Assessment in Behavioral Medicine

Edited by

Ad Vingerhoets
Department of Psychology, Tilburg University, The Netherlands
BIOBEHAVIOURAL PERSPECTIVES ON HEALTH AND DISEASE PREVENTION

From the perspective of behavioural science, the series examines current research, including clinical and policy implications, on health, illness prevention, and biomedical issues. The series is international in scope, and aims to address the culturally specific, as well as the universally applicable.

Series Editor: Lydia R. Temoshok, PhD, Institute of Human Virology, Division of Clinical Research, University of Maryland Biotechnology Center, 725 West Lombard Street, Room N548, Baltimore, Maryland 21201, USA

Volume 1
Stress and Health: Research and Clinical Applications
*edited by Dianna T. Kenny, John G. Carlson, F. Joseph McGuigan and John L. Sheppard*

Volume 2
Psychosocial and Biomedical Interactions in HIV Infection
*edited by Kenneth H. Nott and Kav Vedhara*

Volume 3
Adult Crying: A Biopsychological Approach
*edited by Ad J. J. M. Vingerhoets and Randolph R. Cornelius*

Volume 4
Anxiety Disorders in Children and Adolescents; Epidemiology, Risk Factors and Treatment
*edited by Cecilia A. Essau and Franz Petermann*

Volume 5
Assessment in Behavioral Medicine
*edited by Ad Vingerhoets*
Health-related quality of life (HRQOL) is viewed by patients, clinicians, and society as an important outcome of medical technology and disease control. HRQOL assessment has become necessary for the evaluation of pharmaceutical treatments, medical interventions, and in the tracking of population health. Behavioral medicine clinicians and researchers will benefit from recent advances in the assessment of treatment effectiveness. This chapter is not intended as a review of all HRQOL instruments; instead it provides a review of concepts less likely to be familiar to clinicians—ideas that are quickly determining how we measure health and the impact of behavioral medicine interventions. Further, we consider the role of outcomes assessment in the larger resource allocation process. We conclude by discussing what should be considered when choosing and applying a measurement strategy for assessing quality of life in a behavioral medicine setting.

The term "quality of life" is used inconsistently. Quality of life (QOL) is a broad term often encompassing aspects of life not always directly relevant to a behavioral medicine clinician (i.e., standard of living, job satisfaction). The term "health-related quality of life" (HRQOL) clarifies the health focus of interest and narrows quality of life to aspects relevant to health. Although the term "functional status" is often used interchangeably with the term "health-related quality of life", there is a clear distinction between the two. Functional status is the objective degree of disability caused by an illness, whereas HRQOL also includes subjective evaluation of that disability, referred to as "handicap" by the World Health Organization (WHO, 1980). It may often include symptoms that may not affect functioning (Kaplan & Anderson, 1996). Though the term "health status" is used by many authors, we prefer HRQOL, because it more accurately conveys gradations of life quality and "health status" does not adequately convey these gradations.
It should also be noted that the popularity of the term “quality of life” for both patients and regulatory agencies has led to numerous established measures being relabeled as QOL measures. We prefer to think of HRQOL measures as assessing multiple domains of health, not as an aggregation of separate measures developed independently to assess various aspects of health (e.g., combining a measure of depression and a symptom checklist and referring to it as assessment of QOL). Thus, the discussion below includes only those measures that have been specifically designed to assess health-related quality of life.

Current importance of health-related quality of life

Until recently, common sense and clinical intuition were used to determine treatment effectiveness. Documenting which treatments are efficacious and which are not has been relatively unstandardized. Assessing side effects of medication and examining areas of health that are not the primary focus of treatment seemed unwarranted and burdensome. However, accountability is now a priority in health care. Measuring outcomes in a way that allows broader comparison of interventions has become the standard of care. The United States serves as an important example of how the focus in health outcomes has transformed the health care system. American doctors perform more procedures than their colleagues in any other country in the world, yet there is little evidence that Americans enjoy better health than their peers in other industrialized countries. Brook and Lohr (1987) have suggested that between 30–50% of all medical procedures have little or no documented benefit. To wisely invest health care dollars, it has become clear that we need to measure the product of health care.

The topic of HRQOL has become increasingly popular in the clinical research literature over the past ten years, not only because of extended survival from once terminal diseases, but because traditional measures of mortality and other biological outcomes do not adequately measure effectiveness of interventions designed to improve life quality. There is also a growing appreciation for the need to use such measures to assess which services produce the greatest amount of health for the money spent. The drive to use newer, more expensive, technology will result in soaring costs. Since resources are limited, it is necessary to determine which interventions produce the most benefit relative to their costs (Kaplan, 1993).

A goal in assessing outcomes is to have the ability to evaluate the value of care (a measure of cost and quality of care), and to assist
patients, payers, and providers in selecting treatments based on value. This leads to the question of how value will be defined. Professional societies and medical disciplines, including behavioral medicine, will be increasingly asked to document the value of their services. This is especially important to behavioral medicine because interventions can have such a broad impact: smoking cessation, diet, exercise, and pain management are but a few of the areas to target that affect a broad range of conditions and areas of disability. What remains to be seen is what processes (health care) lead to better outcomes. Understanding how best to provide such documentation is the focus of the remainder of this chapter.

Methods of assessing health-related quality of life

One way to understand the numerous instruments available to assess HRQOL is by crossing two dimensions: the intended use of the information and the breadth of health domains covered in the measure. More specifically, instruments range in regard to their usefulness in individual treatment decisions, and whether the content of the assessment is limited to disease specific expectations or a broader understanding of overall health status. The focus here will be broader applicability and thus disease specific measures will not be reviewed (see Spilker, 1996).

The diversity of problems encountered in most behavioral medicine settings make generic measures a more flexible component of a standardized assessment battery. While HRQOL measures cover a broad range of functioning, disability, and distress, they are often cited as being insensitive to small changes in clinical status (as compared to disease specific measures). Yet, despite the high face validity of disease specific measures, the literature does not uniformly support their greater sensitivity. It is generally regarded as good practice to include both a disease-specific as well as a generic measure in both research and clinical work.

PSYCHOMETRIC-BASED ASSESSMENT
Psychometric approaches to measuring HRQOL are perhaps the most familiar to those trained in the health sciences. These measures often require the respondent to indicate the frequency and/or intensity of symptoms or behaviors. Responses to individual questions are aggregated to create separate homogeneous scales (e.g., physical function,
social, and/or emotional function). Such scales have been successfully used to assess the outcomes of medical treatments and to compare patient outcomes under different systems of care (Deyo & Patrick, 1989; Spilker, 1996).

Strengths of psychometric measures include the fact that assessment of multiple dimensions of health appeal to clinicians and are familiar to those trained in this common assessment strategy. Limitations include the inability to integrate the information in economic analyses of treatment outcome, the subjectivity of reporting perceived ability, and that mortality cannot be incorporated into the data analysis (longitudinal analyses will often exclude any deaths in a sample of cancer patients and thus may bias the results). Clinical interpretation can be difficult, as comparison of different treatments may favor one group on some of the scales, while the comparison group scores more favorably on other scales.

Perhaps the most common outcome assessment procedure in contemporary health services research is the SF-36 (Ware & Sherbourne, 1992). A product of the RAND-sponsored Medical Outcomes Study (MOS), the Short Form 36 (SF-36) is a 36-item general health status assessment questionnaire that has substantial data on its reliability and validity on a wide variety of populations (Ware et al., 1994). For a complete description of this instrument and its applications, see Chapter 15 in this volume.

Stiggelbout et al. (1996) have developed an instrument that directly asks patients to rate their general attitude toward quality of life over quantity of life. The instrument is conceptually similar to the time trade-off method (see below) but is a self-administered paper and pencil questionnaire, not an interview format used in traditional time trade-off tasks. The questionnaire was constructed on the basis of semi-structured interviews with cancer patients, resulting in ten statements concerning tradeoffs between quality of life and length of life, quality of life and chance of survival, and the attitudes of subjects toward discontinuation of treatment. Each statement is rated on a 7-point Likert scale ranging from completely agree to completely disagree. A high score on the “Q” scale implies a limit to the acceptability of reductions in quality of life; a high score on the “L” scale suggests high striving for lengthening life, even if that is associated with reduced quality of life. It appeared that older patients more often stressed quality of life and were less concerned with prolonged survival than younger patients. The authors address issues for future refinement of this scale, but specifically designed the questionnaire to evaluate atti-
tudes toward trade-off of quality versus quantity of life. They argue that just as patients are less willing than healthy subjects to trade-off length of life for quality of life, patients might be less willing to trade-off length of life for quality of life in actual situations than in hypothetical situations.

UTILITY-BASED MEASURES
Behavioral medicine professionals are faced with many of the same challenges as other health care providers. Documenting the effectiveness of their interventions relative to other interventions offered by other providers will become increasingly important in the health care marketplace. In fact, a panel convened in 1993 by the US Department of Health and Human Services suggested that standardized outcomes analyses be conducted to evaluate the cost/effectiveness of medical care (Gold et al., 1996). One way to directly compare relative treatment effectiveness is to examine the impact of interventions on the utility gained. This requires certain characteristics of an assessment instrument. Utility-based measures address this requirement.

There are many controversies and concepts relevant to utility-based measurement, and we cannot review them all here. The interested reader is referred elsewhere for a more extensive discussion of the issues (Gold et al., 1996). Utility theory and measurement were developed, in part, as a normative model for individual decision making. Utilities are numbers that represent the strength of an individual’s preference for various health outcomes under conditions of uncertainty. Preferences are the values people assign to different health outcomes when uncertainty is not a condition of measurement. These numbers reflect a person’s level of subjective satisfaction, or desirability with different health states. The utility approach uses one or more scaling methods to assign numerical values on a scale of 0 (anchored as death) to 1.0 (anchored as optimum health). The resultant score thus allows morbidity and mortality to be combined into a single weighted measure, often referred to as (quality adjusted life years), or QALYs.

Quality Adjusted Life Years
The benefits of health care can be expressed in terms of the years of life they produce, adjusted for reduced quality of life. The most popular term for this concept is “quality adjusted life years” (QALYs) (Weinstein & Stason, 1977), although other terms such as “well years”
have also been used. For example, if a cigarette smoker died of heart
disease at age 50 and we would have expected him to live to age 75, it
might be concluded that the disease cost him 25 life years. If 100 ciga-
rette smokers died at age 50 (and also had life expectancies of 75
years), we might conclude that 2,500 (100 men × 25 years) life years
had been lost. However, we should not assume that these 25 years lost
were disease-free or disability-free. That is, death is not the only
outcome of concern; many cigarette smokers would be expected to
suffer myocardial infarction or develop pulmonary diseases that would
leave them somewhat disabled to varying degrees. Although they are
still alive, the quality of their lives has diminished.

The use of a concept such as QALYs permits all degrees of disability
to be compared to one another. The QALY combines data on the total
life years gained from an intervention with data on the utility (or value)
of health states for those years, to give a single measure of achieve-
ment or output. By calculating the cost per QALY gained for different
clinical procedures or even different social problems, available
resources can be directed toward interventions that maximize health
gain (Chisholm et al., 1997). That is, a disease that reduces the quality
of life by one-half will take away .5 QALYs over the course of 1 year. If
it affects two people, it will take away 1.0 QALY (equal to 2 × .5) over
a 1-year period. A medical treatment that improves the quality of life
by .2 for each of five individuals will result in a production of 1 QALY
if the benefit is maintained over a 1-year period. Using this system, it is
possible to express the benefits of various programs by showing how
many QALYs they produce (Kaplan & Anderson, 1996). Using this
common metric of QALYs also allows one to introduce cost into the
direct comparison of programs. Several alternative terms have been
developed which are fundamentally similar to QALYs yet focus on dif-
f erent elements of utility. Healthy year equivalent (HYE; Mehrez &
Gafni, 1989) and disability adjusted life years, or DALYs, are just two
examples. Using this approach allows for comparison of behavioral
medicine programs to other medical procedures on very different
populations. Thus, this approach provides a framework within which
to make policy decisions that require selection between competing
alternatives.

As mentioned above, a number of methods are used to calculate util-
ities and preferences (Revicki & Kaplan, 1992). Standard gamble or
lottery procedures are used to calculate utilities, while preferences can
be generated by using visual analog rating scales or other scaling
methods. Either holistic or decomposed approaches can be used to
elicit health utilities. Each approach has its benefits and limitations, and each researcher or clinician must decide which approach best fits their situation. Again, given the focus on application of HRQOL to behavioral medicine, interested readers are referred elsewhere for more coverage on this issue (e.g., Gafni, 1994; Kaplan et al., 1993; Testa & Nackley, 1994; Torrance, 1987). However, some understanding of these two approaches should inform decisions about the choice of instruments.

HOLISTIC APPROACH
The holistic method of calculating utilities involves having an individual assign utilities to a number of hypothetical health state scenarios. These scenarios describe important aspects of a health condition in terms of physical, psychological, and social functioning. These scenarios are presented to individuals and the relative preferences are elicited for each health state.

Two common holistic methods explicitly have patients consider decisions under uncertainty: standard gamble and time trade-off. The standard gamble offers a choice between two alternatives: Choice A—living in a suboptimal state of health with certainty, or Choice B—taking a gamble on a new treatment for which the outcome is uncertain. The subject is told that the new treatment will lead to perfect health with a given probability (p) or immediate death with a given probability (p-1). The subject can choose between remaining in a state that is intermediate between wellness and death, or take the gamble and try the new treatment. The probability (p) is varied until the subject is indifferent between the choices A and B. For example, a person is told to imagine they are blind and that a treatment which would enable the patient to regain their sight results in a .001% chance of death. If the patient chooses the treatment then the odds of death are increased to .1%, for example, and the patient is again asked to choose their current state of blindness or the treatment with risk of death involved. The risk continues to be increased until the patient no longer is willing to take the risk. The less risk a patient is willing to take to alleviate a health state the higher the utility of that health state.

A variety of problems with the gamble method have become apparent. Some believe that the standard gamble has face validity because it approximates choices made by medical patients (Mulley, 1989), yet treatment of most chronic diseases does not result in complete cure or death. Also, the cognitive demands of the task are high, as well as the time required to ascertain one individual’s utility/preference.
Another common holistic method is referred to as time trade-off. Here, the subject is offered the choice of living for a defined amount of time in perfect health or a variable amount of time in an alternative state that is less desirable (e.g., moderate disability). Presumably, all subjects would choose a year of wellness versus a year with some health problem. However, by reducing the time of wellness and leaving the time in the sub-optimal health state fixed, an indifference point can be determined. For example, a subject may rate being in a wheelchair for two years as equivalent to perfect wellness for a year. The time trade-off is theoretically appealing because it is conceptually equivalent to a QALY. However, there is a concern whether the tasks can be clearly understood by the average subject (Kaplan, 1995).

DECOMPOSED APPROACH
In the decomposed approach, patients are asked a series of questions about their functioning in specific health domains. Based on their responses, individuals are assigned a utility value. The weights and utilities are developed from previous ratings by samples from the general population, or some other reference group. For example, the Health Utility Index (HUI; Torrance & Feeny, 1989) and the Quality of Well-being Scale (QWB; Kaplan & Bush, 1982) use a decomposed approach to generate preferences. In fact, the HUI and QWB represent hybrids of psychometric and utility-based measures.

The utility/preference approach has several advantages compared to the psychometric approach. First, it incorporates time and risk preferences for different health state outcomes into the measurement process, and scores are easily incorporated into economic analyses. Preference assessment has been successfully incorporated into numerous clinical trials. However, there is controversy over the definition of utilities/preferences and the methods used to derive these values. Preferences for some health states vary widely between individuals as well as a result of how the health states are described.

However, there are unresolved issues with deriving preferences. Namely, the duration of the disease state may influence the preference itself. For example, many patients would value an illness condition as more desirable if the assumption was made that the “stay” in that health state was short (“I can tolerate anything for a day”), and thus relative (and absolute) preferences may change as a result. Another issue is whose preferences are most useful. It appears that the patient’s perspective is most important, though often studies will rely on physician’s rating/preference primarily out of convenience of design. In
addition, measured preferences have been reported to vary with the method of elicitation and respondent populations surveyed. In eliciting preferences for hypothetical health states from the general population, the subjective rating of a respondent's own health state should be considered in determining representative population groups.

The Quality of Well-Being Scale (QWB)

While the SF-36 serves as an example of a psychometric instrument, the Quality of Well-Being Scale (QWB) is a preference-based instrument used with several different populations (Kaplan & Anderson, 1996). While the Health Utilities Index (HUI) is also a psychometrically sound instrument, our experience has been with the QWB and thus is described in more detail here. Interested readers are referred to Feeney and Torrance (1989) for information on the HUI. The QWB assesses a patient's objective level of functioning in three domains: mobility, physical activity, and social activity. A distinction is made between "functional ability" and "functional performance" (Anderson et al., 1989), namely a patient is asked to report activity performed rather than the patient's perception of what could possibly have been performed. The QWB concentrates on functional performance (or what the individual actually did) on the past six completed days.

In addition to these three domains, the QWB assesses the presence of a wide array of symptoms. On any particular day, nearly 80% of the general population is optimally functional, yet over an interval of six days, only 12% experience no symptoms (Kaplan et al., 1976). Even if these symptoms do not affect a patient's functioning, they do lower quality of life. For example, the QWB score is heavily driven by the wide array of symptoms assessed, as compared to functional performance (Kaplan et al., 1976). Our experience has shown that the QWB instrument is, in its operation, sensitive to the health-related issues that are most important to people, and it is thus capable of capturing even small variations in health status. It is in fact the importance of symptoms on the QWB that may make it more useful to the behavioral medicine clinician than most other generic measures.

One of the criticisms of the QWB is that it is more expensive and difficult to administer than competing measures such as the SF-36. The QWB is relatively long and complex because it has some branching and probe questions and requires a trained interviewer. The original QWB used a complex interviewer-administered questionnaire because
self-administered questionnaires produced biases resulting in overestimates of health status (Anderson et al., 1986, 1988). However, refinements in questionnaire design may allow us to get around these problems.

The Quality of Well-Being Scale, Self-Administered (QWB-SA)

Given the above mentioned criticisms, a self-administered QWB—referred to as the Quality of Well-Being Scale, Self-Administered (QWB-SA)—was developed that addresses some of these issues (Kaplan et al., 1996, 1997).

The QWB has several strengths has we wished to maintain. First, the QWB includes assessment of symptoms in addition to various areas of functioning. To help make the instrument more useful in the clinical setting the list of symptoms was expanded. The revised list of symptoms and problems includes several mental health items and all items are arranged in a manner consistent with a medical “review of systems”. Second, the QWB assesses functional status (versus perceived ability) in three areas: mobility, physical activity, and social functioning. This perspective was maintained. Third, the QWB asks a patient to report on symptoms and activity over a 6-day period. To reduce recall bias, the QWB-SA assesses only the three days prior to completion of the questionnaire. Finally, the scoring of the instrument utilizes population-derived preference weights. Given the addition of several items to the QWB-SA, new preference weights were derived from ratings by a new sample of subjects. A series of studies have begun to establish the psychometric properties of this new measure. It is our expectation that the information derived from the QWB-SA should prove useful to clinicians as well as to health care managers and policy makers (Ganiats et al., 1997; Sieber et al., 1997; Sieber et al., 2000).

Quality of Life Time Without Symptoms of Toxicity or Treatment (Q-TWiST)

A recent development has offered another way to calculate QALYs (Gelber et al., 1996). The Q-TWiST methodology was originally developed for cancer research and to describe Quality of Life Time Without
Symptoms and Toxicity of treatment. Instead of survival analysis that scores patients 1.0 for being alive and 0.0 for death, TWiST codes time with symptoms or toxicity as 0.0. The Q-TWiST methodology is an extension of the TWiST method but adds quality of life to the evaluation. Thus, the term Q-TWiST is used for quality-time without symptoms and toxicity. In many ways, the Q-TWiST methodology is identical to quality-adjusted survival analysis. There are three steps to the Q-TWiST analysis. First, health states are defined. These health states typically represent the expected outcomes and side effects of treatment. Each of these states is assigned a utility score. The exact method for assigning utilities may differ from study to study. Sometimes the utilities are simply assigned by the investigators while in other cases the utilities are provided by patients.

The second step involves partitioning of the overall survival time. For example, one component of survival might be TWiST or the time without any symptoms or adverse effects. The second stage might be the time with severe symptomatic adverse effects. A third component might be the amount of time with reductions in wellness due to the progression of the disease. All of these occur prior to death. Usually, survival analysis is used to estimate the duration of these states. For example, in a study of patients with breast cancer the median survival time might be seven years. The analysis would estimate the portion of time spent in each of the defined states.

The third stage involves comparison of treatments using the Q-TWiST methodology. For each treatment group, duration of stay in state is multiplied by the utility and the Q-TWiST is calculated as the sum of TWiST plus the products of remaining utilities and health states. For example, suppose that patients have a median survival of seven years. Three of the years are spent without symptoms (TWiST). Two years are spent with adverse reaction to medication. The utility for this state is .7, so the adjusted years in this state equal 1.4 (obtained by multiplying 2 years x .7 quality weight). Then, the remaining two years are spent in a state of disease progression with a utility of .5. This results in one adjusted year (.5x2=1). So, the total Q-TWiST is 5.4 (obtained as (3 years x 1.0) + (2 years x .7) + (2 years x .5)). Q-TWiST analysis typically compares treatments using these adjusted survival times.

The Q-TwiST was developed for studies in cancer and AIDS. Recently, Schwartz and colleagues have adapted the method for other chronic diseases (Schwartz et al., 1995a, 1995b, 1997). The Schwartz adaptation of Q-TwiST allows the dimensions to be continuous rather
than binary. The preference weighting system uses patient ratings, but can also integrate social costs and the provider perspective. Several studies have shown the value of the modified Q-TWiST for evaluations of outcomes in neurologic diseases such as multiple sclerosis (Schwartz et al., 1997) and epilepsy (Schwartz et al., 1995b).

Assessment in the clinical setting

CHOOSING THE RIGHT INSTRUMENT
HRQOL measures have the potential to become the new standard in medical practice. Without these tests, patient functioning and well-being are unlikely to be discussed during a typical medical visit. A majority of patients feel it is appropriate and desirable to discuss psychological problems with their physicians, but few patients initiate these discussions even when they are experiencing problems (Good-Delvecchio et al., 1987). As a result, doctors are not well-informed about their patient's functional status, or HRQOL. Well-being, especially as it relates to psychological distress, often goes unrecognized and untreated. Collecting this information may be paramount to treatment success in behavioral medicine settings.

Medical personnel are sometimes concerned about the validity and importance of self-rated health collected through self-administered surveys, often preferring physiological and biomedical outcomes. There is a tradition of using highly technological apparatus to obtain extremely precise estimates of aspects of pathology and impairment. However, the majority of clinical measurements used for diagnosis and treatment-response monitoring are low technology and require use of questions (i.e., clinical history) very similar in style to those used in HRQOL scales. In addition, providers are often uncertain about the responsiveness of these questionnaires to detect small, clinically relevant changes (Deyo & Patrick, 1989; Revicki, 1992). On the other hand, those concerned with health care resource allocation and the assessment of the cost-effectiveness of interventions often criticize health status questionnaires for not incorporating mortality, duration of survival, or patient preferences into health outcomes (Feeny & Torrance, 1989; Kaplan et al., 1989). The task is to address these criticisms.

While clinical medicine attempts to do the best for an individual person, regardless of cost, the public concern is with reducing the burden of disease suffered by populations and its ethical standpoint is
one of utilitarianism. However, since populations are made up of individuals and the burden of an illness in a population is the sum total of disease experienced by individuals, there should be some common ground on how to satisfy both clinical and public health medicine goals. Thus, the goal for providers of behavioral medicine should be to use a health status measure that is responsive to clinically meaningful differences in clinical symptoms and functional status, while providing information useful for determining where such changes are cost-effective for population-based health care.

One step toward this goal is the regular use of a HRQOL measure in everyday behavioral medicine practice. Such assessments would help ensure that important dimensions of health are consistently considered, would help track changes in health over time, and thus provide potentially useful information in treatment decisions. Such an assessment should provide a view of the patient's complete status, in order to detect unforeseen effects of treatment. Disease-specific measures do not have this breadth. As an instrument becomes more specific to a disease or particular function, it may no longer meet the goals of population-based assessment. Disease-specific measures aid in identifying patient behaviors that exacerbate a chronic illness and may provide helpful information, but their specificity precludes their sole use as an HRQOL instrument.

To make this approach practical, assessment should vary in length according to the application. Comparability between treatments and populations would be enhanced if short-forms (of more disease-specific symptoms) were imbedded within longer forms assessing more global function. This would allow a portability in both contributing to a normative database (e.g., generic questions), and tailoring the assessment to the particular disease being treated.

From both clinical and public health perspectives, outcome measures are needed that are responsive to changes in services or treatments. The lack of direct congruence between measures of disease activity and subjective feelings of distress is a reflection of the complexity of factors that determine well-being and HRQOL. As clinical practice is concerned with a holistic approach to patients, HRQOL may provide an appropriate way of tapping into patients' experiences. Standardized questionnaires may prove very helpful for those patients who find it hard to verbalize their feelings and for those health professionals who find it difficult to explore the patient's wider experiences.

Deverill et al. (1998) propose three criteria for an acceptable QALY-type measure. First, it must be practical in that it can be completed in a
reasonably short time, facilitating a high response rate to the questions. Second, it must have adequate test-retest and inter-rater reliability. Third, it must have construct and empirical validity (gives results as expected with other measures of the construct). These authors state that QALY measures are rarely used correctly in economic analyses.

Administration and issues determining frequency of assessment

If administration of some standardized HRQOL instrument is accepted into behavioral medicine practice, how frequently should these assessments occur? There are no clear guidelines on how frequently health outcome measures should be given. Some measures ask respondents about an extended time period (e.g., one month), though this increases the likelihood of recall bias and inaccurate reporting. The best evidence suggests that people do relatively well at recalling health events for an interval of approximately one week or less (Kaplan et al., 1978). Thus, shorter intervals will provide a better approximation of current health status, though some events (e.g., symptoms, function) may not be captured if the interval is too short (e.g., 24 hours).

Once a measure and assessment window is selected, the question of how frequently a measure should be used depends on the problem under study. For chronic health conditions, assessment yearly may be appropriate. On the other hand, acute or episodic illnesses might be expected to change considerably over a brief time interval, and thus weekly assessment may be more appropriate. Generally, a matching between the sensitivity of the measure, the burden in completing it (i.e., length of questionnaire, requirement of face-to-face interview in clinic), and the variability of the disease or patient's health status should occur.

WHAT DOMAINS OF HEALTH TO ASSESS?

There is great value in using disease-specific measures of symptoms and problems unique to specific diseases in parallel with generic measures. One must be careful when selecting a disease-specific measure. A comparison of outcomes between two treatments using a disease-specific measure may be biased if the measurement includes a list of side effects likely to be experienced from one treatment but not the other; conversely, using only a generic measure without a broad assessment of symptoms may be insensitive to unanticipated side effects of either
Assessing the Utility of Behavioral Medicine Interventions

While it is important to assess function, most people report little or no dysfunction, yet report reduced HRQOL due to the presence of one or more symptoms. Thus, a generic measure that includes a broad spectrum of symptoms may be ideal.

Most health status measures (e.g., SF-36, QWB, HUI) begin with the World Health Organization definition of health as a complete state of physical, mental, and social well being, and not merely the absence of disease (WHO, 1948). This definition has led many investigators to assume that a general measure must include separate scales for physical, mental, and social health. In fact, some reviewers of the literature have gone as far as reviewing the quality of measures by recording whether there was a separate scale for each of these dimensions. However, this reliance on the labeling of subscales will likely misrepresent the actual content covered on many instruments.

For example, previously the QWB has been criticized because there is no subscale named "sensory function". However, the QWB includes symptoms for loss of vision, loss of hearing, impairment of vision (including wearing glasses or contact lenses), problems with taste and smell, and so on. Another example concerns mental health. Many authors note that the QWB excludes mental health content. Despite widespread interest in the model among practitioners in many different specialties, the QALY concept has received very little attention in the mental health fields until now. We believe that this reflects the widespread belief that mental and physical health outcomes are conceptually distinct. Ware and Sherbourne (1992) emphasized that mental and physical health are different constructs and that attempts to measure them using a common measurement strategy is like comparing apples to oranges. However, a measure without a mental health component does not necessarily neglect mental health. Mental health symptoms may be included and the impact of mental health, cognitive functioning, or mental retardation may be represented in questions about role functioning.

Consider the case of a person with depression. Depression may be a symptom reported by a patient just as a cough is reported by other patients. Depression without disruption of role function would cause a minor variation of wellness. If the depression caused the person to stay at home, his/her score would be lower, with severe depression leading to hospitalization resulting in a lower score still. Certainly, studies have shown the validity of the QWB assessing changes in mental health (see Patterson et al., 1994, 1997).
CEILING AND FLOOR EFFECTS AND OTHER RELEVANT PSYCHOMETRIC ISSUES

Validity to the clinician is a straightforward concept: does the scale measure what it is intended to measure and compare well to a “gold standard” or criterion. However, criterion and construct validity pose certain dilemmas when assessing HRQOL; these important issues have been outlined earlier (Kaplan et al., 1976). Additional relevant performance statistics are the accuracy of the instrument: its sensitivity and specificity. While criticism is often directed toward health questionnaires as being based on a “soft” science, the primary method of data collection for physicians—medical history taking—is plagued by these same problems. The issue for behavioral medicine clinicians should be whether an instrument is sensitive enough to detect meaningful changes expected from an intervention. Using a measure which can detect these differences and be used to compare a population or intervention to other populations or interventions is ideal.

Other issues to consider in the selection of an assessment tool: Will the data collected be comparable to data collected on other populations? Is the data able to be used in decisions regarding resource allocation? Is there coverage of all domains of health that are relevant to this specific population? If the condition being assessed is expected to be stable, then is the measure reliable over time? Ebrahim (1995) suggests that the assessment of reliability and validity of health status measures is carried out inappropriately. Namely, that repeatability of a measure across a population does not adequately address the stability of the measure to assess changes over time for an individual. Has concurrent, construct and convergent validity been demonstrated with the particular disease/population being studied? Is the instrument responsive to observable increments in change? That is, is the instrument able to show an expected dose response curve? Is the instrument sensitive to changes within an individual over time? Are ceiling and floor effects a concern with this instrument or with this particular patient population?

Ceiling and floor effects have been studied in relation to several different measures. Some evidence suggests that perfect scores on the SF-36 and the SIP are common (Anderson et al., 1994). In contrast, perfect scores on the QWB are rare (Ganiats et al., 1997). Similarly, floor effects on a utility-based measure indicate mortality, not a psychometric problem as it would be on a traditional psychometric measure.
INTERPRETATION AND USE OF DATA IN CLINICAL MANAGEMENT DECISIONS

As Remie and Garssen (1997) point out: should a cancer patient's high score on a depression questionnaire be considered abnormal or a healthy ability to acknowledge and disclose one's feelings and reactions to disease?

How does one interpret changes on an HRQOL scale? First, the clinician should not equate reported mean treatment/group differences with changes in an individual patient's score. Even after adjusting for possible mediating factors of age and gender, additional information is needed. Second, we believe the clinician must place a change in scores within the context of patient preferences (whether they be population-based or individually-derived). For example, does the patient more highly value the increased ability to ambulate or attach a greater value on symptom relief? A scale aimed to determine QALYs should have this ability.

Meyerowitz (1993) asks whether data drawn from large-scale studies play a meaningful role in clinical practice where the central concern is with a specific, unique individual. Unfortunately, studies have documented that physicians' perceptions of a patient's needs and concerns often differ from the patient's own report in terms of psychological distress and desire for information. In the absence of sound data, there is a high risk of making inaccurate judgments about what the patient wants. Therefore, an assessment tool that incorporates patients' preferences, either at a population or an individual level, seems most appropriate.

Summary and conclusions

Numerous barriers must be overcome before HRQOL measures can be incorporated into clinical practice. The length and cognitive complexity of questionnaires are key issues. The ideal instrument should be quick to administer, sensitive to detect small changes in health, be interpretable by clinicians, and useful in the development of cost-effective population-based interventions.

The purpose of this chapter has been to summarize major methodological and practical issues associated with the construction and application of QALYs. In addition, we define the problems encountered by behavioral medicine practitioners who apply these
measures. We concluded with some of the issues to be included in the application of QALYs in mental health care evaluation. These issues are particularly relevant to mental health care as interventions are explicitly aimed at improving life, rarely extending it. A critical issue in the future will be to increase utilization of existing HRQOL measures in clinical practice to help document the effectiveness of behavioral medicine interventions and position clinicians more prominently in the health care marketplace.

References


