BASIC SCIENCE REVIEW

Reprinted from ANNALS OF INTERNAL MEDICINE Vol. 118; No. 8, 15 April 1993
Printed in U.S.A.

Measuring Health-related Quality of Life

Gordon H. Guyatt, MD; David H. Feeny, PhD; and Donald L. Patrick, PhD, MSPH

■ Clinicians and policymakers are recognizing the importance of measuring health-related quality of life (HRQL) to inform patient management and policy decisions. Self- or interviewer-administered questionnaires can be used to measure cross-sectional differences in quality of life between patients at a point in time (discriminative instruments) or longitudinal changes in HRQL within patients during a period of time (evaluative instruments). Both discriminative and evaluative instruments must be valid (really measuring what they are supposed to measure) and have a high ratio of signal to noise (reliability and responsiveness, respectively). Reliable discriminative instruments are able to reproducibly differentiate between persons. Responsive evaluative measures are able to detect important changes in HRQL during a period of time, even if those changes are small. Health-related quality of life measures should also be interpretable—that is, clinicians and policymakers must be able to identify differences in scores that correspond to trivial, small, moderate, and large differences.

Two basic approaches to quality-of-life measurement are available: generic instruments that provide a summary of HRQL; and specific instruments that focus on problems associated with single disease states, patient groups, or areas of function. Generic instruments include health profiles and instruments that generate health utilities. The approaches are not mutually exclusive. Each approach has its strengths and weaknesses and may be suitable for different circumstances. Investigations in HRQL have led to instruments suitable for detecting minimally important effects in clinical trials, for measuring the health of populations, and for providing information for policy decisions.

What Is Health-related Quality of Life?

Health status, *functional status*, and *quality of life* are three concepts often used interchangeably to refer to the same domain of "health" (1). The health domain ranges from negatively valued aspects of life, including death, to the more positively valued aspects such as role function or happiness. The boundaries of definition usually depend on why one is assessing health as well as the particular concerns of patients, clinicians, and researchers. We use the term *health-related quality of life* (HRQL) because widely valued aspects of life exist that are not generally considered as "health," including income, freedom, and quality of the environment. Although low or unstable income, the lack of freedom, or a low-quality environment may adversely affect health, these problems are often distant from a health or medical concern. Clinicians focus on HRQL, although when a patient is ill or diseased, almost all aspects of life can become health related.

Why Measure HRQL?

HRQL is important for measuring the impact of chronic disease (2). Physiologic measures provide information to clinicians but are of limited interest to patients; they often correlate poorly with functional capacity and well-being, the areas in which patients are most interested and familiar. In patients with chronic heart and lung disease, exercise capacity in the laboratory is only weakly related to exercise capacity in daily life (3). Another reason to measure HRQL is the commonly observed phenomena that two patients with the same clinical criteria often have dramatically different responses. For example, two patients with the same range of motion and even similar ratings of back pain may have different role function and emotional well-being. Although some patients may continue to work without major depression, others may quit their jobs and have major depression.

These considerations explain why patients, clinicians, and health care administrators are all keenly interested in the effects of medical interventions on HRQL (4). Administrators are particularly interested in HRQL because the case mix of patients affects use and expenditure patterns, because increasing efforts exist to incorporate HRQLs as measures of the quality of care and of clinical effectiveness, and because payers are beginning to use HRQL information in reimbursement decisions.

Annals of Internal Medicine. 1993;**118**:622-629.

From McMaster University, Hamilton, Ontario, Canada; and the University of Washington, Seattle, Washington. For current author addresses, see end of text.

Abbreviations

HRQL	health-related quality of life
MOS	Medical Outcome Study

Table 1. Modes of Administration of HRQL Measures

Mode of Administration	Strengths	Weaknesses
Interviewer	Maximizes response rate Few, if any, missing items Minimizes errors of misunderstanding	Requires many resources, training of interviewers May reduce willingness to acknowledge problems Limits format of instrument
Telephone	Few, if any, missing items Minimizes errors of misunderstanding Less resource intensive than interviewer-administered mode	
Self	Minimal resources required	Greater likelihood of low-response rate, missing items, misunderstanding
Surrogate responders	Reduces stress for target group (very elderly or sick)	Perceptions of surrogate may differ from target group

The Structure of HRQL Measures

Some HRQL measures consist of a single question that essentially asks "How is your quality of life?" (5) This question may be asked in a simple or a sophisticated fashion, but either way it yields limited information. More commonly, HRQL instruments are questionnaires made up of a number of *items* or questions. These items are added up in a number of *domains* (also sometimes called *dimensions*). A *domain* or *dimension* refers to the area of behavior or experience that we are trying to measure. Domains might include mobility and self-care (which could be further aggregated into physical function), or depression, anxiety, and well-being (which could be aggregated to form an emotional-function domain). For some instruments, investigators do rigorous valuation exercises in which the importance of each item is rated in relation to the others. More often, items are equally weighted, which assumes that their value is equal.

Modes of Administration

The strengths and weaknesses of the different modes of HRQL administration are summarized in Table 1. Health-related quality-of-life questionnaires are either administered by trained interviewers or self-administered. The former method is resource intensive but ensures compliance, decreases errors, and decreases missing items. The latter approach is much less expensive but increases the number of missing subjects and increases missing responses. A compromise between the two approaches is to have instruments completed with supervision. Another compromise is the phone interview, which decreases errors and decreases missing data but dictates a relatively simple questionnaire structure. Investigators have done initial experiments with computer-administration of HRQL measures, but this is not yet a common method of questionnaire administration.

Investigators sometimes use a *surrogate respondent* to predict results that would be obtained from the patient. For instance, McKusker and Stoddard (6) were interested in what patients might score on a general, comprehensive measure of HRQL—the Sickness Impact Profile—when they were too ill to complete the questionnaire. The investigators used a surrogate to respond on behalf of the patient but wanted assurance that surrogate responses would correspond to what patients would have said had they been capable of answering. They administered the Sickness Impact Profile to terminally ill patients who were still capable of completing the questionnaire and to close relatives of the respondents. The correlation between the two sets of responses was 0.55, and the difference between the two pairs of responses was greater than 6 on a 100-point scale for 50% of the patients. The results provide only moderate support for the validity of surrogate responses to the Sickness Impact Profile.

These results are consistent with other evaluations of ratings by patients and proxies. In general, the correspondence between respondent and proxy response to HRQL measures varies depending on the domain assessed and the choice of proxy. Proxy reports of more observable domains, such as physical functioning and cognition, are more highly correlated with reports from the patients themselves. For functional limitations, proxy respondents tend to consider patients more impaired (they overestimate patient dysfunction relative to the patients themselves). This is particularly characteristic of those proxies with the greatest contact with the respondent (7). For other sorts of morbidity, patients tend to report the most problems, followed by close relatives, and clinicians report the least. These findings have important clinical implications because they suggest that clinicians should concentrate on careful ascertainment of the reported behaviors and perceptions of patients themselves, and they should limit the inferences they make on the basis of the perceptions of the caregivers.

What Makes a Good HRQL Instrument?

Measuring at a Point in Time versus Measuring Change

The goals of HRQL measures include differentiating between people who have a better HRQL and those who have a worse HRQL (a *discriminative* instrument) as well as measuring how much the HRQL has changed (an *evaluative* instrument) (8). The construction of instruments for these two purposes is different. If we want to discriminate between those with and without thyroid disease, we would be unlikely to include fatigue as an item because fatigue is too common among people who do not have thyroid disease. On the other hand, in

measuring improvement in HRQL with treatment, fatigue, because of its importance in the daily lives of people with thyroid disease, would be a key item. In the next sections, we list key measurement properties separately for discriminative and evaluative instruments. The properties that make useful discriminative and evaluative instruments are presented in Table 2.

Signal and Noise

Investigators examining physiologic end points know that reproducibility and accuracy are the necessary attributes of a good test. For HRQL instruments, reproducibility means having a high signal-to-noise ratio, and accuracy translates into whether they are really measuring what they intended to measure. For discriminative instruments, the way of quantitating the signal-to-noise ratio is called *reliability*. If the variability in scores between patients (the signal) is much greater than the variability within patients (the noise), an instrument will be deemed reliable. Reliable instruments will generally show that stable patients have more or less the same results after repeated administration.

For evaluative instruments, those designed to measure changes within patients during a period of time, the method of determining the signal-to-noise ratio is called *responsiveness*. Responsiveness refers to an instrument's ability to detect change. If a treatment results in an important difference in HRQL, investigators want to be confident that they will detect that difference, even if it is small. Responsiveness will be directly related to the magnitude of the difference in score in patients who have improved or deteriorated (the signal) and the extent to which patients who have not changed provide more or less the same scores (the noise).

Validity When a Gold Standard Exists

Although no gold standard for HRQL exists, instances occur in which a specific target for an HRQL measure exists that can be treated as a criterion or gold standard. Under these circumstances, one determines whether an instrument is measuring what is intended using *criterion validity* (an instrument is valid if its results correspond to those of the criterion standard). Criterion validity is applicable when a shorter version of an instrument (the test) is used to predict the results of the full-length index (the gold standard). Another example is using an HRQL instrument to predict death. In this instance, the instrument will be valid if variability in survival between patients (the gold standard) is explained by the questionnaire results (the test). Self-ratings of health, like more comprehensive and lengthy measures of general health perceptions, include a patient's evaluation of physiologic, physical, psychological, and social well-being. Perceived health, measured through self-ratings, is an important predictor of death (9).

Validity When No Gold Standard Exists

Validity examines whether the instrument is measuring what it is intended to measure. When no gold or criterion standard exists, HRQL investigators have borrowed validation strategies from clinical and experimental psychologists who have dealt with the problem of deciding whether questionnaires examining intelligence, attitudes, and emotional function are really measuring what they are supposed to measure. The types of validity that psychologists have introduced include content and construct validity. *Face validity* examines whether an instrument appears to be measuring what it is intended to measure, and *content validity* examines the extent to which the domain of interest is comprehensively sampled by the items, or questions, in the instrument. Quantitative testing of face and content validity are rarely attempted. Feinstein (10) reformulated these aspects of validity by suggesting criteria for what he calls the *sensibility*, including the applicability of the questionnaire, its clarity and simplicity, likelihood of bias, comprehensiveness, and whether redundant items have been included. Because of their specificity, Feinstein criteria facilitate quantitative rating of an instrument's face and content validity (11).

Construct Validation

The most rigorous approach to establishing validity is called *construct validity*. A construct is a theoretically derived notion of the domain(s) we want to measure. An understanding of the construct will lead to expectations about how an instrument should behave if it is valid. Construct validity involves comparisons between measures and examines the logical relations that should exist between a measure and characteristics of patients and patient groups.

The first step in construct validation is to establish a *model* or theoretical framework that represents an understanding of what investigators are trying to measure. That theoretical framework provides a basis for understanding the behavior of the system being studied and allows hypotheses or predictions about how the instru-

Table 2. What Makes a Good HRQL Measure?

Instrument Property	Evaluative Instruments*	Discriminative Instruments†
High signal-to-noise ratio	Responsiveness	Reliability
Validity	Correlations of changes in measures during a period of time, consistent with theoretically derived predictions	Correlations between measures at a point in time, consistent with theoretically derived predictions
Interpretability	Differences within patients during a period of time can be interpreted as trivial, small, moderate, or large	Differences between patients at a point in time can be interpreted as trivial, small, moderate, or large

* Measure differences within a patient during a period of time.
† Measure differences between patients at a point in time.

ment being tested should relate to other measures. Investigators then administer a number of instruments to a population of interest and examine the data. Validity is strengthened or weakened when the hypotheses are confirmed or refuted. For example, a discriminative HRQL instrument may be validated by comparing two groups of patients: those who received a toxic chemotherapeutic regimen and those who received a less toxic regimen. An HRQL instrument should distinguish between these two groups; if it does not discriminate, something has gone wrong. Alternatively, correlations between symptoms and functional status can be examined; those patients with a greater number and severity of symptoms should have lower functional status scores on an HRQL instrument. Another example is the validation of an instrument discriminating between people according to some aspect of emotional function; results should correlate with existing measures of emotional function.

The principles of validation are identical for evaluative instruments, but their validity is shown when *changes* in the instrument being investigated correlate with *changes* in other related measures in the theoretically derived predicted direction and magnitude. For instance, the validity of an evaluative measure of HRQL for patients with chronic lung disease was supported by the finding of moderate correlations with changes in walk test scores (12).

The responsiveness of evaluative instruments may be compromised by *ceiling* effects in which patients with the best score may have substantial HRQL impairment or *floor* effects in which patients with the worst score may deteriorate further. Bindman and colleagues (13) found that hospitalized patients (who already had the lowest possible score on a generic measure, the Medical Outcome Study Short Form [MOS-20]), reported that their health became worse in the subsequent year. Clearly that deterioration could not be detected by the MOS-20—a floor effect. Ganiats and colleagues (14) found that patients who all had the highest possible scores (representing the best possible function) on a physical functioning scale (the Functional Status Index), varied considerably on their score on a generic utility measure, the Quality of Well-Being. Thus, some patients with the best possible Functional Status Index could still improve in their health status—a ceiling effect.

A Detailed Example of Construct Validation

The Inflammatory Bowel Disease Questionnaire was designed to measure disease-specific HRQL (15) and it includes 30 items directed at 4 domains: bowel symptoms, systemic symptoms, emotional function, and social function. Investigators administered the Inflammatory Bowel Disease Questionnaire (along with global ratings of change in function, global ratings of change by the physician and a relative, a Disease Activity Index, and the emotional function domain of a generic HRQL measure) to 42 patients with inflammatory bowel disease on two occasions separated by 1 month. At the time the investigation was planned, the investigators made predictions about how change in the Inflammatory Bowel Disease Questionnaire score should relate to change in the other measures if this questionnaire was really measuring HRQL. Examples of the predictions and the results are as follows:

1. The patient's global rating of change in disease activity should relate closely (correlation ≥ 0.5) with change in the bowel-symptoms dimension of the Inflammatory Bowel Disease Questionnaire. Correlation observed was 0.42.
2. Some relation (correlation ≥ 0.3) should exist between change in the Disease Activity Index and change in the bowel-symptoms dimension of the Inflammatory Bowel Disease Questionnaire. Correlation observed was 0.33.
3. Some relation (correlation ≥ 0.3) should exist between change in the Disease Activity Index and change in the systemic-symptoms dimension of the Inflammatory Bowel Disease Questionnaire. Correlation observed was 0.04.
4. Change in the emotional-function dimension of the Inflammatory Bowel Disease Questionnaire should relate closely (correlation ≥ 0.5) with change in the emotional-function dimension of the generic questionnaire. Correlation observed was 0.76.

Of the 10 predictions made, three were correct, but in five the correlation was slightly lower than predicted and in two the correlation was much lower than predicted. The results provided moderate support for the validity of the questionnaire, but more data are required before the instrument can be used with confidence.

This example illustrates that validation is not an all-or-nothing process. We may have varying degrees of confidence that an instrument is really measuring what it is supposed to measure. The sort of a priori predictions that characterize the validation of the Inflammatory Bowel Disease Questionnaire strengthen the validation process. Without such predictions, it is too easy to rationalize the observed correlations. Validation does not end when the first study with data concerning validity is published but continues with repeated use of an instrument. The more frequently an instrument is used, and the more situations in which it performs as expected, the greater our confidence in its validity. Perhaps we should never conclude that a questionnaire has "been validated," but rather we should suggest that strong evidence for validity has been obtained in a number of different settings and studies.

Interpretability

A final key property of an HRQL measure is *interpretability*. For a discriminative instrument, we could ask whether a particular score signifies that a patient is functioning normally or has mild, moderate, or severe impairment of HRQL? For an evaluative instrument, we might ask whether a particular change in score represents a trivial, small but important, moderate, or large improvement or deterioration.

A number of strategies are available to make HRQL scores interpretable (16). For an evaluative instrument, one might classify patients into those who had important improvement as well as those who did not and examine the changes in score in the two groups; inter-

Table 3. Characteristics of Measures of Health-related Quality of Life

Approach	Strengths	Weaknesses
Generic instruments		
Health profile	Single instrument	May not focus adequately on area of interest
	Detects differential effects on different aspects of health status	May not be responsive
	Comparison across interventions, conditions possible	
Utility measurement	Single number representing net impact on quantity and quality of life	Difficulty determining utility values
	Cost-utility analysis possible	Does not allow examination of effect on different aspects of quality of life
	Incorporates death	May not be responsive
Specific instruments	Clinically sensible	Does not allow cross-condition comparisons
Disease specific	May be more responsive	May be limited in terms of populations and interventions
Population specific		
Function specific		
Condition or problem specific		

pret observed changes in HRQL measures in terms of elements of those measures that will be familiar to readers (for instance, descriptions of changes in mobility); or determine how scores in HRQL measures relate to marker states that are familiar and meaningful to clinicians. Data suggest that small, medium, and large effects correspond to changes of approximately 0.5, 1.0, and greater than 1.0 per question for instruments that present response options on seven-point scales (17). For instance, in a domain with 6 items, changes of 3 or 4 represent small effects, changes of 5 or 6 represent moderate effects, and changes of 7 or more represent large effects. Investigators used this information to interpret a recent trial that showed use of bronchodilators results in a small but clinically important improvement in dyspnea, fatigue, and emotional function in patients with chronic airflow limitation (18). In a study (19) of patients with arthritis, a change of 0.02 points in the Quality of Well-Being utility instrument was equivalent to all treated patients improving from moving their own wheelchair without help to walking with physical limitations.

In its use as a discriminative instrument, we know how patients in various health states score on the Sickness Impact Profile: patients shortly after hip replacement have scores of 30 that decrease to less than 5 after full convalescence (20); scores in patients with chronic airflow limitation, severe enough to require home oxygen, are approximately 24 (21); scores in patients with chronic, stable angina are approximately 11.5 (22); scores in those with arthritis vary from 8.2 to 25.8 in patients with American Rheumatism Association arthritis class I to class IV (23). The availability of data to improve the interpretability of HRQL measures is likely to increase exponentially in the next decade.

Types of HRQL Measures

Generic Instruments

Health Profiles

Two basic approaches characterize the measurement of HRQL: *generic instruments* (including single indicators, health profiles, and utility measures) and *specific instruments* (Table 3) (24). *Health profiles* are instruments that attempt to measure all important aspects of HRQL. The Sickness Impact Profile is an example of a health profile and includes a physical dimension (with categories of ambulation, mobility, as well as body care and movement); a psychosocial dimension (with categories including social interaction, alertness behavior, communication, and emotional behavior); and five independent categories including eating, work, home management, sleep and rest, as well as recreations and pasttimes. Major advantages of health profiles include dealing with a variety of areas and use in any population, regardless of the underlying condition. Because generic instruments apply to a variety of populations, they allow for broad comparisons of the relative impact of various health care programs. Generic profiles may, however, be unresponsive to changes in specific conditions.

Utility Measures

The other type of generic instrument, *utility measures* of quality of life, are derived from economic and decision theory; they reflect the preferences of patients for treatment process and outcome. The key elements of utility measures are that they incorporate preference measurements and relate health states to death. Thus, they can be used in *cost-utility analyses* that combine duration and quality of life. In utility measures, HRQL is summarized as a single number along a continuum that usually extends from death (0.0) to full health (1.0) (although scores less than zero, representing states worse than death, are possible [25]). Utility scores reflect both the health status and the *value* of that health status to the patient. The usefulness of utility measures in economic analysis is important when health care providers are asked to justify the resources devoted to treatment.

Utility measures provide a single summary score of the net change in HRQL—the HRQL gains from the treatment effect minus the HRQL burdens of side effects. Utility measures are useful for determining if patients are, overall, better off, but they do not show the domains in which improvement or deterioration occurs.

The simultaneous use of a health profile or specific instruments can complement the utility approach by providing this valuable information. The preferences in utility measurements may come directly from individual patients who are asked to rate the value of their health state. Alternatively, patients can rate their health status using a multiattribute, health-status classification system (such as the Quality of Well-Being scale). A previously estimated scoring function (derived from results of preference measurements from groups of other patients or from the community) is then used to convert health status to a utility score (26).

Specific Instruments

The second basic approach to quality-of-life measurement focuses on aspects of health status that are specific to the area of primary interest. The rationale for this approach lies in the potential for increased responsiveness that may result from including only important aspects of HRQL that are relevant to the patients being studied. The instrument may be specific to the disease (such as heart failure or asthma), to a population of patients (such as the frail elderly), to a certain function (such as sleep or sexual function), or to a problem (such as pain). In addition to the likelihood of improved responsiveness, specific measures have the advantage of relating closely to areas routinely analyzed by clinicians.

Choosing the Appropriate HRQL Measure

Health Status Surveys

The choice of an HRQL measure depends on the purpose of the study (27). Generic measures may be particularly useful for surveys that attempt to document the range of disability in a general population or a patient group. In one survey, investigators used the Sickness Impact Profile to examine the extent of disability in patients with chronic airflow limitation (4). Their striking finding was that the effect of chronic airflow limitation in patients' lives was not restricted to areas such as ambulation and mobility but was manifested in virtually every aspect of HRQL. This included social interaction, alertness behavior, emotional behavior, sleep and rest, as well as recreation and pasttime activities. For surveys investigating the range of disability, specific measures are unlikely to be of use, and investigators must rely on health profiles or the closely related, multiattribute, health status classification and utility function approaches.

Clinical Trials

Clinical investigators are, with increasing frequency, choosing HRQL measures as primary and secondary outcomes in clinical trials. Initially, when studying a new therapy (such as a new drug), investigators rely on disease-specific measures. Disease-specific measures are clinically sensible in that patients and clinicians intuitively find the items directly relevant; their increased potential for responsiveness is particularly compelling in the clinical trial setting. Investigators will have additional reasons for choosing a disease-specific measure if no other outcomes exist that are directly clinically relevant to the patient. A recent study (28) used a questionnaire designed specifically for patients with chronic renal failure and showed that erythropoietin-induced increases in hemoglobin levels improved HRQL in renal-failure patients.

A number of specific measures can be used together in a battery to obtain a comprehensive picture of the impact of different interventions on HRQL. A variety of instruments, including measures of well-being, physical function, emotional function, sleep, sexual function, and side effects, were used to show that antihypertensive agents have a differential impact on many aspects of HRQL (29). This trial showed that an angiotensin-converting–enzyme inhibitor was not as effective, when used alone, as a beta-antagonist or methyl-dopa. The angiotensin-converting–enzyme inhibitor was, however, found to have substantially less adverse effects on HRQL. One would adduce substantially different treatment recommendations from this trial if one considered only the effect of medication on blood pressure rather than both the effects on blood pressure and HRQL. The potential disadvantages of this approach are that the multiple comparisons made and the lack of a unified scoring system may lead to difficulties in interpretation. A study examining multiple outcomes runs the risk that, simply by chance, one or two outcomes will favor an experimental treatment. When this happens, a possibility exists that a useless or marginally effective treatment will be interpreted as showing an important improvement in HRQL.

A number of situations exist in which generic measures are highly appropriate for clinical trials. If a clinical outcome of direct relevance to patients already exists (such as myocardial infarction or stroke), a generic HRQL measure can provide complementary information about the range and magnitude of treatment effects on HRQL. Previously unrecognized adverse experiences may be detected. If the efficacy of an intervention is established, the purpose of a clinical trial may be to elucidate the full impact of a treatment. Utility measures are particularly relevant if the economic implications of an intervention are a major focus of investigation. In one randomized trial, investigators (30) showed that a compliance-enhancing maneuver for patients with chronic lung disease having exercise rehabilitation improved HRQL, and the cost was approximately $25 000 per quality-adjusted life-year gained.

Another instance in which generic measures may be particularly appropriate is when a real trade-off may exist between length of life or length of remission and quality of life. Such situations include chemotherapy for malignant disease and anti-viral agents for patients with human immunodeficiency virus (HIV) infection. A recent trial of zidovudine for mildly symptomatic HIV infection showed that the drug increased the period of progression-free survival by an average of 0.9 months. However, when the investigators (31) used a technique called "Quality-Adjusted: time without symptoms or toxicity" (Q-TWIST), which counts either disease progression or severe adverse events as negative out-

comes, patients treated with zidovudine did poorly. Thus, the HRQL perspective can reverse the treatment decision.

Having presented situations in which specific and generic measures are likely to be particularly appropriate, it is worth pointing out that use of multiple types of measures in clinical trials yields additional information. A randomized trial of patients with severe rheumatoid arthritis showed not only that patients receiving oral gold were better off in terms of disease-specific functional measures but also that they had higher utility scores than did patients receiving placebo (32). The investigators showed the impact of the treatment using measures of direct relevance to both patients and health workers and provided the information necessary for an economic cost-utility analysis. Perhaps health profiles and utility measures should be included in any clinical trial in which the major focus is patient benefit. Disease-specific measures are of greatest interest to the patients themselves and to the clinicians who treat them, whereas generic measures, because they permit comparisons across conditions and populations, are of greatest interest to the policy or decision maker. Thus, use of both categories of measures will be appropriate when the results could interest both audiences. Health-related quality of life measures may be used in clinical practice, providing clinicians with information they might not otherwise obtain. Forms that can be self-administered and immediately scanned by computer can be used to provide rapid feedback of HRQL data to clinicians.

Shortening a Long Instrument

Distilling the measurement of HRQL into a few key questions is a goal for clinical investigators. One approach is to develop a long instrument, test it, and use its performance to choose key questions to include in a shorter index. This approach has been used to create shorter questionnaires based on the lengthy instruments from the Medical Outcomes Survey (33).

How would one determine if the shortened questionnaire is an adequate substitute for the full version? The issue for discriminative purposes is the extent to which people are classified similarly by the short and long forms of the questionnaire. Statistically, one would examine the extent to which variance, or variability in scores, in the full instrument is predicted or explained by scores of the short version. If the rating of people's quality of life by the shorter instrument corresponds to ratings by the longer version, we should be comfortable with the substitution.

For evaluative purposes, the responsiveness and validity of the shorter version should be tested against the full instrument. If both correlations of change with independent measures and instrument responsiveness were comparable, it is appropriate to substitute the shorter instrument. If measurement properties deteriorated, the investigator needs to decide whether trading off respondent burden is worth the increases in sample size necessitated by a less responsive instrument.

Translating HRQL Questionnaires

If a questionnaire in a different language is used, a simple translation is unlikely to be adequate. Although experience with translations is still limited, we know that without rigorous *back-translation* and *pretesting* the instrument may be interpreted differently in the new language (34). Even if the translation is adequate, cultural differences can adversely effect an instrument's measurement properties (35). To be fully confident of an instrument's validity in a new language or culture, a complete repetition of the validation process is required (36).

Information Sources for HRQL Measurement

Many generic and specific HRQL measures now exist that have good validation data. The use of HRQL measures facilitates the choice of optimal treatment for individual patients, development of clinical and public policy guidelines, as well as conduct of economic analyses. Compendia of available measures (37), including critical reviews (38, 39), can facilitate choice of an instrument for a specific setting or purpose. HRQL measures are likely to become methodologically more sophisticated as well as simpler to use and to interpret (40). Clinical investigators with a strong interest in determining the effects of medical interventions on HRQL may find that collaboration with an expert in HRQL measurement, most often a social scientist, will enhance the quality of their work.

Grant Support: Dr. Guyatt is a Career Scientist of the Ontario Ministry of Health.

Requests for Reprints: Gordon H. Guyatt, MD, Room 2C12, McMaster University Health Sciences Centre, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5.

Current Author Addresses: Drs. Guyatt and Feeny: Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Centre, 1200 Main Street West, Hamilton, Ontario, Canada, L8N 3Z5.
Dr. Patrick: University of Washington, School of Public Health and Community Medicine, Department of Health Services, Sc-37, Seattle, Washington 98195.

References

1. **Patrick DL, Bergner M.** Measurement of health status in the 1990s. Annu Rev Public Health. 1990;11:165-83.
2. **Patrick DL, Erickson P.** Health status and health policy: quality of life in health care evaluation and resource allocation. Oxford University Press. New York, 1993.
3. **Guyatt GH, Thompson PJ, Berman LB, Sullivan MJ, Townsend M, Jones NL, et al.** How should we measure function in patients with chronic heart and lung disease? J Chron Dis. 1985;38:517-24.
4. **Wennberg JE.** Outcomes research, cost containment, and the fear of health care rationing. N Engl J Med. 1990;323:1202-4.
5. **Torrance GW.** Measurement of health state utilities for economic appraisal. J Health Econ. 1986;5:1-30.
6. **McCusker J, Stoddart AM.** Use of a surrogate for the Sickness Impact Profile. Med Care. 1984;22:789-95.
7. **Rothman ML, Hedrick SC, Bulcroft KA, Hickam DH, Rubenstein LZ.** The validity of proxy-generated scores as measures of patient health status. Med Care. 1991;29:115-24.
8. **Kirshner B, Guyatt GH.** A methodologic framework for assessing health indices. J Chronic Dis. 1985;38:27-36.
9. **Mossey JM, Shapiro E.** Self-rated health: a predictor of mortality among the elderly. Am J Public Health. 1982;72:800-8.
10. **Feinstein AR.** Clinimetrics. New Haven, Conn.: Yale University Press; 1987:141-66.
11. **Oxman AD, Guyatt GH.** Validation of an index of the quality of review articles. J Clin Epidemiol. 1991;44:1271-8.
12. **Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW.**

A measure of quality of life for clinical trials in chronic lung disease. Thorax. 1987;42:773-8.
13. **Bindman AB, Keane D, Lurie N.** Measuring health changes among severely ill patients. The floor phenomenon. Med Care. 1990;28: 1142-52.
14. **Ganiats TG, Palinkas LA, Kaplan RM.** Comparison of Quality of Well-Being Scale and Functional Status Index in patients with atrial fibrillation. Med Care. 1992;30:958-64.
15. **Guyatt GH, Mitchell A, Irving EJ, Singer J, Williams N, Goodacre R, et al.** A new measure of health status for clinical trials in inflammatory bowel disease. Gastroenterology. 1989;96:804-10.
16. **Guyatt GH, Feeny D, Patrick D.** Proceedings of the International Conference on the Measurement of Quality of Life as an Outcome in Clinical Trials: Postscript. June 14-17, 1989. Controlled Clin Trials. 1991;12:266S-269S.
17. **Jaeschke R, Singer J, Guyatt G.** Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clin Trials. 1989;10:407-15.
18. **Guyatt GH, Townsend M, Pugsley SO, Keller JL, Short HD, Taylor DW, et al.** Bronchodilators in chronic airflow limitation, effects on airway function, exercise capacity and quality of life. Am Rev Respir Dis. 1987;135:1069-74.
19. **Thompson MS, Read JL, Hutchings HC, Peterson M, Harris ED Jr, et al.** The cost effectiveness of auranofin: results of a randomized clinical trial. J Rheumatol. 1988;15:35-42.
20. **Bergner M, Bobbitt RA, Carter WB, Gilson BS.** The Sickness Impact Profile: development and final revision of a health status measure. Med Care. 1981;19:787-805.
21. **McSweeney AJ, Grant I, Heaton RK, Adams KM, Timms RM, et al.** Life quality of patients with chronic obstructive pulmonary disease. Arch Intern Med. 1982;142:473-8.
22. **Fletcher A, McLoone P, Bulpitt C.** Quality of life on angina therapy: a randomised controlled trial of transdermal glyceryl trinitrate against placebo. Lancet. 1988;2:4-8.
23. **Deyo RA, Inui TS, Leininger JD, Overman SS.** Measuring functional outcomes in chronic disease: a comparison of traditional scales and a self-administered health status questionnaire in patients with rheumatoid arthritis. Med Care. 1983;21:180-92.
24. **Patrick DL, Deyo RA.** Generic and disease-specific measures in assessing health status and quality of life. Med Care. 1989;27:5217-32.
25. **Boyle MH, Torrance GW, Sinclair JC, Horwood SP.** Economic evaluation of neonatal intensive care of very-low-birth-weight infants. N Engl J Med. 1983;308:1330-7.
26. **Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S.** A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol. 1992; 10:923-8.
27. **Patrick DL.** Health-related quality of life in pharmaceutical evaluation. Forging progress and avoiding pitfalls. PharmacoEconomics. 1992;1:76-8.
28. **Laupacis A.** Changes in quality of life and functional capacity in hemodialysis patients treated with recombinant human erythropoietin. The Canadian Erythropoietin Study Group. Semin Nephrol. 1990;10(2 Suppl 1):11-9.
29. **Croog SH, Levine S, Testa MA, Brown B, Bulpitt CJ, Jenkins CD, et al.** The effects of antihypertensive therapy on the quality of life. N Engl J Med. 1986;314:1657-64.
30. **Toevs CD, Kaplan RM, Atkins CJ.** The costs and effects of behavioral programs in chronic obstructive pulmonary disease. Med Care. 1984;22:1088-100.
31. **Gelber RD, Lenderking WR, Cotton DJ, Cole BF, Fischl MA, Goldhirsch A, et al.** Quality-of-life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection. Ann Intern Med. 1992;116:961-6.
32. **Bombardier C, Ware J, Russell IJ, Larson M, Chalmers A, Read JL.** Auranofin therapy and quality of life in patients with rheumatoid arthritis. Results of a multicenter trial. Am J Med. 1986;81:565-78.
33. **Stewart AL, Hays RD, Ware JE Jr.** The MOS short-form general health survey. Reliability and validity in a patient population. Med Care. 1988;26:724-35.
34. **Berkanovic E.** The effect of inadequate language translation on Hispanics' responses to health surveys. Am J Public Health. 1980; 70:1273-81.
35. **Deyo RA.** Pitfalls in measuring the health status of Mexican Americans: comparative validity of the English and Spanish Sickness Impact Profile. Am J Public Health. 1984;74:569-73.
36. **Nord E.** EuroQol: health-related quality of life measurement. Valuations of health states by the general public in Norway. Health Policy. 1991;18:25-36.
37. **Spilker B.** Quality of life assessments in clinical trials. Raven Press, 1990.
38. **McDowell I, Newell C.** Measuring Health. Oxford University Press, 1987.
39. **Patrick D, Erickson P.** Health status and health policy: quality of life in health care evaluation and resource allocation. New York. Oxford University Press, 1992.
40. **Feeny D, Guyatt GH, Patrick DL.** Proceedings of the International Conference on Quality of Life as an Outcome in Clinical Trials. Controlled Clin Trials. 1991;12(4 Suppl):79S-280S.