.6 (12.0)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>21</td>
<td>13</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Censored</td>
<td>0</td>
<td>17</td>
<td>9</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Lost</td>
<td>73.5 (34.5)</td>
<td>95.8 (40.2)</td>
<td>53.7 (26.3)</td>
<td>65.6 (21.1)</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

7. REPORTING AND INTERPRETATION OF CONCLUSIONS

Simple, accurate and lucid reporting of conclusions is the goal, to facilitate the combination of results across similar studies and the use of results for individual decision-making by clinicians and patients.

QOL assessments are likely to be reported in medical journals in which there are strong restrictions on the number of tables and figures. A reporting strategy is set out below: it concentrates on the content of tables and figures and focuses on results pertaining to QOL dimensions. In general, detailed reporting of items within dimensions is likely to be possible only exceptionally.

Reports of QOL measures in clinical trials should address

(a) summary measures (over individuals at each assessment time or over time for each individual) against the backdrop of individual profiles (see Table 3),
(b) important differences of, say, d units between treatments or over time, by displaying typical profiles which exhibit such differences (see below),
(c) censoring of QOL assessment by death or other cause, to indicate whether censoring is informative and whether informativeness is related to treatment or time (see Section 6.4 and Table 4),
(d) conflict between dimensions, or between survival and measures of quality (see Sections 6.1 and 6.4),
(e) components of variance (assuming assessments are missing at random) (see Section 6.2) and
(f) sensitivity analysis to quantify the dependence of the results obtained on the weighting scheme employed (see Section 6.3).

Implicit in requirement (b) is an important difficulty associated with an interpretation of systematic differences between patient groups, namely the subject-matter importance of a difference of, say, $d$ units between treatments. What magnitude of difference in score can be judged substantively significant either from a clinical viewpoint or from the perspective of an individual patient? A systematic approach to elicitation of, say, clinicians' views on the scale of important changes can sometimes be adopted (Freedman and Spiegelhalter, 1983). Alternatively, gross changes can sometimes be inferred when using questionnaires with defined cut points to indicate severity; for example, a high score on the Goldberg health questionnaire (Goldberg and Williams, 1988) suggests that further psychiatric evaluation may be necessary. More often, however, we are concerned with more subtle changes, such as those of mood.

One approach is to examine typical score profiles which exhibit the difference in question. If a component of variance between times within patients is approximately known, it is plausible that a systematic difference of more than one or two corresponding standard deviations would be perceptible by individual patients. The latter (two standard deviations) is effectively the recovery of mean QOL to the patient's previous maximum: assessment times may be chosen to estimate this component of variance by incorporating a dual base-line (see Section 4.6), or replication at some time after randomization. Overall it can be argued that values of $d$ small compared with the standard deviation between patients within times are unlikely to have much importance.

The use of QOL results from trials in individual decision-making requires further comment. Methods of eliciting patients' preferences in a clinical rather than research context have begun to emerge (McNeill et al., 1978). However, it is less clear how the aggregated values and preferences incorporated into QOL instruments play a role in this context, other than in generally sensitizing clinicians to the issue. Indeed it has been suggested that, for the patient, data regarding probabilities of various outcomes are of greater importance than the values that others have attached to outcomes (Drummond, 1989).

Clinical trial results may be used to estimate prognostic probabilities, but it is important to adopt the patient's own values into any decision in which competing objectives, such as QOL and length of life, are being traded off (Simes, 1986). A general analysis allows recommendations to be expressed conditional on any individual's values (Simes, 1986), these values to be inserted later in specific cases. Since we are concerned with decisions in the face of uncertainty, values placed on states of health should be interpretable as utilities; in other words, any mean QOL values used should be considered as personal expectations.

8. CONCLUSIONS

Both their subjectivity and their multiplicity increase difficulty and complexity of QOL assessments compared with the use of a single objective measure. It is thus
essential that basic tenets and principles of design, analysis and presentation be adhered to; it cannot be overemphasized that they do apply here. In particular, we note the following.

(a) Simplicity should be the keynote wherever possible (whether with respect to design, analysis or presentation): hence our advocacy of simple weighting schemes, supplemented by sensitivity analyses.

(b) Distributions of responses to QOL assessments are at least as important as mean (etc.) values, in view of the subjectivity of the measures and the need to allow some personalization of scales, scoring systems and interpretation.

(c) Despite the desirability of allowing (some) personalization or special choices for particular applications, the use of standard (evaluated) scales is advocated as a core with more detailed customized extensions.

(d) A presentation of results of QOL assessments should, as well as being as simple as possible, facilitate their use in treatment decisions for individual patients, and interpretation of the clinical significance of effects.

(e) Sensitivity analyses should play an important role in the presentation and interpretation of the results of QOL assessments.

Despite practical and analytical difficulties, satisfactory assessment of QOL in clinical trials is worth pursuing, since the relevance of traditional approaches is limited and their objectivity to some extent illusory. To our original question 'Quality-of-life assessment: can we keep it simple?' our answer is a qualified 'yes'.


DISCUSSION ON THE PAPER BY COX ET AL.

Deborah Ashby (University of Liverpool): I am pleased to open the discussion on this paper, which represents collaboration at its best, between a group of individuals with a wealth of experience in quality-of-life assessment.

In a recent review of quality-of-life assessment in clinical trials, Schumacher et al. (1991) document the dramatic rise in the number of clinical trials publications with quality of life as a key theme. A simple extrapolation would lead us to expect nearly 1000 publications in the next year. This paper has done a great service in clarifying the main issues which run through quality-of-life assessment in clinical trials. My initial remarks are mainly relevant to clinical trials, but then I would like to widen the debate by considering the other contexts, namely individual patient decisions, resource allocation and the population perspective, which are more briefly addressed in the paper.

Because of the subjective nature of quality-of-life assessment, variables are likely to be measured with error. There is a huge literature on errors in variables; see for example Byar and Gail (1987). Phillips and Davey Smith (1991) have shown that for cohort studies with a binary end point, and multiple explanatory factors, errors in covariates can seriously bias estimation of main effects of interest. If we are looking at independent effects of correlated variables, then repeat measurements will be needed. The clinical trials described tonight reverse this set-up, with typically two groups, but multiple end points. For anybody contemplating multivariate analysis, there may be some parallels, and Section 2.2.2 on reliability is particularly relevant. It reinforces the authors' message of Section 4.1, that potential biases are better addressed at the design stage.

The authors emphasize clarity of presentation. Pocock (1991) discusses simple presentation of changes in scores between two occasions, but when the data are essentially times until the onset of some state the graphical techniques in tonight's paper are very valuable.

Much of the paper is about issues which are relevant whichever school of inference one prefers, but the paper has a predominantly classical flavour. In the clinical trials context I presume that a Bayesian approach could be taken in a way analogous to the more conventional trials (Spiegelhalter and Freedman, 1988), and using Bonferroni-type adjustments for multiple end points is essentially equivalent to taking priors for those end points with a large dollop of probability on 0. In the wider context at certain levels decisions need to be taken which depend partly on subjective judgments. I would have thought that
a Bayesian approach would come into its own. My colleague Dr Jane Hutton is involved with clinical psychologists in work taking this approach to individual patient decisions. In taking a public health perspective, I see scope for work using a Bayesian approach to decisions of resource allocation, which would benefit from collaboration with health economists already addressing these issues.

There is an explosion of interest in quality in many areas (Box, 1989). In the health context, audit is a growing area. To quote from Working Group on Clinical Audit in Primary Dental Care (1991)

‘Clinical audit in primary dental care may be defined as the systematic, critical analysis of the quality of dental care, including the procedures used for diagnosis and treatment, the use of resources and the resulting outcome and quality of life for the patient’.

A post-graduate student in Liverpool, Collette Bridgman, is currently planning to look at morbidity following out-patient and community general anaesthesia. Because the majority of the patients are children, the use of the delighted-terrible face scale (Bowling, 1991) seems promising, and may be more generally applicable in studies of children.

‘Quality’ is an explicit part of the government’s discussion document, ‘The health of the nation’ (Department of Health, 1991). How does quality-of-life assessment relate to this? In clarifying my own thoughts, I found the division into structure, process and outcome (McLachlan, 1991) useful. Structure and process are in many ways easier to study, and the National Health Service collects a large amount of information on these. However, as McPherson (1992) points out, it is only where the relationships between process and outcome are well understood that indicators of process have any direct relevance to health. Hence, in this new climate where ‘health gain’ is a key phrase, outcome needs to be measured directly, and this is why I expect quality-of-life measurement to increase in importance.

Quality-of-life assessment draws on concepts from psychology, economics, clinical medicine, public health and statistics. The authors’ contribution in this paper is to tease out the specifically statistical issues, and to address them systematically. The key message to take away from the paper is to improve quality of life by remembering to KISS, which stands for ‘keep it simple, stupid’. I shall therefore simply say that it gives me great pleasure to propose the vote of thanks.

Anthony Hopkins (Royal College of Physicians, London): Dr Ashby started by saying that all the authors have wide experience of quality of life. But of course all of us do on the basis of our own life experience. However, there are not only methodological problems in measurement but also difficulties in deciding exactly what it is we want to measure. To take Dr Ashby’s example of young children going in hospital for dental extraction, I would say that we have to consider not only the quality of the child’s experience, but also of the effect of the arrangements on the quality of life of the mother. For example, although short stay day case surgery might give the child a better experience, a working mother might have to make special arrangements to pick a child up an hour after the procedure, causing many more difficulties than if she were able to leave the child there for the whole day.

All of us are aware that we are attempting to measure health-related quality of life. Even this limited concept has several components—fixed morbidity, which cannot be changed by medical care, movable health status which, by definition, can be shifted by care, and health-related quality associated with non-medical aspects of socioeconomic life. It is well recognized that housing and income both have major impacts on health. Medical research workers are now concentrating on health-related outcomes, with special reference to functional status. An alternative is to identify outcome measures which are of importance to patients. Danny Ruta, a Health Service researcher in Aberdeen, is looking at patient-derived outcome measures. For example, one patient with varicose veins may be concerned about aching in the legs and another by the cosmetic appearance. If the person who is bothered about aching has a wonderful cosmetic result, but the aching continues, then the outcome has been ineffective with regard to the symptoms of particular importance to him. Dr Ruta finds that people are often not very articulate in formulating their ideas in concrete terms, and if they cannot do so there is a checklist of common symptoms to remind them of those which they would particularly like to see addressed. Such individualized patient outcome measures are of great importance for the individual patient, but are less easy to aggregate in terms of the current fashionable concept of health gain. This adds another dimension to the difficulties raised by David Spiegelhalter who drew attention to the dangers of comparisons between different outcomes, all of which gained the same added number of quality-adjusted life years.

As an academic clinician, I am alarmed about some issues of measurement which do not seem to be adequately considered. For example, when brought into routine use, I am not sure that adequate attention will be paid to the timing of administration of various questionnaires designed to measure health...
status. To take a simple example, whether a questionnaire is given immediately before or after the consultation will have an important bearing on the patient's perceived health depending on the content of the consultation, and yet, if the nurse giving out the questionnaire just regards it her duty to administer it on a particular day, major technical issues of medicine arise.

Most of us would accept that we cannot add up scores on various dimensions of a complex health profile, but if we try to look at each of the six or seven domains in, for example, the Nottingham health profile separately for large numbers of patients in a trial we shall become totally confused. Some patients will achieve improvements in one domain, but losses in another: hence the temptation to seek simple aggregate unidimensional scores of health status.

When we come to valuations of different health states, then important ethical issues arise, particularly here with regard to allocation of resources. If resources are to be allocated according to some measure of quality of life, we have to ask whose valuation it should be. Should we use the patient-specific valuation, as for varicose veins, or valuations by the community at large, as is being tried in Oregon, or the valuations of a particular sector of society such as doctors or nurses? It is noteworthy that the commissioners in Oregon had to alter the valuations derived from their community surveys to make the rankings look more sensible, and ethical. For example, it may turn out in some similar survey in this country that the valuations of the health status of mentally handicapped children might not be high, and yet they are expensive to care for. In these circumstances, the 'gain' to the community by medical interventions has to be accepted as an ethical gain, rather than an improvement in measured health status.

Whatever these difficulties, I am sure that we should use a wider range of outcome measures in clinical trials, alongside traditional biological measures such as blood pressure. Although biologically important, no-one is aware of their blood pressure (until it reaches inordinately high levels). In fact the quality of life of someone with hypertension seems to depend more on the way that the individual is told about his high blood pressure, and the approaches adopted, than the height of the blood pressure itself. We need to add to trials additional measures which reflect personal reality.

It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

J. L. Hutton (University of Liverpool): I visited an exhibition, 'The art of death', on the way to this meeting. Good death and bad death were defined by moral criteria in earlier centuries. Quality of life was considered by the ancients to be a question of virtue. A spiritual dimension is rarely addressed in quality-of-life assessment. Movements to assess and improve quality of dying (QOD) and death in modern society have considered spirituality. As yet there are no recognized scales for measuring QOD. With the move to community care and increasing numbers of the hospices, simple, valid and practical measures of QOD need to be developed.

In response to the recent initiative by psychiatrists and general practitioners to increase the proportion of people whose depression is treated, a Times editorial pointed out that depression could be a spur to achievement. What social effect might the treatment of Winston Churchill's or Schubert's depression have had?

The value of patient's choice of treatments has been studied with breast cancer patients. With guidance from a clinical psychologist, patients were able to assess possible relevant outcomes of lumpectomy and radical mastectomy, and to assign their own utilities. Formal calculation of subject-expected utilities was not often necessary (Owens et al., 1987). The opportunity to make their own treatment decision also enhances the patient's quality of life (Leinster et al., 1990).

A veterinarian colleague has expressed interest in this paper. Little assessment has been made of the quality of life of animals after treatment. Although a veterinarian in general practice might be able to follow up his patients informally, my colleagues at the university equine centre generally have no further contact once a horse is returned to its owner. A common reason for referral is colic. Anecdotal evidence suggests that some horses suffer repeated bouts of colic after being returned home and the owners have them put down. Thus there are two main dimensions on which the quality of life of an animal can be measured: the usefulness of the animal after treatment and the perceived happiness and freedom from pain of a pet. In addition, owners could be asked to fill in an Eysink personality inventory to enable the veterinarian to assess whether a pet needs treatment, and perhaps also as an assessment of the treatment.

More seriously, there are veterinarians concerned to ensure that their treatments enhance animals' quality of life. Simplicity will be encouraged by the patient's limited ability to communicate. Deming's
message and the message of this paper, to keep evaluation simple so that all can participate in improving quality, is well worth emphasizing.

Christopher J. Bulpitt (Royal Postgraduate Medical School, London): The authors make several important points concerning the assessment of quality of life (QOL). They have stressed the necessity for standard, valid, repeatable and sensitive measures. Certain instruments may be suitable for inclusion as core data in several studies. They conclude that there is no gain beyond 5-7 categories of response to individual items and suggest that more than one series of questions could be completed by different patients. Patients could therefore be randomized to different questionnaires with certain common material. This would appear very promising when large numbers of subjects are available. The authors consider that sophisticated weighting systems are a waste of time and the unweighted average score should be calculated and standardized, for example as a percentage of the maximum. The authors consider that the data should be presented according to the different responses (negative, positive, etc.) and, moreover, presented through time.

In addition, they suggest that the data could be broken down according to whether the subject attends for a further evaluation, misses the next visit but not later visit(s), is lost to follow-up or later dies.

My criticisms of the paper include the fact that their example in Section 3.2 ‘trials of anti-hypertensive therapy’, was never subsequently referred to! The authors also believe that ‘clinical trial results . . ., are best left multidimensional’ but not for resource allocation where they originally said that any QOL assessment needs to be collapsed to a single scale to allow direct comparisons between options. I believe that those responsible for resource allocation can understand a simple table giving details of the different options. Such a table may include information on the physical, psychological and social scores for QOL, a health status index scored between 0 and 1, years of expected survival, the age of the subjects involved, and the costs of the various options. The authors also consider that secondary outcome variables are primarily ‘hypothesis generating’. Results for all outcome variables may need to be confirmed and I would welcome an adjustment of p-values when several outcome variables are examined. However, the authors rely on the severe Bonferroni adjustment for the few (how many?) key issues that they choose.

I prefer the term ‘repeatability’ to ‘reliability’. Under the heading of reliability, the authors mention a split-test reliability which has more to do with validity than with repeatability.

Lastly, have the authors really kept it ‘simple’? They advocate simple weighting together with sensitivity analyses and ‘some personalization of scales, scoring systems and interpretation’. The answer to the question could be a qualified ‘no’, but the paper has clarified many important issues.

H. M. Goodare (Horsham): As a breast cancer survivor I find it most encouraging that the authors take such a refreshingly client-centred approach to this difficult subject. I agree that a pragmatic approach is preferable to aggregating quality-adjusted life years, and I would give considerable weight to the avoidance of distress and suffering: the healer primarily should do no harm (Table 2, treatment F).

I would also endorse whole-heartedly their view that ‘it is important to adopt the patient’s own values into any decision in which competing objectives, such as QOL and length of life, are being traded off’ (Section 7, p. 371). Clinical judgment in such matters can only be made on the basis of patient choice, and this implies a much greater willingness on the part of clinicians to involve the patient in the decision-making process than is common, in my experience.

The great singer Kathleen Ferrier was given a choice. When she was offered hormone treatment for her advanced breast cancer, a laryngologist told her that ‘the long-term effect on the cancer would be negligible but that the effect on her voice might be disastrous’. In reply, Ferrier ‘said she regarded her voice as a divine gift and would go to the grave with the voice as it had been given her. She had this overpowering desire to sing in Orpheus and said that she would go ahead’ (Leonard, 1988). The choice was her privilege, and our immeasurable gain.

I was not so fortunate with my own hormone treatment. My oncologist assured me that Tamoxifen had no side-effects, but after two years on the drug my singing voice has lost its top octave. My QOL now lacks a dimension of great importance to me, and I was not given the opportunity to make my own cost–benefit analysis of the treatment on offer.

There is a spiritual element in QOL that will always be nearly impossible to assess. Some people may value hearing more than sight, the ability to read books more than physical mobility, and so on. But if the patient is asked to select his own QOL objectives, researchers are unlikely to err.

I am very hopeful that the authors’ sensitive, careful and commonsense approach to QOL assessment will be widely adopted. No other will do, in my view.
Roy A. Carr-Hill (University of Hull): There is an additional point which arises from Dr Ashby’s presentation. The reference to ‘The health of the nation’ also provides the correct back-cloth to debates about improving health-related quality of life: the movement towards ‘health for all’ laid out at Alma Ata and the growing realization that these targets are only attainable with through-going community involvement in health.

I would like to extend the authors’ arguments for simplicity. They focus mainly on the problems of comprehensibility to clinicians or to individual patients. Although this is the correct focus for evaluating quality-of-life assessments in specific clinical contexts, the major focus of debate in policy terms is the use of quality-of-life assessments to derive quality-adjusted life years for comparison across conditions to inform resource allocation decisions.

In that context, the issue of simplicity assumes other dimensions. For the use of quality-of-life assessments derived via complex statistical methods to inform decision-making about resource allocation means that both the ‘manager’ and the ‘public’ are dependent on apparently ‘scientific’ evidence. Democracy suffers not only because few can understand the manipulations involved but also because of the pseudoscientificity itself. The claim that the value judgments involved in deciding priorities about health care can and should be organized into a technical framework is very contentious (see Carr-Hill (1991)); it also runs counter to the growing realization that, only by involving the community in health-making decisions about health and health care can the ‘health-for-all’ targets be reached (World Health Organization, 1989).

Involving the community in this way imposes a much stricter requirement of simplicity: for the reporting of quality of life, assessments have to be made not only directly relevant to the patient in a clinical context but also to (representatives of) the community who should be involved in decisions about the allocation of resources. Instead of focusing on efficiency via the cost accountant, any reform of the National Health Service should focus on equity through the extension of democracy.

It is important to be clear about the role of professional statisticians in public policy debates. Whereas it used to be thought appropriate for professionals (including statisticians) to make ex cathedra pronouncements about the state of the world and how their findings should be taken into account in making policy, in a world where the practice of health development is increasingly influenced by the concept of participation, it is antediluvian. The corollary of the authors’ injunction to ‘make it simple’ is ‘power to the people’. The sooner that is taken seriously, especially in decadent capitalist economies such as the UK’s, the better.

Trevor Sheldon (University of Leeds): I am concerned that the rigour which the authors impose on the use of health-related quality-of-life (QOL) measures in clinical trials is not extended to their use for resource allocation. In Section I they appear to accept that the dimensions of QOL need to be collapsed to a single scale to allow direct comparison in individuals and for resource allocation. This will be seen by many economists and others keen on league tables comparing cost and quality-adjusted life years (QALYs) as a welcome escape clause.

For individuals, collapsing QOL dimensions into a single score is reasonable in that they can be involved in a practical decision analytic exercise. The weighting will reflect individual preference sets in the context of real treatment options. The results are not generalized to others. However, economists, via the QALY, are developing a standardized, non-disease-specific instrument—a generic single-score outcome measure which, when combined with cost data, can be used to rank treatments for rationing.

In clinical trials subjects are relatively homogeneous with respect to disease state and the dimensions of health in which change is expected. In the derivation of QALYs, however, estimated valuations of health states are for application across disease states and a more heterogeneous population with respect to, among others, experience of illness and degree of risk averseness.

Valuations are often elicited by presenting hypothetical questions about abstract health states. The assumptions of underlying consistency and rationality of personal evaluations on which the welfare economic framework rests are also unrealistic in this context. The high interpersonal variability of valuations limits the usefulness of averages for large-scale decision-making. QALYs are used for ranking treatments which affect very different combinations of dimensions of QOL subsumed under a single scale.

Though it would make choices between health service interventions easier, a sensible combination of dimensions of QOL is both conceptually and technically more problematic in resource allocation than in clinical trials and should not be done other than on a rigorously scrutinized experimental basis.

Ultimately perhaps there is no scientific basis on which final choices about resource allocation can be made. It is a fundamentally political process which must take into account a range of factors...
such as equity. Statisticians must be vigilant and build on this important paper in applying strict criteria for the use of health-related QOL measures, not only in clinical trials but also in resource allocation.

Jack Hibbert (formerly Central Statistical Office, London): My interest in this subject stems, not from any direct knowledge or experience of it, but from the intriguing nature of the questions that arise when one attempts to apply statistical method in a field of this kind. My lack of knowledge may be illustrated by my initial reaction to the hypothetical scale for angina referred to in the verbal presentation by David Spiegelhalter. This was that there should perhaps be a further value of –1 for those who were alive, but wished that they were dead. There may be another reason which is that, having recently reached the age of 60, I am starting to take a greater interest in the potential effects of medical treatment on my future quality of life than I have hitherto.

As I read the opening sections of the paper, I became increasingly troubled because the methods being discussed seemed to imply an approach which failed to provide for decision-taking based on personal choice. By the time that I had finished the paper, however, I was considerably reassured, even though it remains unclear to me how precisely health-related quality-of-life (HRQOL) assessments—as distinct from estimates of the probabilities of different outcomes—might play a useful part in decision-taking by the individual patient.

This, however, is only one of four contexts identified by the authors for the potential use of HRQOL assessments. The paper deals very comprehensively with their uses in comparisons of interventions in a clinical trial, but much less fully with the two remaining contexts: measuring the health of populations and assessing the benefit of alternative uses of resources. There is nothing wrong with this, of course, but I suspect that it reflects the fact that the use of HRQOL assessments in these contexts may be of more doubtful value.

But the question of whether HRQOL assessments are of value in these two contexts is very important, because the pressures to use them are considerable and are likely to become even stronger. Those seeking to describe complex phenomena, or to take decisions based on them, will inevitably be drawn towards summary measures with an apparently scientific basis, as their simplicity is attractive and their pedigree provides support for their use. So, whether we like it or not, HRQOL assessments are, and will be, used in political contexts. Work which can throw more light on the relevance of HRQOL assessments in these contexts is thus needed to put such uses in better perspective. I see this paper as a milestone along that road, as well as being an important addition to the literature on the uses of clinical trials.

Keith Abrams (University of Liverpool): I would like to make three brief comments regarding the analysis of quality-of-life data.

For many chronic diseases length of life and quality may be kept separate as the authors have stated. However, in cancer, and especially for certain forms of cancer, where length of life can be limited, we may be interested in the combination of both quality of life and length of life. In the cases where a cure is attempted there is no question that length of survival is the primary outcome measure, and equally, in patients diagnosed with advanced disease, straightforward palliation is required. This leaves a substantial proportion of patients for whom both quality and length of life are inextricably linked. For example, Fayers et al. (1991) reported the results of a quality-of-life study for small cell carcinoma of the lung when changes in quality occur over short periods of time, and where the main aim was a reduction in treatment toxicity at no expense of survival. In such cases it may be desirable to combine quality and length of life in some way.

The authors have described various techniques for combining quality-of-life data with survival data. I would like to expand briefly on their description of multistate models that have been used and advocated by Andersen (1988), Kay (1982) and Olschewski and Schumacher (1990). Fig. 2 shows the type of data that have been obtained for a group of lung cancer patients in a pilot quality-of-life study at Clatterbridge Hospital, Wirral. Leaving aside the issues surrounding how the quality-of-life states were formed, the key question is—how best can quantitative statements be made about these patients that would be beneficial in future patient–treatment decisions? One way to describe such data would be to use a multistate model, such as that shown in Fig. 3. The risk of the various transitions could then be estimated yielding clearly interpretable quantities, rather than the more vague statements obtained by using quality-adjusted survival techniques.

The authors have rightly stated that multistate models can require extensive data, and that various assumptions have to be made. However, as quality of life becomes increasingly more important, and is more routinely collected, the issue will become not the lack of data but rather how to handle the
Fig. 2. Lung cancer patients in a pilot quality-of-life study at Clatterbridge Hospital, Wirral: ———, 'good'; ······, 'medium'; ————, 'poor';  , death

Fig. 3. Simple multistate model for a quality-of-life study

data that are being collected. This situation is not unique to quality-of-life studies, but merely highlights the need for suitable data handling software for many different types of event history data, as previously recognized by Clayton (1988).

Peter Fayers (Medical Research Council Cancer Trials Office, Cambridge): Firstly, another way to think about the nature of the trade-off is to consider a two-dimensional plot with quality-of-life score on one axis and survival benefit on the other (assuming that there is no censoring). Thus patients are represented by points in two-dimensional space. If we ignore the separate issues of palliative treatments, in most cases we are interested in improved survival which may be to be at the expense of poorer quality of life, i.e. we are concerned with a trade-off in which the points show a negative correlation. Thus, to answer whether patient A is 'better' than patient B, we are imposing an unnatural ordering on points scattered in two dimensions. If we then add cost, we obtain a three-dimensional space in which we would have to seek a ranking of points. This seems totally unreasonable, and it is surely preferable to avoid combining such different dimensions into a single score.

Secondly, an implicit assumption that underlies most clinical trials is that the results can be generalized to future patients, in particular that, if a group difference favouring treatment A is detected, then most
future patients will either benefit from A or, at worst, experience no overall difference between A and B. When the reverse effect applies, where patients who receive A have a disadvantage, there is usually a recognized biological explanation. This principle of transferability of results does not in general apply to quality-of-life assessments; it is common for some patients to prefer one treatment, whereas others feel that the toxicity outweighs its benefits in treating the disease.

Thirdly, it is curious that there is a long tradition of clinical trials reporting treatment toxicity alongside cure rates or survival rates, with hardly even an attempt to combine them into a global score. Why does merely changing the word ‘toxicity’ into ‘quality of life’ make so many people want to use a different approach to analysis? I can only offer the reason that a clinician would hope to monitor and even to control specific toxic reactions, and therefore regards toxicity as a completely separate concept—or dimension—from cure. The approach to quality-of-life assessments should be similar.

Finally, like the authors of this paper, I regard sensitivity as being the most important property of any quality-of-life index. If an index can detect differences between treatments, it must be measuring something; if it also has face validity, and therefore seems sensible, it becomes potentially useful. And, if it is based on a simple general question and yields a score that both clinicians and patients can understand and regard as meaningful, there are clear advantages.

R. G. Newcombe (University of Wales College of Medicine, Cardiff): It is possible, and arguably appropriate, to examine duration of survival, and quality-of-life (QOL) conditional on survival, both separately and in combination. I propose an alternative to the quality-adjusted life years approach which may be used for each of a number of scales and time points separately; it is appropriate for relatively short follow-up periods. The method may be termed ‘pessimal substitution’.

To illustrate the method, first consider the data in Table 3. Here, proportions scoring 0, and distributions of non-0 scores, are presented separately, but this is not the only possible representation. Suppose that such results, relating to a single time point, require to be compared between two treatment groups. As well as comparing proportions scoring 0, and distributions of non-0 scores, it would be natural to compare the full distributions in the two samples, typically by the Mann–Whitney test, and corresponding confidence intervals.

An analytically similar approach can take into account deaths before the chosen time point. Bayer et al. (1987) performed a randomized trial comparing intravenous glycerol against a placebo in elderly patients who had suffered an acute stroke within 48 h of entry. Several neurological and functional assessment scales were evaluated at 6 months and 12 months. Three strategies were used in the statistical analysis.

(a) Survival data were compared between treatment groups.
(b) At each follow-up time $t$, each assessment measure $X$ was compared between the two groups for survivors only.
(c) At each follow-up time, each assessment measure was compared between the two groups by using results for all patients.

To do this, initially each subject who had died was assigned the lowest score for the survivors at that assessment. Other missing observations due to (credibly) uninformative censoring are disregarded. The effect of altering the score assigned to intervening deaths, to different values at the unfavourable end of the QOL scale, may then be examined; if the conclusions are robust under such perturbations, the analysis provides a sensitive, holistic, assessment of QOL when not all the cohort has survived. A Mann–Whitney-based approach may then lead to a summary statistic of estimative form, estimating

$$\Pr[X|RX_1 > X|RX_2] + \frac{1}{2}\Pr[X|RX_1 = X|RX_2],$$

$RX_1$ and $RX_2$ denoting the two treatments—a rational basis for future clinical decision-making.

The method is most suitable for short follow-up intervals, when relative duration of survival in those dying by time $t$ is comparatively unimportant—these may have had very poor QOL between entry at death, anyway. For longer follow-up, the method is less appealing: there is a strong reason to award different scores for patients dying at different intervening times such as 0.9 and 0.1, and robustness with respect to moving such a variety of scores for deaths past scores for poor quality survival becomes less plausible and more difficult to assess.

The following contributions were received in writing after the meeting.

Morag Farquhar (St Bartholomews Hospital Medical College, London): I congratulate Cox et al.
on their paper calling for simplicity in the measurement of quality of life in a field where there is currently no standard measure of the concept. However, I am concerned that they provide no explicit definition of what they mean by 'quality of life'.

It has been noted elsewhere that a profusion of papers deal with quality of life, but in a heterogeneous manner, making comparison difficult (Ebbs et al., 1989). A closer inspection of papers that mention quality of life in their titles often reveals a rather limited perception of this concept; many have either avoided defining what they purport to measure or have limited their definitions to what the investigators have seen as large components of the whole concept (Ebbs et al., 1989).

There is an implicit assumption among some researchers that we all know what we mean by quality of life. For the individual researcher, or team of researchers, this is probably true; however, within and across disciplines there is no firm consensus about the meaning of the term or its theoretical construct (Schipper and Clinch, 1988).

The term quality of life has been described as ambiguous and aggravating (Edwards, 1985) or, perhaps more constructively, as a rubric (Schipper and Clinch, 1988). Where researchers fail to define the term we can only gain some insight into what they mean by how they put it into operation.

Cox et al. list a series of, what are essentially, health and functional status scales as their measures of quality of life. However, health and functional status are just two of a (potentially infinite) number of dimensions of quality of life. For example, it has been said that other dimensions include life satisfaction, self-esteem and socioeconomic status (George and Bearon, 1980). If we must use the term quality of life in such contexts, then what Cox et al. are referring to is 'health-related quality of life', not quality of life itself.

One of the reasons that there is no standard quality-of-life measure is that there is no consensus definition of quality of life. To reach such a consensus we must open the debate on the definition of the term; the only way to achieve this is for researchers to say what they mean when they use the term. Equally, where we are referring only to the dimensions health and functional status we must either make this explicit or use the term health-related quality of life. To use the term quality of life is misleading.

R. D. Gelber (Harvard Medical School, Harvard School of Public Health and Dana–Farber Cancer Institute, Boston) and A. Goldhirsch (International Breast Cancer Study Group, Lugano): Quality-of-life (QOL) assessment is indeed gaining in importance. This paper is another testimony to this fact and discusses many of the same issues raised by Schumacher et al. (1991).

Throughout their paper, Cox et al. caution against combining QOL and length of life, and recommend that they be presented as separate multiple end points of a trial. An extension of this warning to the economic sphere would argue that global indices concerning the state of an economy should never be calculated because they fail to highlight the status for an individual industry. By rejecting all attempts to factor aspects of QOL into the primary clinical trial results we abdicate our responsibility to assist our clinical colleagues with treatment decision-making. The quality-adjusted time without symptoms and toxicity (Q-TWIST) method was developed to emphasize rather than to obscure the spectrum of interpretations that might affect individual members of the population.

One aspect of particular research interest to us is the changing face of QOL adaptation over time. For treatment comparison, as patients experience events which significantly reduce their QOL, adjustment and coping begin to modify the future impact of the incident event. Focusing on the duration of time and its quality brings the analysis of clinical trials data closer to the main interests of the patient (Gelber et al., 1989). We have made assumptions with Q-TWIST to keep it simple and to provide a first step towards incorporating QOL into the usual assessment of clinical trials results. The approach begins by partitioning the overall survival experience into average time periods coloured with a QOL type of assessment. By showing this partitioning for only one treatment in Fig. 1, Cox et al. fail to recognize that the main contribution of Q-TWIST is to provide a treatment comparison. The method also provides treatment comparisons which illustrate the evolution of QOL gains over time, and threshold utility analyses which explicitly summarize how value judgments influence conclusions about therapeutic benefits. Recent applications of Q-TWIST have illustrated the trade-offs between toxic effects and disease control for longer versus short duration chemotherapy in breast cancer (Gelber et al., 1991), chemotherapy versus no adjuvant treatment for node negative (relatively low risk) breast cancer (Gelber et al., 1992a) and zidovudine versus placebo for mildly symptomatic individuals infected by the human immunodeficiency virus (Gelber et al., 1992b).
The next step in the development of Q-TWIST is to incorporate individual patient-derived assessments. We look forward to the increased use of the other approaches suggested by Cox et al. to highlight QOL as a primary consideration for treatment comparisons.

David J. Girling (Medical Research Council Cancer Trials Office, Cambridge): I should like to draw attention to the value, in assessing the quality of life of patients in randomized cancer trials, of a simple diary card completed by the patients. In many such trials we have asked patients to complete a card every evening after their last meal, recording how they have been feeling during the previous 24 hours (Fayers et al., 1991). The card that we use typically contains 6–8 questions on physical symptoms, and the card that we are using in current trials on the treatment of lung cancer contains questions on loss of appetite, nausea, vomiting, cough, coughing up blood, chest pain, shortness of breath and difficulty in swallowing (dysphagia). Against each symptom, the patient records a score according to the following instruction. ‘Please complete every evening after your last meal, even when you are in hospital, by writing the number of your answer in the appropriate box as follows: 0—not at all; 1—a little; 2—moderately; and 3—very much.’

In comparisons of treatment policies, this instrument has been both sensitive and reliable when its use is confined to periods when day-to-day changes are likely, e.g. during the administration of chemotherapy or radiotherapy. As an example, in a randomized trial of palliative thoracic radiotherapy with two fractions given one week apart (F2-regimen) versus a single fraction (F1-regimen) in patients with inoperable non-small-cell lung cancer (Medical Research Council Lung Cancer Working Party, 1992), we were expecting that the two regimens would be equivalent in their palliative action and adverse effects. In the event, as assessed by the clinicians, the two regimens were essentially identical. In marked contrast, on daily assessment by the patients using the diary card, those treated with the F2-regimen experienced substantially more dysphagia (Fig. 4), which was recorded as a transient adverse effect of radiotherapy in 56% of 117 patients compared with 23% of 118 in the F1-group (difference, 33%; 95% confidence interval, 17–48%). The clinicians, who assessed the patients at the start of treatment, making their next assessment one month later, reported that dysphagia had occurred during the first month in 25% of the F2-patients and 17% of the F1-patients. These findings led to our recommending the F1-regimen.

We have shown that this simple diary card is reliable; for example, the patterns of dysphagia with the F2-regimen in this trial and in a previous trial in which the same regimen was included (Medical Research Council Lung Cancer Working Party, 1991) were virtually identical. We are one of the few groups comparing quality of life in different treatment groups in randomized cancer trials.

Fig. 4. Percentage of patients reporting dysphagia on their diary cards: the arrows show assessments by clinicians and bars the doses of radiotherapy
Swend Kreiner (Danish Institute for Educational Research, Copenhagen) and Niels Keiding (University of Copenhagen): We whole-heartedly support several of the general points put forward, especially the point that properties of a measurement instrument should always be tested before it is used in a new context.

Although simplicity is important, it should always be subordinate to the requirement that measurements are valid, generalizable and objective. Construct validity should be regarded as the overriding concern in the initial development of measures. Construct validity is much more than just a question of the overall pattern of relationships between the instruments and other measures. It is also a question of whether or not items included in one measure are correlated because they all reflect the same underlying construct, and a question of whether they correlate with items from other measures because the underlying constructs correlate (Bohrnstedt, 1983). It should be an integral part of the statistical analysis to assess empirically these further aspects of construct validity also.

Several frames of inference have been developed for analysis of construct validity. In addition to factor analysis the family of models belonging to the so-called item response theories (IRTs) should be mentioned (Hambleton, 1989). A recent discussion by Rosenbaum (1989) gives an excellent introduction to the analysis of construct validity within an IRT frame of reference.

If one is concerned about simplicity one should consider the special IRT model referred to as the Rasch (1960) linear logistic model for binary items and its generalizations to more multifarious responses (Andersen, 1990). These models meet some of the requirements of construct validity and imply both objectivity and simplicity. They yield statistically sufficient unweighted raw scores that are objective measures of the latent construct. The Rasch models are attractive from an applied point of view as well. There is a rich literature on the technical aspects of analysis by Rasch models, especially on the many different approaches to examination of the adequacy of the models. The reader is referred to Andersen (1980, 1990), Gustaffson (1980) and Andrich (1988) for additional references on this subject.

Rasch (1977) coined the term ‘specific’ objectivity for the type of objectivity associated with the Rasch model, stressing that measures may be objective within some specific frames of reference, but not within others. We believe that the notion of specific objectivity could be particularly helpful in quality-of-life contexts, because of the ambiguous nature of quality of life.

Manfred Olschewski, Gabi Schulgen and Martin Schumacher: (Universität Freiburg): We congratulate the authors on their comprehensive and sophisticated discussion of methodological principles in quality-of-life (QOL) assessment. The paper is particularly important for the development of an adequate statistical methodology for the analysis of QOL data. A recent survey (Schumacher et al., 1991) on analyses of QOL trials published in leading journals revealed that only few used adequate statistical procedures.

The authors propose the use of simple averages over items for a specific dimension of QOL to yield comparability over different trials. This implies an equal weighting of all items belonging to a particular QOL dimension. Where patients’ weights for particular items differ between trials, comparability of QOL results is not assured by the formal restriction to unweighted summary scores. Factor analytic techniques based on the empirical correlation structure among the items can be applied to check the consistency of the weighting schemes used in different trials and at different points in time within the same trial.

The statistical analysis procedures proposed by the authors assume normally distributed QOL data. Extensions of these methods to longitudinal data (Zeger et al., 1988) based on generalized linear model terminology might be appropriate for dealing with the non-normality which often arises in data from QOL questionnaires.

The authors pointed out that the use of multistate models in the analysis of QOL data is hindered by the requirement of extensive data. Usually the exact transition times are not observable; only the state at a preplanned follow-up visit is known. Kay (1986) proposed a method for analysing such incomplete data, accounting for unobserved transitions between two follow-up visits. Grueger et al. (1991) investigated the validity of statistical inferences based on such incomplete information.

We agree with the authors’ arguments advocating the use of validated standardized instruments for QOL assessment to facilitate meta-analysis of several trials. One way of assuring practicability of QOL assessment is by use of short and simple questionnaires. At the expense of higher personal and financial resources also non-simple QOL assessment using structured interviews can result in complete, high quality data, as our experience in a multicentre clinical trial on over 200 patients suffering from heart failure.
DISCUSSION OF THE PAPER BY COX ET AL.

has proven. The response rate on more than 100 items over the whole follow-up period of 2 years was almost 100\%, the proportion of missing data being nearly 0.

Although not arguing against the attempt to keep QOL assessment as simple as possible, one should be careful not to miss important features of QOL by restriction to too simple methods as a sacrifice to medical interpretability.

**Gwen Parr** (Ciba-Geigy, Horsham): The efforts of the Oregon Health Services Commission in its attempts to prioritize 800 basic health services illustrate the practical problems encountered in the use of a formula based on quality-adjusted life years and its combination with weightings derived from various sources, including members of the public (Klein, 1991). Initially cosmetic breast surgery was ranked higher than treatment for an open fracture of the thigh; a real example of one of the extraordinary conclusions that you suggest in Section 5.2 and ample proof of the need for simple scales that appear reasonable.

Your heading to Section 4.3.1 refers to rheumatoid arthritis studies but at least one trial is cited which was confined to osteoarthritis, a very different condition. I think that Section 4.3.1 would be better headed ‘Arthritis studies’, to encompass both diseases. (See Section 3.3.)

**Stephen Senn** (Ciba-Geigy, Basle): I particularly welcome the perspicacious and useful discussion of treatment-by-patient interaction in Section 6.2. This is a topic which is very poorly dealt with in general in publications on clinical trials, even in papers written by statisticians. The authors’ analysis is relevant to all outcomes, not just those measuring quality of life, and deserves to be widely read.

It is ironic (although not inappropriate) that Cox et al. should begin their discussion of this problem with crossover trials, since any perusal of the statistical literature will show that the models commonly employed for multiperiod designs assume that once patient, treatment, period and (inevitably) carry-over have been ‘eliminated’, everything is pure ‘noise’ (Senn and Hildebrand, 1991). In terms of the model of Section 6.2 this implies that $\sigma^2 \leq \sigma^2$, that all patients respond identically and that, however tempting it may be to attempt to do so, the individual within-patient comparisons from an AB-BA design, period corrected or not, cannot be used to identify responders and non-responders. Similar attempts to define response in parallel group trials by comparing values under treatment with the baseline are even more problematic and indeed, in my view, the authors’ model also implies that many common ‘response’ measures used in clinical trials, e.g. ‘time to return to base-line’, are conceptually false. It is interesting to note, in passing, that the ratio $\tau/\sigma_\tau$ is the effect size, much beloved of meta-analysts (Hedges and Olkin, 1985), and that the exposition here has implications for analysis and interpretation in that field also.

The burgeoning literature on ‘N-of-1’ trials, much of it in the medical press, shows, however, that physicians have other obsessions. For them nearly all error is treatment-by-patient interaction and hardly any is noise. Clearly the truth lies somewhere between the statisticians’ and the doctors’ interpretations. I hope that the authors’ discussion of components of variance, which shows quite clearly how difficult it is to extract the treatment-by-patient interaction but that it may nevertheless be important, will provide the inspiration for some more reasonable work in N-of-1 trials, where the current level of debate is abysmal. Studying series of N-of-1 trials, or equivalently multiperiod crossovers using the components-of-variance approach as outlined here, would be a considerable improvement on the current habit of peering at $P$-values for individual patients. This is one area where, the authors’ title notwithstanding, things have been kept far too simple for too long.

**P. J. Solomon** (University of Adelaide): An assessment of the overall ‘well-being’ of a patient often includes investigating the occurrence of a particular complaint. Some suggestions for the analysis of complex data on many complaints arising from a clinical trial are the following.

A simple way of identifying subgroups of patients with many complaints would be to let $k$ be the number of complaints that the patient reports present, so that $k = 0, 1, 2, \ldots \ldots$. The distribution of $k$ could then be investigated for homogeneous groups of patients, e.g. by treatment group.

When repeated responses are given over time on severity as well as presence or absence of a complaint, individual patient trajectories can detect patterns of variation and give signals of sudden change. Simple measures of linear trend, serial correlation or seasonal variation, and associated components of variance, may be useful. For example, if $y_{imr}$ is the $r$th score for the $i$th individual in calendar month $m = 1, \ldots, 12$, then we can take

$$E(Y_{imr}) = \mu_i + \alpha_i \cos \left( \frac{2\pi m}{12} \right) + \beta_i \sin \left( \frac{2\pi m}{12} \right).$$
Estimates of $\mu_i$, $\alpha_i$ and $\beta_i$ can be obtained by least squares and examined for any relationships or trends. $
abla$ and $\beta$ and the components of variance $\sigma^2$ and $\sigma^2$ could also be estimated for each treatment group.

A simpler approach to summarizing individual patient trajectories is to find the proportion of time that the complaint is present and the corresponding average severity. The distribution of both statistics within treatment groups could then be examined.

If interest lies in investigating the relationship between the patterns of response over time for two or more complaints or items, perhaps within a given dimension, a starting point would be to plot the severity of one complaint against another at each time point with the aim of describing the plots by one or more simple summary statistics.

A major issue in the analysis of large-scale clinical trials is investigating relationships among important features of the data usually to compare treatment groups. An example would be to calculate within treatment groups the proportion of patients for whom both the patient and the doctor report the complaint absent, the proportion for whom the patient reports absent but the doctor reports present, and conversely, and the proportion for whom both the patient and the doctor report the complaint present. Then present some weighted overall result, say standardized over status at entry (e.g. Table 5).

In Table 5 $p_{i...}$ is the proportion of patients in treatment group $i$ with complaint present at entry, $a_{i...}$ is the proportion with complaint absent at entry and $x$ is the overall proportion of patients with complaint present at entry (which could be defined as either the patient or the doctor reporting the complaint present).

The three groups (patient, doctor) = (0, 1), (1, 0) and (1, 1) could be combined to simplify the comparison to complaint absent versus complaint present in the study.

R. J. Stephens, S. P. Stenning, M. K. B. Parmar and D. Machin (Medical Research Council Cancer Trials Office, Cambridge): It is recognized that in many cancer sites there is often very little prospect that new therapies will offer large gains in survival of patients. In such circumstances differences in quality of life (QOL) become increasingly important. There are clear advantages in standardizing the

![Graph](image_url)

Fig. 5. Changes in anxiety scores obtained from the HAD scale in patients grouped according to the number of assessments made before death.
measures of QOL across successive trials and following the recommendation of Maguire and Selby (1989) we have been using the Rotterdam symptom checklist (de Haes et al., 1990) in combination with the hospital anxiety and depression (HAD) scale (Zigmond and Snaith, 1983). Using these instruments we have collected QOL data on more than 1500 cancer patients in current randomized trials.

In cancer trials intercurrent deaths and non-compliance may provide a strong and potentially informative censoring mechanism for QOL studies. To illustrate the influence of this censoring we describe some exploratory analyses from a randomized trial comparing two palliative chemotherapy regimens in patients with lung cancer. QOL is important for these patients during their short survival period, and between-treatment comparisons need to continue well beyond the median survival time.

Fig. 5 shows the mean anxiety profiles (from the HAD) obtained from all those patients that have so far died (those still alive are omitted to simplify the presentation) grouped by the number of assessments that their lifespan allowed. In Fig. 5 higher scores represent greater levels of anxiety. Fig. 5 suggests that the mean anxiety scores are approximately parallel for the five death-censored groups. This may then allow us to obtain a pooled estimate of ‘anxiety’ by simply averaging the results at each assessment from all the available patients. If the profiles were not parallel then it would be necessary to summarize anxiety for each censored group separately. However, the mean profiles themselves may disguise substantial variation between patients. This is illustrated in Fig. 6 for the patients who only completed their first three follow-up assessments.

We can conclude from this simple exploratory approach that analysing QOL data, whether for individual or combined measures, is not straightforward. In particular, the need to allow for informative censoring causes practical analytical difficulties which may only be detected if data from a large number of patients are available.

George W. Torrance (McMaster University, Hamilton): Cox and colleagues identify four contexts in which health-related quality of life (QOL) is measured:

(a) population health,
(b) resource allocation,
(c) clinical trials and
(d) clinical decision-making.

They suggest that the different contexts have different objectives and, therefore, may require different QOL measures. For example, they specifically suggest that resource allocation and clinical decision-making require a single summary measure of QOL, whereas in clinical trials results are best left
multidimensional. Their reason is that this makes ‘it straightforward for future users to weigh up the alternatives with their own, currently unspecified, values’ (p. 2).

It is useful to carry on this line of reasoning. The four contexts are of two specific types—(b) and (d) involve decision-making, whereas (a) and (c) involve information, presumably to aid decision-making in (b) and (d), i.e. the purpose of the QOL information from a clinical trial is to aid decision-making by the ‘future users’. These may be clinical decisions being made by a patient and physician, or resource allocation decisions being made by the responsible decision maker. Clinical trials may often be called on to provide the relevant QOL information for both of these uses.

In this paper, Cox and colleagues focus only on the traditional use of clinical trials to provide information for clinical decision-making. Thus, for example, they advocate disease-specific QOL instruments. If trials are also to provide QOL information for resource allocation decisions, they will require generic QOL instruments that enable comparisons across diseases and programmes.

The authors criticize quality-adjusted life years (QALYs) as inappropriate for clinical decision-making, clinical trial reporting and resource allocation. QALYs were never intended for clinical decision-making—they were developed for use in resource allocation. They would be appropriate for clinical trial reporting if they are useful for resource allocation. Their appropriateness for resource allocation is an empirical question—do they provide useful information to the responsible decision maker?

They may. Ultimately, in resource allocation, the responsible decision maker must formally or informally combine the QOL, the length of life and the number of individuals affected to reach a decision. A good QALY analysis should show all the results first in disaggregated form, then a template for combining the results using values, followed by the combined results using specific values gathered carefully from appropriate individuals and finally a thorough sensitivity analysis and thoughtful discussion, including caveats. The goal is simplicity and transparency. Such a QALY analysis may indeed be informative.

**Nanny Wermuth (University of Mainz):** The authors are to be congratulated on their illuminating discussion of available measures of quality of life. I expect it to be most useful for everybody who wishes to construct or interpret summary measures, i.e. scales, based on responses to questionnaires.

The paper contains not only a lucid description of why are such measures wanted?, what are their desirable properties?, how can standard principles for design of experiments aid in obtaining useful responses?, which analyses will provide basic insights?, but also a most convincing and devastating critique of using complex schemes of weighting individual answers to obtain a single overall score.

Simple weighting schemes have also aided interpretation in a different context (Cox and Wermuth, 1992; Wermuth and Cox, 1991) where it is proposed to compute derived response variables having special relations of linear conditional independence with a set of explanatory variables.

I appear to disagree with the authors on two specific points: on ‘subgroup analysis’ and on ‘standard scales’.

**Subgroup analysis**

Though warnings of misuse are helpful, I find that much insight can often be gained by subgroup analyses. One example concerns blood glucose control of diabetic patients (Kohlmann et al., 1991). A formally correct but uninformative summary of a linear regression analysis for 68 patients would be that a particular metabolic parameter, the glycosylated haemoglobin (Y, GHb), depends directly on knowledge about the illness (X1, score), duration of illness (X2, months) and number of years of formal schooling (X3). Table 6 shows instead data summaries for two subgroups of patients. From these and the plots in Fig. 7, we can conclude that for all patients with less than 8 years of illness the level of education discriminates almost perfectly between a good metabolic adjustment (GHb < 9)

```
<table>
<thead>
<tr>
<th>Results for at most 10 years of formal schooling</th>
<th>Results for at least 13 years of formal schooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 38</td>
<td>n = 30</td>
</tr>
<tr>
<td>( \bar{y} = 9.90 )</td>
<td>( \bar{y} = 8.64 )</td>
</tr>
<tr>
<td>( s_y = 1.96 )</td>
<td>( s_y = 1.88 )</td>
</tr>
<tr>
<td>( \bar{x}_1 = 33.26 )</td>
<td>( \bar{x}_1 = 38.10 )</td>
</tr>
<tr>
<td>( s_{x_1} = 7.95 )</td>
<td>( s_{x_1} = 5.55 )</td>
</tr>
<tr>
<td>( \bar{x}_2 = 143.03 )</td>
<td>( \bar{x}_2 = 101.03 )</td>
</tr>
<tr>
<td>( s_{x_2} = 87.96 )</td>
<td>( s_{x_2} = 76.47 )</td>
</tr>
<tr>
<td>( r_{x_1} = -0.343 )</td>
<td>( r_{x_1} = -0.117 )</td>
</tr>
<tr>
<td>( r_{x_2} = -0.560 )</td>
<td>( r_{x_2} = 0.302 )</td>
</tr>
<tr>
<td>( r_{12} = 0.063 )</td>
<td>( r_{12} = -0.227 )</td>
</tr>
</tbody>
</table>
```

This content downloaded from 163.1.41.27 on Fri, 07 Apr 2017 11:20:40 UTC
All use subject to http://about.jstor.org/terms
and a poor adjustment. This appears to call for a better level of communication between physicians and patients already in the early years of the illness.

**Standard scales**

Though desirable properties of scales are readily listed, I think that we have to admit that statisticians have failed to provide widely accepted methods for obtaining scales with such properties. Therefore it still appears rather unclear when we can safely speak of ‘standard evaluated scales’. I am convinced that no simple methods will be appropriate for constructing good scales and for establishing their high quality, especially if it is likely that responses depend on cultural context, on knowledge, or on the education of a respondent.

**Alan Williams** (University of York): Although the authors warn against regarding their ‘practical recommendations’ as ‘tablets of stone’, their collective professional experience and reputation guarantee that their views for or against particular methodological positions will be cited as authoritative in this rapidly developing field. It is therefore important to note the self-imposed limitations of their review, namely that they are concerned only with ‘QOL assessment in clinical trials comparing treatments’, which ‘differs from the resource allocation and individual treatment decision contexts’, where ‘any QOL assessment has often been collapsed to a single scale to allow direct comparison between options’. This narrowing of their scope leads them to recommend that ‘clinical trial results are best left multidimensional and reported so as to make it straightforward for future users to weigh up the alternatives with their own, currently unspecified, values’.

They concede, however, that this still does not enable them to escape the problem of weighting items and/or dimensions, and after criticizing those who seek to derive such weights from patients’ own relative valuations of different health states they settle for an entirely arbitrary all-purpose integer scale. This will have all the equivalence problems with which they castigate Torrance, plus the additional difficulty that it has no external justification beyond ‘appearing reasonable’. But this defence is also presumably available to those who systematically seek patients’ valuations, a procedure which appears even more reasonable than (though certainly not as simple as) their solution. An unweighted average of percentile scores within equally weighted dimensions will only coincidentally correspond (even in direction of movement, let alone magnitude of change) with the views of patients, and yet it is envisaged that such a scale would be used to choose the better treatment in a clinical trial. Should this be encouraged when patients’ values play no part in the assessment?

Leaving the difficult bits of QOL measurement to some unspecified future user (who would have to unpick the entire ‘simple’ weighting scheme to use the results in such a way as to reflect patients’ values) may be ‘simple’ and ‘practical’ for the researchers, but what evidence is there that a future user could do a better job in filling the gaps than the research community itself? Should not the message be to encourage more work on the sophisticated (but relevant) weighting systems, and in the meantime to regard all simple systems as suspect, rather than the other way round?

---

**Fig. 7.** Dependence of metabolic adjustment \((Y, \text{GHb})\) on duration of illness \((X_2)\) for diabetic patients with (a) at most 10 years and (b) at least 13 years of formal schooling.
We are very grateful to all the contributors for their constructive and encouraging comments, which illustrate the wide ranging implications of the topic. We were particularly glad that several specific applications were described. We shall not comment on these but they add much to the paper.

The broad qualitative aspects are in most respects more important than the technical details of statistical analysis so it is not surprising that many of the comments concerned these issues of principle. The objective of providing evidence for sensitive choice by patients, requested so eloquently by Mrs Goodare, is an objective that we wholly share. Dr Hopkins's argument for the isolation of aspects considered crucial by individual patients is important; scientific progress, however, depends on public discussion and transferability of results and for this some use of standard instruments seems unavoidable. In addition scientific understanding, and hence most long run progress, may require a careful study of the relation between quality-of-life (QOL) and biochemical measurements, blood pressure, etc. and for this standardized measurement again seems essential.

Ms Farquar and to some extent Dr Hutton are concerned about the general philosophical concept of QOL. We agree that we are dealing here with a very limited aspect of something probably too nebulous for fruitful study as a whole; in many ways a much less pretentious term than QOL would be preferable for our purposes.

Our paper focuses principally on clinical trials and also on the implications for individual patient choice, as have other recent contributions (Feeny et al., 1991). It is not surprising, however, that the discussion broadened into medical audit and resource allocation. Many contributors, Sir Jack Hibbert in his thoughtful warning about pressures from policy makers, Dr Carr-Hill, Mr Sheldon, Dr Parr and Mr Fayers, shared our generally sceptical attitude to quality-adjusted life years (QALYs), Professor Torrance and Professor Williams being the exceptions. Mr Sheldon and Professor Torrance and Professor Bulpitt refer to a sentence in Section 1 which they interpret as supporting the view that a single summary measure of QOL is essential both in the context of decision-making for the individual patient and in resource allocation. That this is not our view in the resource allocation context will be clear from the arguments in Section 5.2 and elsewhere, and we have taken the opportunity to amend the misinterpreted sentence slightly in proof to avoid further confusion.

The role in audit, mentioned for example by Dr Ashby, is less controversial; here the primary requirement at the moment seems to be to set up procedures for the collection of reliable and meaningful data and many of the general points about design in the paper are apposite.

Several contributors have taken up the question of combining QOL measures with survival times. We agree with Dr Abrams that multistate models, which have a long history (Fix and Neyman, 1951) can be useful although, even when QOL is collapsed onto a small number of states, the possible types of dependence in transition rates are considerable once we move away from simple Markov processes to the various kinds of semi-Markov process (Cox, 1986). Such models do, however, have the advantage of allowing the study of the effect of explanatory variables on transition rates. The case made out in the paper for not combining survival time and QOL into a single measure does not mean that the two should be considered only separately, but rather that some incisive way should be found for describing their joint distribution; see Sections 6.4.1 and 7 of the paper. The best methods may well depend strongly on context. Dr Gelber, Dr Newcombe and Professor Torrance describe possible approaches. Dr Newcombe's 'pessimal substitution' methods may prove valuable when the follow-up period is short, but even then may be unduly influenced by outlying (low) QOL scores among survivors. We welcome the suggestions by Dr Gelber and Professor Torrance of different approaches to QALY analyses, which involve keeping the constituent dimensions essentially disaggregated until the time that a decision has to be made, with emphasis on the importance of sensitivity analyses at that stage.

We agree with Dr Ashby's implicit comment that, as in so many statistical problems, the issues of formulation, hopefully leading to the asking of the right question, are more important than issues of philosophical approach to formal inference.

We are glad that Dr Senn found the discussion of components of variance helpful, in particular in clarifying the relation between 'average' conclusions and those for individuals. Professor Wermuth's enlightening discussion and specific example illustrate a different kind of interaction. Dr Solomon's interesting suggestions are also relevant to this theme. Although the key ideas of components of variance are 50-60 years old at least, they, and the related notions of components of covariance, seem to be underused in such contexts. There is a connection also with Dr Ashby's concern with errors of measurement.
The role of relatively formal psychometric methods was defended in particular by Dr Olschewski and by Dr Kreiner and Professor Keiding. The value of, for example, the Rasch model and its generalizations can be considerable in developing new instruments but we stick to our general point that in specific trials with established instruments simple methods should be used. This is challenged by Professor Williams who, correctly, points out the arbitrariness of simple scoring and unweighted addition. Our defence of simple scoring briefly is that

(a) the effect of any single item on the overall score is easily assessed,
(b) any increased weighting of a particular item is easily achieved but would have to be explicitly justified,
(c) 'accidental' domination by one or two items is avoided,
(d) what we regard as illusions of objective precision are avoided (see also Dr Parr's first point) and
(e) less importantly, technical statistical assumptions of normality etc. are more likely to be satisfied, although we do not agree with Dr Olschewski that our analyses assume or require normality.

We agree with Professor Williams that further research on all these issues is desirable and re-emphasize the need for formal or informal sensitivity analysis.

Finally we agree strongly with those contributors, in particular Professor Bulpitt, Dr Girling and Dr Stephens et al., who among other things stress the importance of careful measurement.

REFERENCES IN THE DISCUSSION


