METHODS FOR QUALITY-OF-LIFE STUDIES

Marcia A. Testa
Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115

Johanna F. Nackley
Phase V Technologies, Inc., Wellesley Hills, Massachusetts 02181

KEY-WORDS: quality-of-life methods, patient outcomes, clinical trials, statistics

INTRODUCTION

The interpretation of quality-of-life data for therapeutic decision making and policy planning requires a comprehensive understanding of the methodologies employed in the design, data collection, and analysis phases of the research. The assumptions and limitations of the quality-of-life measures and the associated methods of study design and data analysis should be carefully reviewed prior to using quality-of-life outcomes as dependent variables of interest. To date, the primary focus of quality-of-life, methodologic research has been on defining the construct of “quality of life” (3, 57, 69, 93), deliberating its role in outcomes research (71, 78, 87), and choosing the appropriate measurement instruments (4, 40, 70). Much less attention has focused on the design and statistical methodology used to substantively evaluate and interpret quality-of-life treatment effects and differences. And yet, rigorous evaluation of therapeutic interventions such as pharmacologic treatment, surgery, and preventive programs depends almost entirely upon the adequacy and appropriateness of the methodologies employed and the manner in which analytical models are used to interpret the results. Without adequately defining and applying the methodologic assumptions, analytical

1Publication of this article does not necessarily represent approval, concurrence, or official endorsement of the statements and opinions herein by the AHCPR, the Public Health Service of the US Department of Health and Human Services.
models, and statistical techniques, valid conclusions regarding treatment
effects and the relative risks, benefits, and costs of alternative therapies are
difficult to achieve, if not untenable.

Analytical and statistical methodology in quality-of-life research is fragment­
ated and lacks careful integration among the diverse but related fields
of health, behavioral, social, therapeutic, and quantitative research methods.
Techniques to measure quality of life are drawn from evaluation, health
survey, nursing, clinical, psychometric, and sociometric research. Study
designs are adapted from survey, observational, and interventional method­
ologies. Statistical methods are often based upon simple linear models and
treatment contrasts, but more sophisticated techniques include multivariate
linear models, tests of global hypotheses, longitudinal methods of data
analysis, nonparametric rank statistics, failure time and survival methods,
and health-state Markov processes. The objective of this report is to evaluate
current measurement, design, and statistical methodology used in quality­
of-life outcomes research by critically reviewing the published literature in
the field. Because quality-of-life research spans many diverse disciplines,
this review focuses on methodologies related primarily to medical and public
health interventional and observational studies, with an emphasis on thera­
peutic clinical trials.

Quality of life is widely perceived to have substantial potential as an
endpoint in medical and health-outcomes research. Health-related quality­
of-life endpoints are being used increasingly to evaluate pharmacologic
agents in clinical trials of cancer (50, 64), HIV (81, 91), arthritis (5, 59),
heart failure (11, 92), and hypertension (10, 17, 47, 83, 85). They are also
being employed to adjust measures of effectiveness for therapeutic decision
making and to plan allocation of resources (24, 29, 32, 34). In addition to
the primary statistical methods used in estimating the quality-of-life param­
eters, major methodologic issues arise in secondary analyses that attempt
to incorporate the quality-of-life estimates into adjusted measures of "effec­
tiveness".

The relevant methodologies used in quality-of-life outcomes research must
address the design, measurement, analysis, and interpretation of the qual­
ity-of-life measures and the corresponding effects of treatments. With regard
to these issues, Guyatt et al (39) asked two relevant and practical questions:
(a) how should health-related quality-of-life measures be compared, and (b)
how can we make health-related quality-of-life results from controlled trials
meaningful to the intended audience? To address some of the specific issues
of design, measurement, and analyses, the methodologic areas involving
quality-of-life measurement, estimation, hypothesis testing, and use of
summary indices in effectiveness models will be reviewed from both a
qualitative and quantitative perspective.
MEASUREMENT OF QUALITY-OF-LIFE OUTCOMES

The measurement properties of the quality-of-life indices and scales used in therapeutic trials affect their ability to detect meaningful treatment differences. These properties are a function of both the theoretical framework from which the quality-of-life constructs are derived, and how well the scales perform in measuring those constructs. A scale can be evaluated by several indices that measure performance, including reliability, validity, responsiveness, and sensitivity. A scale's level of performance can have a profound impact on the conclusions for a particular clinical trial or case-control study. In addition, the appropriate use of summary quality-of-life estimates in subsequent models of cost-effectiveness and cost-utility of alternative drug, interventions, and treatments (84) requires an understanding of how performance levels can influence secondary analyses of overall treatment effectiveness.

Conceptualizing Quality of Life

Our ability to measure quality of life depends to a great extent upon how it is conceptualized (7, 68). In medical and health survey research, the term quality of life is an organizing concept that brings together a set of domains related to the physical, functional, psychological, and social health of the individual. When used in this context it is often referred to as "health-related quality of life" to differentiate it from its use in other contexts, including references to the level of crime, adequacy of housing, fairness of taxes, and cultural environment. An extremely comprehensive and thorough review of the field of quality-of-life research as it pertains to the development of health policy is given by Patrick & Erickson (72). Table 1 depicts the process of translating the concept of quality of life into more measurable constructs.

According to the representation in Table 1, health-related quality of life involves the five broad dimensions of opportunity, health perceptions, functional status, morbidity or impairment, and mortality. Depending upon the specific target population and the purposes of the study, it may not always be necessary to measure all dimensions to fully evaluate quality of life, if certain assumptions hold. For example, in a multicenter, randomized, double-blind clinical trial comparing the effects of antihypertensive medications (captopril, methyldopa, and propranolol) on quality of life, the primary focus was on the domains of general health perceptions, functional status, and self-reports of symptoms (17) because the other dimensions were assumed to be relatively constant. Randomization was used to ensure equal balance for opportunity, the disease was asymptomatic, and all patients were titrated to produce similar effects for efficacy and safety, with the assumption
Table 1  A multidimensional conceptualization of quality of life

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Domains</th>
<th>Indicators (Indices, Scales, Subscales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opportunity</td>
<td>Social or cultural</td>
<td>Access to care, societal stigma, support</td>
</tr>
<tr>
<td></td>
<td>Coping</td>
<td>Ability to withstand stress, psychological or physical</td>
</tr>
<tr>
<td>Health Perceptions</td>
<td>General health perceptions</td>
<td>Self-rating, worry, concern</td>
</tr>
<tr>
<td></td>
<td>Expectations/satisfaction</td>
<td>Satisfaction with functioning</td>
</tr>
<tr>
<td>Functional</td>
<td>Social</td>
<td>Work and daily role</td>
</tr>
<tr>
<td></td>
<td>Psychological</td>
<td>Distress (anxiety, depression, loss of behavioral and emotional control)</td>
</tr>
<tr>
<td></td>
<td>Cognitive</td>
<td>Well-being (positive affect, emotional ties, life satisfaction)</td>
</tr>
<tr>
<td></td>
<td>Physical</td>
<td>Memory, alertness, reasoning</td>
</tr>
<tr>
<td>Morbidity</td>
<td>Signs</td>
<td>Objective clinical findings directly observable</td>
</tr>
<tr>
<td></td>
<td>Symptoms</td>
<td>Subjective evidence indirectly observable</td>
</tr>
<tr>
<td></td>
<td>Self-reports</td>
<td>Patient self-reports of symptoms and conditions</td>
</tr>
<tr>
<td></td>
<td>Physiologic</td>
<td>Laboratory measures, pathology</td>
</tr>
<tr>
<td></td>
<td>Diagnosis and severity</td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>Death</td>
<td>Survival, years of life lost</td>
</tr>
<tr>
<td></td>
<td>Duration of life</td>
<td></td>
</tr>
</tbody>
</table>

1Adapted from a model proposed by Patrick & Erickson (72)

that length of life and mortality would be comparable in all groups. However, in other studies, equal efficacy among therapies might not be a reasonable assumption and therefore morbidity and mortality would have to be factored into the evaluation model. The range of dimensions and the added requirements for disease-specific evaluation should always be considered.

The range of responses and extent of coverage of domains will influence the accuracy of the measures, especially as they reflect changes in quality of life during intervention studies. For example, a generic health survey instrument might be sensitive to picking up distinctions between individuals who differ by clinical status; however, a more disease-specific questionnaire might be needed to detect changes within individuals in a specific clinical class. Turner approached the issue of measurement using a three-dimensional model to evaluate patient outcomes in rehabilitation (89). The first dimension includes the areas of assessment such as impairment, disability, and handicap. The second dimension included the domains of assessment that are generally accepted as relevant for rehabilitation outcomes (physical, mental, emotional, and social), and the third dimension, the type of measure classified according to its applications as evaluative, predictive, and dis-
METHODS FOR QUALITY-OF-LIFE STUDIES

criticize (54). The model is important because it relates the measure's purpose, structure, and application.

For example, a functional health status measure commonly used in cancer clinical trials such as the Karnofsky Performance Status Scale (52), covers a very broad range of functioning from death (0 points) to full health (100 points indicating normal: no complaints: no evidence of disease). The average asymptomatic patient with hypertension or who is HIV-positive might only be able to function daily to his or her satisfaction between a score of 90 (able to carry on normal activity, minor signs, or symptoms of disease) and 100. Changes of 10 points might have a very large impact on the ability of the asymptomatic patient to function and might reduce his or her usual day-to-day functional capacity by more than 50%, even though 10 points only represents 10% of the total Karnofsky scale. Hence a seemingly "small effect" under one set of assumptions of 10 percent can translate into a very large effect of 50 percent based upon individual expectations. Small differences between individuals on the overall scale can translate into relatively large differences within individuals because they are calibrated against the patient's own internalized expectations of quality of life.

Three properties of quality-of-life measures influence how and how well the constructs are measured and analyzed. As depicted in Table 1, the hypothetical constructs and corresponding measurement scales are multidimensional and multilayered. Thirdly, they are most often measured indirectly. To address the multidimensional property of the measures one must address the multivariate nature of the data and the multiple and global hypotheses that they will generate. Alternatively, one must find ways to combine the various constructs and domains through the use of summary parameters or utility weighting procedures that make it possible to incorporate quality of life into general models of risk-benefit and cost-effectiveness.

In addition, the measures are multilayered and nested. That is, we measure constructs by forming single questions or items; these are grouped into sub-scales, which, in turn, form broader scales that may themselves be part of even broader composites jointly reflecting overall quality of life. Hence, sub-scales are nested within scales, which are nested within composites, which are nested within domains, which are nested within the major dimensions. What level does one use for the primary analysis? Certainly the layers and dimensions are not independent, with inner layers being subsets of outer layers. Summary parameters from factor analysis or other linear functions of the individual layers attempt to deal with both the multilayering and multidimensionality.

The third property of quality-of-life scales involves the indirect nature by which one must measure the quality-of-life constructs. In addition to the random variability between individuals associated with most biomedical
variables which can be measured directly, indirect measures contain other random effects associated with the inherent variability within questionnaires/forms, between forms, within-subject, between interviewers, and over time.

The Validities of Health-related Quality-of-life Measures

Validity is defined as the capacity of the instrument to measure what it intends to measure. Kirshner & Guyatt (54) recognized that the performance indices or the "validities" of a scale have different implications across the three types of quality-of-life measures, namely, evaluative, predictive, or discriminant. While discriminative indices should demonstrate large and stable between-subject variation so that groupings of similar patients can be contrasted to other groupings, evaluative indices must possess high test-retest reliability and responsiveness to true changes in quality of life. Much of the current debate surrounding the interpretation of the meaning of quality-of-life data is due to applying the same performance standards appropriate for discriminate measures to evaluative measures for which such standards are inappropriate.

Nonintervention-Based Validities

The validities of a scale tell us how well our measure quantifies the true response (62). We assess validity according to the degree of relationship of a test or scale with another variable of interest. The validity coefficient of a measurement $Y$ with respect to a second measurement $X$ is defined as the absolute value of the correlation coefficient

$$\rho_{yx} = \frac{\sigma_{yx}}{\sigma_y \sigma_x}$$

where $\sigma_y, \sigma_x$ and $\sigma_{yx}$ are the corresponding standard deviations and covariance. As seen here, the validity coefficient of a measurement can only be stated in relation to a second measure. Thus, although claims are often made concerning the fact that a quality-of-life scale has been "validated", one cannot speak meaningfully about a "valid quality-of-life scale" or a "valid quality-of-life instrument" without first defining the measure $X$ against which the scale $Y$ is being correlated or compared.

Nonintervention-based validities tell us how well the measures are functioning at a single point in time. For example, reliability, including measures of within-form homogeneity of items (internal consistency) and between-forms stability (test-retest), are indications of how precise and stable the measures are. Construct validity, including convergent, trait, and discriminant validity, assess how well the scales relate to other measures that purport to measure the same construct. A thorough review of measurement theory can be found in a number of texts (63, 80).
Methods for Quality-of-Life Studies

Intervention-Based Validities

The accurate measurement of change in quality-of-life scales is paramount in interventional studies of health (37, 38, 41). To this end, researchers are interested in measuring both the overall effects of treatments as well as in examining the individual differences in treatment response. In interventional and observational studies quality of life is usually measured longitudinally and involves multidimensional scales describing the physical, emotional, and social health status of the patient. While the reliability of the scales is established by evaluating internal consistency (6, 16) and test-retest reproducibility as cited above, the ability to detect meaningful treatment effects is more difficult to define and evaluate.

The instrument's responsiveness (37) to true change and its sensitivity to treatment effects are being used increasingly for validation of evaluative measures such as those used to assess changes due to intervention. Knowledge of a scale's responsiveness and sensitivity are essential in determining both the power of the statistical analyses and the interpretation of the findings.

A demonstration of responsiveness is necessary in a therapeutic intervention study because it helps to discriminate between scales that are likely to change and those that are not. This knowledge is especially important when designing "equivalence trials" or when attempting to prove the null hypothesis of no treatment effect. One could easily be unaware of a large Type II error (assuming no difference, when a true difference actually exists) because the error could be due to low responsiveness rather than small sample size. Recently, Guyatt (38) and Tuley (88) proposed a responsiveness index using the change from baseline to the final evaluable double-blind visit. The index is a scaled measure of this change and is defined as the ratio between the mean change from baseline to endpoint after treatment and the standard deviation of change within untreated, stable subjects.

Low responsiveness, lack of sensitivity, and significant confounding are often not taken into account when designing a clinical trial. When these omissions occur, conclusions based upon acceptance of the null hypothesis of "no effect" can be extremely misleading. Alternatively, even highly statistically significant treatment effects have been criticized because it is difficult to relate the quality-of-life change units employed into a known or measurable health or social consequence. Do the statistically significant changes reflect a substantial change in quality of life, or one that is minimal? In a cost-effectiveness analysis conducted by Edelson (24) comparing anti-hypertensive medications, the incremental cost effectiveness (per quality-adjusted life year gained using the more expensive therapy) varied by as much as $1.4 m based upon reductions in quality of life as small as 1
percent. What was so striking about this quality-adjusted, cost-effectiveness analysis was the very narrow range of quality-of-life changes that caused a reversal in the final conclusions (84). It is certainly possible that the estimates of incremental costs suffer from high error when they attempt to distinguish between such small quality-of-life effects.

How responsive the quality-of-life scale is to therapeutic effects is a function of the validity and precision of the scale and the magnitude of the actual drug effect. Using Guyatt's responsiveness index, Testa et al (85) computed the quality-of-life change from baseline to the final evaluable double-blind visit to represent the response of the patient to treatment. Most importantly, the investigators then went on to evaluate the sensitivity of their measures to detect differences in treatment response by analyzing the longitudinal changes between drug treatments in relationship to the scale's responsiveness to changes in stressful life events as reported by patients during the course of the study. As shown in Figure 1, changes in General Perceived Health could be calibrated to stressful life events. Pooled data from two clinical trials of antihypertensive therapies in elderly men found that a 0.1 responsiveness unit change (approximately .06 between-individual standard deviation units) on the General Perceived Health scale corresponded to 32 Life Change Units (LCUs) on the Holmes and Rahe Social Readjustment Scale (48; MA Testa, unpublished observations). Therefore, a quality-of-life change of this magnitude could be translated into an effect size comparable to changing to a different line of work (36 LCUs) or major change in arguments with spouse (35 LCUs).

While the published research supporting a negative relationship between stressful life events, disease, and mental health was abundant (8, 9, 43, 44, 53, 60, 86, 94), using this relationship to estimate the responsiveness and sensitivity of the quality-of-life measures represents new ground for intervention-based validation for comparative clinical trials. However, the methods had been established and used much earlier in observational studies of medical outcomes (7, 101). In the HIS study by Brook et al, investigators found that for the subscales of general health and vitality (two of the three subscales used in the General Perceived Health in the study by Testa et al (85) cited above), a 31-point LCU change resulted for every .08 between-individual standard deviation change in those subscales. Despite the differences in the study design, and length of follow-up, the calibrations were remarkably similar to the study by Testa et al (85), indicating the stability of the relationships between stressful life events and functional health. Such findings make objective stressful life events an excellent criterion variable for assessment of the responsiveness of a quality-of-life scale. Other studies have assessed the responsiveness of health status measures to change using relative efficiency statistics (a ratio of paired $t$ statistics) (58), and receiver-
Figure 1  Linear trend in General Perceived Health (responsiveness units change since last visit) as a function of stressful life events (LCU score change reported since last visit). Note that changes in the range of 0.10 to 0.20 responsiveness index units reflect the influence on quality of life of important life events such as loss of a job. (Reprinted from Testa et al (85) with permission.)

operating characteristic curves that describe the ability of a scale to detect improvement using an external, dichotomous criterion (21, 22).

Respondent Variability

The question of validity brings up a related issue of intra- and interrespondent variability. Jachuck first noted that patients, relatives, and physicians all had different perceptions of what constituted improvement in the patient’s quality of life during antihypertensive therapy (49). Physicians consistently rated patients as having improved, whereas relatives uniformly agreed that
they had worsened. The sources of variability associated with such factors as who is reporting, how reports are obtained, and where assessments are conducted all contribute to respondent variability.

In a study comparing nifedipine GITS and atenolol on aspects of the patient’s quality of life, three respondents, patient, clinician, and spouse, were used to assess between-form interrater variability (83). The spouse was a more sensitive discriminator of treatment differentials in two areas, predicting withdrawal from the study and sexual dysfunction. For the physician, the correlation between the frequency of each symptom reported by the patient and the frequency reported by the physician was fairly high ($r = .81$). However, the physician had a much higher threshold for reporting side effects, related to their severity, specificity, and medical implications (1). The percent of symptoms present by usual safety reporting (spontaneous reports by physicians) was less than 10 percent of that reported by patients, and physician checklists still only provided 25–33% of the levels reported by patients. These results indicate that even clinically straightforward measures such as adverse events and symptoms contain a large portion of subjectivity, which is an important component of the patient’s quality of life.

Parallel forms developed for the patient to complete in the clinic (using a long-form version) and at home (using a shorter-form version) were used to study intra-individual, between-form variability (82). The take-home short forms were nearly identical in reliability to those administered in the clinic, as measured by Cronbach’s standardized alpha coefficient; but the clinic-based assessments proved to be the more sensitive discriminator for the quality-of-life treatment differentials overall.

**HYPOTHESIS TESTING**

One of the weakest areas in quality-of-life research is hypothesis testing and the subsequent interpretation of the results. Conflicting reports and claims are made due to a lack of understanding of design and analysis methodologies, including (a) the role of measurement error on subsequent statistical tests; (b) inadequate power; (c) failure to deal with confounding effects such as age, sex and severity of illness; (d) length of study bias; (e) early withdrawal from the study; and (f) problems with missing data.

**What Constitutes a Meaningful Quality-of-Life Treatment Effect?**

When making medical decisions or when designing and adopting summary parameters for cost-effectiveness and cost-utility analysis studies, it is often difficult to determine what constitutes a meaningful quality-of-life treatment
**METHODS FOR QUALITY-OF-LIFE STUDIES**

Even relatively small changes can have a fairly large impact on the incremental cost-effectiveness of different therapeutic alternatives. Similarly, claims of improved quality of life, stable quality of life, or a relatively better quality of life imply knowledge concerning how great the impact will be for an effect of a predetermined magnitude. How do clinicians, policy planners, and health economists decide what is a "small" or a "large" effect? This question bears directly on very practical and important design and statistical issues. How do we set appropriate effect sizes for alternative hypotheses such that our hypotheses test what we want them to test? What is the magnitude of a quality-of-life decrement that has a dispositive effect on the individual and subsequently on society? These questions must be answered before valid interpretation of quality-of-life data can be made.

**Estimating the Power of a Quality-of-Life Study**

As cited previously, Brook et al conducted extensive research on health status measures in the Health Insurance Study (HIS) involving 3924 individuals (7). When studying the power of quality-of-life measures they recommended that external criteria be reported along with the usual measures of effect sizes (standard deviations and percent change in the mean). Brook and colleagues determined that the HIS study should be powerful enough to detect differences for all physical health measures in terms of the effects of five years of aging, and for all mental health measures, differences equal to the amount of stress associated with a small debt [17 (LCU) on the Holmes-Rahe scale]. Thus they defined the associated changes in quality of life that would accompany these events as being clinically meaningful and important.

Table 2 summarizes the results of the HIS power analysis and the calibration with external criteria for a number of scales. The data indicated that the HIS study had the prerequisite power to detect fairly small between-individual standard deviation effect sizes of between .07−.08 units. For a discriminant measure used to classify patients, tests powered to detect effects of .07−.08 between-individual standard deviations would seem to be very powerful. Cohen’s guidelines for cross-sectional data for the relative magnitude of effect sizes for behavioral measures (15) recommend .2 between-individual standard deviations as a small effect, .5 as a medium effect, and .8 as a large effect. In contrast to these recommended values, the effect sizes in Table 2 between .07 and .08 do appear relatively very small. However, as also pointed out by Cohen, the relative magnitude of longitudinal effect sizes required for detection by evaluative measures should be based upon within-individual standard deviations and will vary from those required by descriptive measures.

In the HIS study the minimally clinically significant effect of 17 LCUs
Table 2  Summary of the statistical power and calibration between longitudinal effects (between-individual standard deviations units) and external criteria (years of aging and the Holmes and Rahe Life Change Units (LCU) scale) in the Rand Health Insurance Study

<table>
<thead>
<tr>
<th>Scale</th>
<th>Effect sizes</th>
<th>External criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Limitations</td>
<td>.073</td>
<td>3.6 years</td>
</tr>
<tr>
<td>Self-Care</td>
<td>.075</td>
<td>10.1 years</td>
</tr>
<tr>
<td>Function Status Index</td>
<td>.071</td>
<td>3.5 years</td>
</tr>
<tr>
<td>Positive Well-Being</td>
<td>.080</td>
<td>30.7 LCU</td>
</tr>
<tr>
<td>General Health</td>
<td>.081</td>
<td>30.9 LCU</td>
</tr>
<tr>
<td>Vitality</td>
<td>.081</td>
<td>30.4 LCU</td>
</tr>
<tr>
<td>Depression</td>
<td>.087</td>
<td>19.9 LCU</td>
</tr>
<tr>
<td>HIS-GWB</td>
<td>.081</td>
<td>20.4 LCU</td>
</tr>
<tr>
<td>Mental Health Index</td>
<td>.080</td>
<td>19.9 LCU</td>
</tr>
</tbody>
</table>

The HIS Health status measures had the power to detect effect sizes of this magnitude as computed with $\alpha = .05$, $\beta = .10$, assuming pretests and covariates and a sample of 3924 adults.

*Source: Brook et al (7)*

would correspond to between .04-.08 between-individual standard deviations depending upon the particular scale. Note that even if the .2 "small effect" guidelines for the estimation of power had been applied, the HIS study would have missed changes in general health corresponding to 75 LCU, which is approximately equivalent to the impact of divorce (73 LCU) on general health.

**Simultaneous Test Procedures**

Three types of questions might be asked with respect to multivariate quality-of-life data: (a) multiple univariate; (b) structured multivariate; and (c) intrinsically multivariate. As described by Hand & Taylor (42), “the last case is addressing the question of whether or not a comparison is significant on a group of variables simultaneously, and the individual identities of the variables is regarded as irrelevant to the basic question”. Such is the case when answering the question of whether a “quality-of-life” difference exists. Because it is the “construct” quality of life that is under investigation and not whether each of its subcomponents is different, primary attention should focus on the intrinsically multivariate question. However, should the intrinsically multivariate question show significance, then further questions would be appropriate: on which variables or combinations of variables is the comparison significant? Since by design, each of the measures that make up quality of life is correlated with all other measures, standard ways of controlling for error inflation, such as Bonferroni, are
inappropriate and would only provide the absolute maximum overall error rate leading to overly conservative testing and inflated Type II error.

Many psychometric measures such as the Mental Health Index (12, 90) contain several scales. If one wishes to promote a specific scale or subscale as important in its own right and apart from the overall concept of quality of life, then Gabriel's simultaneous test procedures could be used (31). However, it may not be the primary intent of a quality-of-life study to promote or examine a particular scale such as sleep disturbance, anxiety, or vitality separate and apart from its contribution to overall quality of life. Therefore, one might choose to concentrate on univariate summary measures derived from factor analysis or prespecified weighting procedures and the corresponding multivariate profiles. Factor analysis identifies weights based upon statistical correlations and variances among the measures and reduces the total set of measures to linear functions that contain the essence of the information. Individual examination of multivariate subscales or computation of summary measures might be desired to examine the relative strength of the treatment effects across the scales, as well as between treatments. At the higher levels, how one combines the functional domains of psychological, social, general perceived, and physical health is still a matter of debate. At this level, the conceptualization regards domains within levels as being of relatively equal importance.

Global Hypotheses Tests for Multidimensional Responses

To define a "change" in quality of life, either positive or negative, one must contend simultaneously with issues of the multiplicity of responses, the multivariate nature of the variables, and the longitudinal nature of the data. Several analytic techniques have been proposed for analyzing clinical trials with multiple endpoints (65, 75, 96). In clinical trials of quality of life, multivariate test statistics such as Hotelling's \( T^2 \) are often used to determine treatment effects (76). However, this statistic is intended to detect any departure from the null hypothesis and therefore might lack the power to detect specific types of departure (e.g. therapeutically favorable, not favorable). For a set of \( p \) quality-of-life endpoints without prespecified priorities, the question of how can significance testing be used while preserving Type I error and allowing correlated endpoints has been addressed in different ways (74).

Pocock addressed several issues of hypothesis testing in quality-of-life clinical trials including the repeated use of significance tests comparing multiple outcomes (74). He proposed that a "directional" global test such as that proposed by O'Brien (65) would be most suitable for combining measures related to the same aspect of quality of life. He also cited the need to calibrate a change in quality of life to a "minimally clinically
important mean difference”. A series of articles recently reported various opinions on the types of analytical and statistical analyses that should accompany quality-of-life multiple endpoints research (27, 28, 33, 66, 77).

However, quality-of-life measures are not strictly “multiple endpoints” in the sense that they are supposed to measure a more global construct through subsets of interrelated constructs. While the multiple endpoints procedures may be suitable for combining variables for which no single syndrome or pathologic process has been defined (e.g. weight, reductions in blood pressure, increases in insulin sensitivity, changes in creatinine level, and lipid levels), quality-of-life endpoints that are linked through a careful conceptualization of dimensions, domains, and indicators deserve to be treated as interrelated constructs rather than unrelated entities. This has led investigators to rely on more sophisticated methods of longitudinal modeling, as reviewed briefly in the following sections.

**Longitudinal Models of Quality of Life**

The most relevant method for dealing with intervention and observational studies involve models that take into account the longitudinal nature of the data. Quality-of-life studies involve problems and sources of bias found in most longitudinal designs including selection bias, heterogeneity, directionality, and confounding. Repeated measures over time on the same individuals are not independent. In addition to the problem of withdrawals and persons lost to follow-up, the conceptual problem of choosing the appropriate model characterizing the changes over time is complex.

Longitudinal assessments of quality of life involve repeated sampling, especially in assessing cognitive functioning, depression, and anxiety where the cyclic trends are important. In addition, long-term surveillance involves prospective repeated evaluations of several domains over time. Ekstrom et al reviewed the statistical methods applied to psychiatric research studies that used parallel groups and multivariate repeated measures designs (25) and recommended increased use of multivariate analysis of variance.

Even when the autocorrelation structure of a times series is not of direct interest in longitudinal studies of quality of life, it has an indirect bearing on the estimation of the parameters that are subsequently calculated. Failure to account for the autocorrelation function in relevant cases results in potential bias when estimating test statistics for quality-of-life studies (14).

A thorough and comprehensive review of statistical models for longitudinal studies of health was compiled by Dwyer et al (23). Waternaux & Ware (95) reviewed linear models for analysis of longitudinal data and cited several methodologies including longitudinal random-effect models that take into account two corresponding sources of error, including within- and between-subject variance components. Appropriate estimation methods for
these types of models include the restricted maximum likelihood estimation (REML) methods (45). Laird (55) and Laird & Ware (56) used the EM algorithm proposed by Dempster, Laird & Rubin (20) to obtain REML estimates.

Other investigators use a more global longitudinal approach based upon the stochastic processes describing the various transitions between health states (2, 67). Those models that use continuous observations depicting health states at discrete points in time can be considered autoregressive processes if the current state depends only on the previous state. Discrete observations can be modeled by a Markov chain. Such models have been used to model disease processes such as HIV infection in other contexts (61).

Still other investigators rely upon an approach using structural equation models to model the complex interrelationships among biopsychosocial variables (51). These models allow estimation of stochastic models with continuous state space and the incorporation of latent variables and error terms. Latent variables are essentially "indirect measures" and thus the model might be fitting for the analysis of quality of life. The modeling of these indirect constructs by a set of indicator variables reduces the error from measuring the construct only through single variables. These methods also have the advantage of allowing for the modeling of explicit assumptions about the measurement error. Zwinderman modeled the measurement of change in quality of life using a latent logistic regression model that allowed for the inclusion of parameters for the time process, the effects of clinical treatments, and the interaction parameters (102).

**Effect of Missing Data, Early Withdrawal, Length Bias and other Confounders**

The problems associated with early withdrawal from a quality-of-life study are particularly troublesome because of the high degree of association between decreases in quality of life and missing data and early withdrawal from the study. This is particularly true in therapeutic trials comparing agents that cause side-effects. In this case, one cannot adopt the assumption of noninformative loss to follow-up. Patients might miss a visit because of some event unrelated to the treatment or disease. Here the missing data is assumed to be random and noninformative. However, if patients miss a visit because they are too sick (physically or emotionally) due to either side-effects of therapy or disease progression, the missingness is informative. Informative missingness occurs when the probability of response depends on the outcome or other covariates. When patients withdraw permanently from the study and no further quality-of-life follow-up information is available, the problem is even more serious and complex.
Choi & Stablein reviewed the mechanisms by which missing data occur (13) and identified three different mechanisms: (a) the treatment or the group to which the subject is assigned; (b) the outcome; and (c) both the treatment and the outcome. In a study comparing two antihypertensive agents, nifedipine GITS and atenolol, the change in quality of life from baseline to last-available visit showed a much larger deterioration for those who withdrew early from the trial as compared to those who completed the study protocol (83). When treatment differences were calculated based upon all cases (last observation carried forward (LOCF)) or 24-week completers, the conclusions were quite different. The LOCF analysis showed no treatment differences whereas the completer analysis showed a more favorable effect for nifedipine GITS, as shown in Figure 2. What is the correct conclusion concerning the treatment differences? One could insist upon maintaining the clinical trial model and the “intent to treat” principle using the LOCF method or other forms of missing value imputation. Or one could adopt the pharmacoepidemiologic or pharmacoeconomic model that would strive to estimate quality of life in those individuals who would practically continue on medication and whose collective experiences with loss of work, side-effects and the positive effects of therapy would contribute to future functioning in society. Heyting et al (46) reviewed statistical complexities that arise

![Figure 2](image-url)

*Figure 2* Factor score 20-week changes from baseline for patients receiving nifedipine GITS and atenolol. Statistical contrasts: nifedipine statistically different from atenolol, *p* < .01, **#p** < .05. Reprinted from Testa et al (83) with permission.
from outcome-related drop-outs in longitudinal clinical trials. The authors proposed methods derived from sample surveys with nonresponse and for observational studies. In a comparison of five different methods for dealing with early withdrawal, they concluded that methods from sample surveys and observational studies were the most advantageous. However, they cautioned that the ability to answer explanatory, rather than pragmatic, questions, arising from a comparative trial with drop-outs is still a complex issue.

Several studies have been designed with too short a follow-up period to make valid conclusions concerning the steady-state functional health of the patients under study. For example, a study by Steiner et al (79) with four groups and 360 male patients made conclusions after only four weeks of stable treatment with antihypertensive therapy. Another study by Dahlof & Dimenas drew conclusions after only two weeks about 77 hypertensive patients who were treated in two groups (18). Length bias is a particularly relevant issue in studies of this type, whether due to early withdrawal or study design, because changes in quality of life are not instantaneous and might emerge only after several months.

Another issue that affects the sensitivity of the measurement instruments deals with the ceiling/floor effects experienced by quality-of-life scales, which is dependent upon baseline quality of life. This is related to the issue of responsiveness. That is, patients cannot improve if they are already at the top of the scale, or deteriorate when they are already at the bottom of the scale. Responsiveness, and therefore sensitivity to treatment effects, is dependent upon the mean values of the scale itself. If patients fall outside of the range for which the scales have meaning where estimation of the true score is not possible, the scale will have both low responsiveness and sensitivity. If the effects of treatment are dependent upon the values of covariates that characterize those groups of individuals, then the ability to detect treatment effects may be seriously confounded. The study by Testa et al recommended controlling for baseline scores in the longitudinal analysis as a way of controlling for differences in responsivity within groups (85).

Other confounders are more obvious. Conclusions based upon single group open-label post-marketing trials claiming absolute "improvement" or "worsening" are unfounded because the studies lack both a randomized control design and the blinding required to rule out a placebo effect. One study, which assessed the effect of enalapril on hypertension and quality of life in a single-arm, open-label study of 4988 patients, made the claim that "Quality of life, determined by mean of the Nottingham Health Profile questionnaire, was favourably influenced in patients participating in this study" (19) even though it was neither unblinded nor contained a control...
group. Statements of absolute change in terms of worsening or improvement in uncontrolled studies should always be viewed cautiously.

**QOL SUMMARY INDICES FOR USE IN EFFECTIVENESS MODELS**

Several types of economic analyses can be used in the evaluation of new medical technologies, including pharmacologic agents (30, 99, 100). Effectiveness measures convey the overall benefit that the agent has on the treated population. When a drug is more effective and more costly than a comparative agent, cost-effectiveness analysis estimates the incremental gain in the therapeutic benefit per unit cost. Quality of life can affect the value of the effectiveness measure, such as survival or time to a clinically defining event, thereby readjusting incremental cost-effectiveness. In a comprehensive review, Weinstein (97) cites several examples that reinforce the need for incorporation of quality of life in cost-effectiveness analyses.

**Quality-Adjusted Measures of Effectiveness**

Several methods of analysis have been proposed to quality-adjust measures of efficacy for use in therapeutic decision making and economic evaluation. These techniques are commonly referred to as cost-utility analyses because of the utility weights placed on the outcomes of treatment. Pliskin (73) considered five health states in a study of hemodialysis versus kidney transplantation in the treatment of end-stage renal disease with corresponding utilities 1, .88, .7, .55, and 0, with utility = 1 describing a patient carrying a functioning transplant, working full time and returning to preuremic levels of activity, and utility = 0 describing a patient who is chronically ill and almost totally disabled (on dialysis or carrying a transplant).

Another model for evaluating coronary artery bypass surgery made utilities dependent on both the current health state as well as the baseline life-style of the patient (98). That is, individuals who were active had a greater reduction in life quality when experiencing pain with strenuous activity, than those patients who were sedentary. Such models are important because they account for the very real differences in health perceptions among patients with the same illness in common.

Both models have been applied to descriptive simulations using data on efficacy and quality of life from prior studies and evaluations. However, a recent model called *Quality-Adjusted Time Without Symptoms and Toxicity* (Q-TWIST), which relies on utility-weighted partitions of survival, has been used for estimation purposes in the evaluation of treatments for breast cancer (35, 36). The approach is similar to that of Pliskin (73) and divides survival time into three health states: time spent during periods of toxicity; time spent during the periods following progression of disease; and time spent
free from both toxicity and progression. Survival time is quality-adjusted according to the utilities assigned to the two health states of toxicity, $\mu_{\text{tox}}$, and progression, $\mu_{\text{prog}}$, assuming a utility of 1 for Time Without Symptoms and Toxicity (TWIST).

Recently, the Q-TWIST model was applied to data from a clinical trial (26) comparing the use of zidovudine (AZT) versus placebo in patients with AIDS-related complex (34). The study illustrated that the decision to treat can be affected by the relative values the patient assigns to $\mu_{\text{tox}}$ and $\mu_{\text{prog}}$. While the original analysis determined that zidovudine resulted in delay of the progression to an AIDS-defining clinical event and incurred “minimal” side effects (26), the Q-TWIST risk-benefit analysis concluded that the decision to treat should be dependent upon the value of utility (ranging from 1 = no toxicity or disease symptoms to 0 = worst health state) that the patient or physician gives to the quality of life experienced during toxicity, $\mu_{\text{tox}}$, when compared to the utility given to the time after an AIDS-defining disease progression, $\mu_{\text{prog}}$. Depending upon patient preference, zidovudine could be “effectively” better than placebo or “effectively” worse than placebo. The estimation of suitable indices or global measures appropriate for use in pharmacoeconomic models is necessary if such models are to be used in health-policy planning.

**Health State Transition and Utility Models**

Most quality-of-life trials are designed to measure quality of life at specific points in time rather than allowing time to vary randomly. In the fixed-time model, given a large enough patient population, it is possible to estimate the rate of transitions between health states by Markov models typically employed in compartmental analysis using the number of patients in each health state at fixed timepoints. In these models, $n(tj)$, the number of individuals in health state $j$ at time $t$ is the random variable, rather than the amount of time as required for the Q-TWIST model.

Rate parameters $L(i,j)$, representing the rate at which individuals transit from health state $j$ to health state $i$, can be estimated along with treatment effects using population-parameter random effects estimation techniques. The estimated time spent in each compartment or health state can be obtained from a function of the rate parameters. Since it is possible to estimate the amount of time spent in each health state from the estimated transition rates and since it is possible to obtain utility weights using a rating scale, a global summary parameter GQOL is estimable from the linear combination of utilities and times given by

$$GQOL = \sum_{j=1}^{p} \mu_j \tau_j$$
where \( \mu \) is the utility associated with health state \( j (j = 1, 2, 3, \ldots p) \) and \( \tau \) is the amount of time spent in health state \( j \) as estimated from the rate parameters.

**SUMMARY**

Methodologies involving the use of quality-of-life patient outcomes in observational and interventional studies of health are drawn from a large and diverse field of research methods. The multidimensional way in which quality of life is conceptualized will affect the way it is measured and the complexity of the measurement. At the earliest stages of research, one must rely on methods common to the fields of tests and measurement, survey research, psychometrics and sociometrics to measure constructs that are not directly observable. Indices measuring performance can either focus on the scale’s ability to perform in noninterventional, cross-sectional studies or interventional, longitudinal studies. Indices of stability, internal consistency, responsiveness with respect to true changes in quality of life, and sensitivity to treatment effects can be used to assess the scale’s adequacy as a dependent variable of interest. Respondent variability can occur due to factors such as different reporters (patient, spouse, physician), the manner and form of administration (long form vs short form; self-administration vs interview) and the assessment environment (clinic, home).

Finally, since quality-of-life research often involves inferential statistics and hypothesis testing, the statistical and epidemiologic principles of good study design should be followed. In addition, one should account for the reliability, responsiveness, and the sensitivity of the scale when designing the scientific hypotheses, and should specifically address the meaning of quality-of-life effect sizes by interventional-based validation. Design considerations must address the statistical issues of power, the determination of effect sizes through validation by external criteria, longitudinal data, effects of withdrawal and early termination, ceiling and floor effects, and heterogeneity of responsiveness and sensitivity among individuals.

The problem of estimating quality-of-life summary parameters for use in pharmacoeconomic models is receiving increasing attention in this era of health-care reform and fiscal restraint. While medical decision theory has used cost-effectiveness models and quality-adjusted life years since the early 1970s, estimation of population parameters to differentiate among different medical interventions is relatively new. The assessment of the patient outcomes associated with medical interventions in terms of the risks, benefits and costs will clearly be a major focus of health-care reform. Development of new methodologies in quality-of-life research should build upon the strong foundation already established in the areas of clinical research, epidemiology,
biostatistics, economics and behavioral science. The purpose of quality-of-life research is ultimately to change the way we deliver and intervene in health care. Improvement in the health-related quality of life is a major objective in public health research. Research findings in this area have the potential to alleviate human suffering, minimize discomfort and morbidity, and positively affect health and well-being.

Acknowledgments

We thank Dr. Donald C. Simonson, Brigham and Women's Hospital and the Joslin Diabetes Center, and Harvard Medical School, Boston, who served as the physician consultant and medical reviewer for this report.

Support for this research was provided by a grant from the Agency for Health Care Policy and Research (AHCPR), No. R01 HS07767-01, and NIAID Contract No. NO-A1-95030.

Literature Cited


41. Guyatt GH, Veldhuizen Van Zanten


97. Weinstein MC. 1981. Economic as-


