Measuring health-related quality of life: General issues

Gordon H Guyatt MD
Department of Clinical Epidemiology and Biostatistics and Department of Medicine, McMaster University, Hamilton, Ontario

GH Guyatt MD. Measuring health-related quality of life: general issues. Can J Respir J 1997;4(3):123-130.
Clinicians and policy makers recognize the importance of measuring health-related quality of life (HRQL) to make informed patient management and policy decisions. Self- or interviewer-administered questionnaires can be used to measure cross-sectional differences in quality of life among patients at a point in time (discriminative instruments) or longitudinal changes in HRQL within patients over time (evaluative instruments). Both discriminative and evaluative instruments must be valid (ie, measure what they are supposed to measure) and have a high ratio of signal to noise (reliability and responsiveness for the two instruments, respectively). Reliable discriminative instruments are able to differentiate reproducibly among persons. Responsive evaluative measures are able to detect important changes in HRQL over time, even if those changes are small. HRQL should also be interpretable B that is, clinicians and policy makers must be able to identify differences in scores that correspond to trivial, small, moderate and large differences.

There are two basic approaches to quality of life measurement: generic instruments that attempt to provide a summary of HRQL and specific instruments that focus on problems associated with individual 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 under different circumstances. Investigations of HRQL have led to the development of instruments suitable for detecting minimally important effects in clinical trials, for measuring the health of populations and for providing information for policy decisions.

Key Words: *Clinical epidemiology, Clinical trials, Functional status, Quality of life*

Mesurer la qualité de vie liée à la santé : généralités

RÉSUMÉ : Les cliniciens et les décideurs reconnaissent l'importance de mesurer la qualité de vie liée à la santé (QVLS) pour assurer une prise en charge éclairée des patients et prendre des décisions politiques. Des questionnaires remplis par des interrogateurs ou des auto-questionnaires peuvent être utilisés pour mesurer les différences transversales dans le niveau de la qualité de vie des patients à un certain point dans le temps (instruments discriminants) ou les changements longitudinaux dans la qualité de vie liée à la santé chez des patients dans le temps (instruments évaluatifs). Les deux sortes d'instruments doivent être valides (c'est-à-dire mesurer ce qu'ils sont sensés mesurer) et démontrer un rapport élevé signal : bruit (respectivement la fiabilité et la sensibilité). Les instruments discriminants fiables peuvent différencier la reproductibilité parmi les individus. Les mesures d'évaluation sensibles peuvent détecter des changements importants dans la qualité de vie liée à la santé dans le temps, même si ces changements sont mineurs. La qualité de vie liée à la santé doit également être interprétée, ce qui veut dire, que les cliniciens et les décideurs doivent pouvoir identifier les différences dans les résultats qui correspondent à des différences importantes, modérées, petites et insignifiantes.

Il existe deux approches de base pour mesurer la qualité de vie : les instruments généraux dont le but est de fournir un résumé de la qualité de vie liée à la santé ; et des instruments spécifiques qui se concentrent sur les problèmes associés aux stades de la maladie chez un individu, sur des groupes de patients ou des domaines de fonction. Les instruments généraux incluent des profils de santé et des instruments qui génèrent des services de santé. Les approches ne s'excluent pas mutuellement. Chacune d'entre elles a ses forces et ses faiblesses et peut s'appliquer dans différents contextes. Les recherches concernant la qualité de vie liée à la santé ont entraîné la mise au point d'instruments capables de déceler des effets importants mais de faible intensité dans des essais cliniques, de mesurer la santé des populations et de fournir de l'information pour la prise de décisions politiques.

Correspondence and reprints: Dr G Guyatt, Room 2C12, McMaster University Health Sciences Centre, 1200 Main Street West, Hamilton, Ontario L8N 3Z5. Telephone 905-525-9140,ext 2900, fax 905-577-0017, e-mail guyatt@fhs.mcmaster.ca

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 positively valued aspects such as role function or happiness. The boundaries of definition usually depend on why health is being assessed and the particular concerns of patients, clinicians and researchers. We use the term 'health-related quality of life' (HRQL) because there are widely valued aspects of life that are not generally considered as 'health', including income, freedom and quality of the environment. While low or unstable income, lack of freedom or a low quality environment may adversely affect health, these problems are often distinct from health or medical concerns. For clinicians, HRQL is the appropriate focus, keeping in mind that when disease and illness are experienced by a patient, almost all aspects of life can become health-related.

WHY MEASURE HRQL?

The important role of HRQL in measuring the impact of chronic disease is increasingly acknowledged (2). Physiological measures provide important information to clinicians but are of limited interest to patients and often correlate poorly with functional capacity and well-being, the areas in which patients are most interested. In patients who are very disabled with chronic heart and lung disease, for instance, differences in exercise capacity among patients studied in the laboratory are only weakly related to differences in the capacity to perform day-to-day activities (3). Another reason to measure HRQL is the commonly observed phenomenon that two patients with the same clinical criteria often have dramatically different responses. For example, two patients with the same forced expiratory volume in 1 s (FEV_1), and even the same exercise capacity, as measured in the laboratory, may have different role functions and emotional well-being. While one patient may continue to work without depression, another patient may quit her job and experience major depression.

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

THE STRUCTURE OF HRQL MEASURES

Some HRQL measures consist of a single question, which essentially asks "how is your quality of life?" (5). This question may be asked in a simple or a very sophisticated fashion, with either method yielding limited information. More commonly, HRQL instruments are questionnaires made up of a number of *items* or questions. Items are combined into *domains* (also sometimes called dimensions). A domain or dimension refers to the area of behaviour or experience that is being measured. Domains may include mobility and self-care, which may be aggregated into physical function, depression, anxiety or well-being, which may be further aggregated to form an emotional function domain. For some instruments, investigators have undertaken rigorous valuation exercises that rate the importance of each item in relation to the others. More often, items are equally weighted, implying that their values are equal.

MODES OF ADMINISTRATION

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

Investigators often consider using surrogate respondents to predict results that they would get from the patients themselves. For instance, McCusker and Stoddart (6) were inter-

TABLE 1
Modes of administration of health related quality of life measures

Mode of administration	Strengths	Weaknesses
Interviewer-administered	– Maximal response rate – Few, if any, missing items – Minimal errors of misunderstanding	– Requires a lot of resources, training of interviewers – May reduce willingness to acknowledge problems
Telephone-administered	– Few, if any, missing items – Minimal errors of misunderstanding – Less resource intensive than interviewer-administered	– Limits format of instrument
Self-administered	– 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

TABLE 2
What makes a good health-related quality of life measure?

Instrument property	Evaluative instruments (Measuring differences within subjects over time)	Discriminative instruments (Measuring differences among subjects at a point in time)
High signal:noise ratio	Responsiveness	Reliability
Validity	Correlations of changes in measures over time consistent with theoretically derived predictions	Correlations between measures at a point in time consistent with theoretically derived predictions
Interpretability	Differences within subjects over time can be interpreted as trivial, small, moderate or large	Differences between subjects at a point in time can be interpreted as trivial, small, moderate or large

ested 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 wished to use 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 six, on a 100-point scale, in 50% of the cases. 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. As might be expected, 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 (ie, 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, lastly, clinicians. These findings have important clinical implications because they suggest that clinicians should concentrate on careful collection of reported behaviours and perceptions of patients themselves, limiting the inferences that they make based on perceptions of the caregivers.

WHAT MAKES A GOOD HRQL INSTRUMENT?

Measuring at a point in time versus measuring change: The goals of HRQL measurement include differentiating between people who have better or worse HRQL (*discriminative instruments*) and measuring to what extent HRQL has changed (*evaluative instrument*) (8). The construction of instruments for these two purposes may be quite different. For instance, let us consider an item such as walking up a flight of stairs. Many patients with chronic respiratory disease avoid stairs altogether. Therefore, when deciding which patients are worse or better off, this item may provide limited help (after all, one can't compare people with respect to an activity they never undertake). The item would therefore not be very useful as a discriminative instrument. On the other hand, there may be many patients for whom climbing stairs is a very important activity, and an intervention that reduced dyspnea when climbing stairs would be very beneficial. A question regarding dyspnea when climbing stairs may therefore be crucial as an evaluative instrument intended for use as an outcome in clinical trials. Properties that make useful discriminative and evaluative instruments are presented in Table 2.

Signal and noise: Investigators examining physiological end-points have long been aware that they must minimize the noise in their instruments and maximize the signal they are trying to detect. For discriminative instruments, quantifying the signal:noise ratio is known as 'reliability'. If the variability in scores among patients (the signal) is much greater than the variability within patients (the noise), an instrument is reliable. Reliable instruments generally demonstrate that stable patients show more or less the same results on repeated administration.

For evaluative instruments designed to measure changes within individuals over time, the method for determining the signal: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 wish to be confident they will detect that difference, even if it is small. Responsiveness will be directly related to the magnitude of the difference in score of patients who have improved or deteriorated (the signal) and to the extent that patients who have not changed obtain more or less the same scores (the noise).

Validity when there is a gold standard: Although there is no gold standard for HRQL, there are instances when there is a specific target for a HRQL measure 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 only valid

if its results correspond to those of the criterion standard. For instance, 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 a HRQL instrument to predict mortality. In this instance, the instrument will be valid to the extent that 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 individual evaluations of physiological, physical, psychological and social well-being. Perceived health, measured through self-ratings, is an important predictor of mortality (9).

Validity when there is no gold standard: Validity refers to whether an instrument is measuring what it is intended to measure. When there is no gold or criterion standard, HRQL investigators borrow validation strategies from clinical and experimental psychologists who have, for many years, 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* refers to whether an instrument appears to be measuring what it is intended to measure, and *content validity* refers to the extent that 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) has reformulated these aspects of validity by suggesting criteria for what he calls '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's criteria facilitate quantitative rating of an instrument's face and content validity (11).

The most rigorous approach to establishing validity is called 'construct validity'. A construct is a theoretically-derived notion of the domain(s) to be measured. An understanding of the construct will lead to expectations about how an instrument should behave if it is valid. Construct validity, therefore, involves comparisons between measures, and examination of the logical relationships that should exist between a measure and characteristics of patients and patient groups.

The first step in construct validation is to a 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 how the system being studied behaves and allows hypotheses or predictions about how the instrument 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 according to the extent the hypotheses are confirmed or refuted.

Consider an instrument designed to discriminate among patients with chronic airflow limitation according to their dyspnea in daily living. One might anticipate that such an instrument would correlate highly with other instruments designed to measure the same construct, and with patients' global ratings of their dyspnea in daily life. The questionnaire might have a moderate correlation with 6 min walk test distance. One might find a weaker relationship with FEV_1, with laboratory exercise capacity, or with clinicians or relatives' ratings of the patients' day-to-day dyspnea. If investigators adduced data consistent with these predictions, there would be confidence that the instrument was actually discriminating among patients according to their dyspnea from day-to-day activities. If data differed substantially from these predictions, the instrument would be viewed with considerably more skepticism.

The principles of validation are identical for evaluative instruments, but their validity is demonstrated by showing that 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 where patients with the best score nevertheless have substantial HRQL impairment, or 'floor' effects where patients with the worst score may yet deteriorate further. Bindman and colleagues (13) found that of hospitalized patients who already had the lowest possible score on the Medical Outcome Study Short Form (MOS-20), many reported 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 had the highest possible score (representing the best possible function) on a physical functioning scale, the Functional Status Index, varied considerably in their scores on a generic utility measure, the Quality of Well-Being (14). This implies that some patients with the best possible Functional Status Index could still improve their health status (a ceiling effect).

There may be varying degrees of confidence that an instrument is really measuring what it is supposed to measure. A priori predictions strengthen the validation process. Without such predictions, it is too easy to rationalize whatever correlations between measures are observed. Validation does not end when the first study with data concerning validity is published, but continues with repeated use of an instrument. The more an instrument is used, and the more varied the situations where it performs as expected if it were really doing its job, the greater the confidence in its validity. Perhaps one should never conclude that a questionnaire has 'been validated'; the best one can do is 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, one may 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 one may ask

TABLE 3
Taxonomy of measures of health-related quality of life

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

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 (15). For an evaluative instrument, one might ask patients to make global ratings of the degree of improvement or deterioration that they have experienced. One could then look at changes in questionnaire scores for those who have remained stable, experienced small but clinically important changes, or moderate and large changes. A second strategy involves assembling groups of patients and asking them to discuss their problems with one another. The patients then rate themselves as having the same degree of HRQL impairment, or a better or worse HRQL to a small, moderate or large degree, as the people with whom they have spoken. Questionnaire interpretability is enhanced when questionnaire scores are related to global ratings from patients.

Both these strategies have been applied to instruments to measure HRQL in patients with chronic airflow limitation (16,17). The results proved consistent: response options comprised seven-point scales in which small, medium and large effects corresponded to changes of approximately 0.5, 1.0 and greater than 1.0 per question. It appears that a similar interpretation can be applied to a questionnaire using seven-point scale response options for asthma, for both adults (18) and children (19). Investigators used this information to interpret a trial that showed that bronchodilators result in small but clinically important improvements in dyspnea, fatigue and emotional function. The availability of data that improve the interpretability of HRQL measures is likely to increase greatly in the next decade.

TYPES OF HRQL MEASURES

Generic instruments – health profiles: Two basic approaches characterize HRQL measurement: generic instruments (including single indicators, health profiles and utility measures) and specific instruments (Table 3) (20). 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, body care and movement); a psychosocial dimension (with categories including social interaction, alertness behaviour, communication and emotional behaviour); and five independent categories (including eating, work, home management, sleep and rest, and recreations and pastimes). Major advantages of health profiles are that they deal with a wide variety of areas and can be used in virtually any population, irrespective of the underlying condition. Because generic instruments apply to a wide variety of populations, they allow for broad comparisons of the relative impact of various health care programs. Generic profiles may, however, be less responsive to changes in specific conditions.

Generic Instruments – utility measures: The other type of generic instrument, utility measures of quality of life, are derived from economic and decision theory, and 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. This allows them to be used in cost-utility analyses, which 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 (21). Utility scores reflect both health status and the *value* of that health status to the patient. The usefulness of utility measures in economic analysis is increasingly important in an era of cost constraints in which health care providers are being asked to justify devoting the resources to treatment.

Utility measures provide a single summary score of the net change in HRQL – the difference between HRQL gains from the treatment effect and the HRQL burdens of side effects. Utility measures are, therefore, useful for determining whether patients are, in net terms, better off because of therapy, but may fail to reveal on which dimensions of HRQL the patients improved versus those they worsened. The simultaneous use of health profiles and specific instruments may complement the utility approach by providing this information.

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 multi-attribute 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 (22).

Specific instruments: The second fundamental 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 explored by clinicians.

CHOOSING THE RIGHT HRQL MEASURE

Health status surveys: The choice of an HRQL measure depends very much on the purpose of the study (23). 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). They found 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, including social interaction, alertness behaviour, emotional behavior, sleep and rest, and recreation and pastime activities. For surveys investigating range of disability, specific measures are unlikely to be useful, and investigators will therefore rely on health profiles or on the closely related multi-attribute 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. In the initial stages of studying a new therapy such as a new drug, investigators are likely to rely on a disease-specific measure. Disease-specific measures are clinically sensible because patients and clinicians intuitively find the items directly relevant. The increased potential for responsiveness of disease-specific measures is particularly compelling in the clinical trial setting. Investigators will have additional reasons for choosing a disease-specific measure if there are no other outcomes that are directly clinically relevant to the patient. Recent randomized trials, in which questionnaires designed specifically for patients with chronic respiratory disease demonstrated that respiratory rehabilitation resulted in small but important improvement in HRQL, illustrate the important information that disease-specific measures can provide (24,25).

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 wide variety of instruments, including measures of well-being, physical function, emotional function, sleep, sexual function and side effects, were used to demonstrate that antihypertensive agents have a differential impact on many aspects of HRQL (26). The trial showed that an angiotensin-converting enzyme (ACE) inhibitor was not as effective, when used alone, as a beta-antagonist or methyl-dopa (26). The ACE inhibitor was, however, found to have substantially less adverse effects on HRQL. Substantially different treatment recommendations would be adduced from this trial if only the effect of medication on blood pressure, rather than both the effects on blood pressure and HRQL, were considered. Potential disadvantages of this approach are that the multiple comparisons being made and the lack of a unified scoring system may lead to difficulties in interpretation. A study examining multiple outcomes runs the risk of suggesting a spurious advantage of treatments for one or two outcomes of chance. When this happens, it is possible that a useless or marginally effective treatment will be erroneously presented as demonstrating an important improvement in HRQL.

In a number of situations, generic measures are highly appropriate for clinical trials. If there is already a clinical outcome of direct relevance to patients, such as myocardial infarction or stroke, a generic HRQL measure may provide complementary information about the range and magnitude of treatment effects on HRQL. Previously unrecognized adverse experiences may, for instance, 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, for instance, investigators demonstrated that a compliance-enhancing manoeuvre for chronic lung disease patients undergoing exercise rehabilitation improved HRQL, and the cost was approximately US$25,000 per quality-adjusted life-year gained (27).

Generic measures may also be particularly appropriate when there may be a trade-off 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 demonstrated that the drug lengthened the period of progression-free survival by an average of 0.9 months. However, when the investigators used a technique called 'quality-adjusted time without symptoms or toxicity' (Q-TWIST), which counts either disease progression or severe adverse events as negative outcomes, patients treated with zidovudine actually fared less well (28). In this instance, the HRQL perspective could reverse the treatment decision.

Having illustrated 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 that may prove important. For instance, a randomized trial of patients with severe rheu-

matoid 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 patients receiving placebo (29). The investigators were able to demonstrate the impact of the treatment using measures of direct relevance to both patients and health workers and to provide the information necessary for an economic cost-utility analysis. An argument can be made for inclusion of a specific measure, health profile and utility measure 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, while generic measures, because they permit comparisons across conditions and populations, may be of greatest interest to the policy or decision maker. Therefore, use of both categories of measures is most appropriate when the results could interest both audiences. HRQL measures may also find a place in clinical practice, providing clinicians with information that 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 would be a dream come true for clinical investigators. One approach to achieving this goal 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, for example, to create shorter questionnaires based on the lengthy instruments from the Medical Outcomes Survey (30).

How does one determine whether the shortened questionnaire is an adequate substitute for the full version? For discriminative purposes, the issue is the extent to which people are classified similarly by the short and long forms of the questionnaire. Statistically, one can examine the extent to which variance or variability in scores in the full instrument is predicted or explained by scores from the abbreviated version. The more that quality of life ratings by the shorter instrument correspond with ratings by the longer version, the more comfortable investigators can be with the substitution.

For evaluative purposes, the responsiveness and validity of the shorter version should be tested against the full instrument. If correlations of change with independent measures and instrument responsiveness are comparable, substitution of the shorter instrument is desirable. If measurement properties deteriorated, the investigator faces a decision about trading-off respondent burden with increases in sample size necessitated by a less responsive instrument.

Translating HRQL questionnaires: If a questionnaire is required in a different language, a simple translation is not likely to be adequate. Without rigorous back-translation and pre testing, the instrument may be interpreted very differently in the new language (31). Even if the translation is adequate, cultural differences can adversely affect an instrument's measurement properties (32). To be fully confident of an instruments' validity in a new language or culture, a complete repetition of the validation process is required (33). A questionnaire for chronic lung disease has been translated into a number of European languages and, for at least two of them (Dutch and Spanish), has proved valid and responsive (34, unpublished data).

Information sources for HRQL measurement: There are now a large number of generic and specific HRQL measures with demonstrated strong measurement properties. These include specific instruments for adults and children with asthma, for rhinitis and for patients with chronic airflow limitation. The use of HRQL measures can help with decisions concerning the optimal treatment for individual patients, development of clinical and public policy guidelines, and conduct of economic analyses. As a result of ongoing work in this rapidly evolving area, HRQL measures are likely to become methodologically more sophisticated and simpler to use and interpret (35).

REFERENCES

1. Patrick D, 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. New York: Oxford University Press, 1993.
3. Guyatt GH, Thompson PJ, Berman LC, 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;29:1151-224.
8. Kirshner B, Guyatt GH. A methodologic framework for assessing health indices. J Chron Dis 1985;38:27-36.
9. Mossey J, Shapiro E. Self-rated health: a predictor of mortality among the elderly. Am J Public Health 1982;72:800-9.
10. Feinstein AR. Clinimetrics. New Haven: Yale University Press, 1987:146-66.
11. Oxman A, 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. Med Care 1990;28:1141-52.
14. Ganiats TG, Palinkas LA, Kaplan RM. Comparison of Well-Being Scale and Functional Status Index in patients with atrial fibrillation. Med Care 1992;30:958-64.
15. 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. Controlled Clin Trials 1991;12:266S-9S.
16. Jaeschke R, Guyatt G, Keller J, Singer J. Measurement of health status: Ascertaining the meaning of a change in quality-of-life questionnaire score. Controlled Clin Trials 1989;10:407-15.
17. Redelmeier D, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996;49:1215-9.
18. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal import change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81-7.
19. Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children with asthma. Qual Life Res 1996;5:27-34
20. Patrick CL, Deyo RA. Generic and disease-specific

measures in assessing helath status and quality of life. Med Care 1989;27:F217-32.
21. Boyle MG, 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.
22. Feeny D, Barr RD, Furlong W, et al. A conprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol 1992;10:923-8.
23. Patrick DL. Health-related quality of life in pharmaceutical evaluation. Forging progress and avoiding pitfalls. PharmacoEconomics 1992;1:76-8.
24. Wijkstra PJ, Ten Vergert EM, van Altena R, et al. Long term benefits of rehabilitation at home on quality of life and exercise tolerance in patients with chronic obstructive pulmonary disease. Thorax 1995;50:824-8.
25. Goldstein RS, Gort EH, Guyatt GH, Stubbing D, Avendano MA. Prespective randomized controlled trial of respiratory rehabilitation. Lancet 1994;344:1394-7.
26. Croog SH, Levine S, Testa MA, et al. The effects of antihypertensive therapy on the quality of life. N Engl J Med 1986;314:1657-64.
27. Toevs CD, Kaplan RM, Atkins CJ. The costs and effects of behavioral programs in chronic obstructive pulmonary disease. Med Care 1984;1088-100.
28. Gelber RD, Lenderking WR, Cotton DJ, 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.
29. Bombardier C, Ware J, Russell IJ, Larson M, Chalmers A, Read JL. Auronofin therapy and quality of life in patients with pheumatoid arthritis. Am J Med 1986;81:565-78.
30. Stewart AL, Hays RD, Ware JE. The MOS Short-form General Health Survey. Med Care 1988;26:724-31.
31. Berkanovic E. The effect of inadequate language translation on Hispanics' response to health surveys. Am J Public Health 1980;70:1273-81.
32. 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.
33. 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.
34. Wijkstra PJ, Ten Vergert EM, Van Altena R, et al. Reliability and vailidity of the chronic respiratory questionnaire (CRQ). Thorax 1994;49:465-7.
35. Feeny D, Guyatt GH, Patrick SL. Proceedings of the International Conference On Quality Of Life As An Outcome In Clinical Trials. Controlled Clinical Trials 1991;12(Suppl):795-805.