Volume 7 • Number 1 • 2004
VALUE IN HEALTH

A Comprehensive Strategy for the Interpretation of Quality-of-Life Data Based on Existing Methods

Patrick Marquis, MD,[1] Olivier Chassany, MD, PhD,[2] Linda Abetz, MA[1]

[1]Mapi Values, Boston, MA, USA; [2]Service de Medicine Interne A. Hôpital Lariboisiere, Paris, France

ABSTRACT

Objectives: Health-related quality of life (HRQL) instruments generally undergo a rigorous development and validation process. In contrast, methods for interpreting HRQL data are varied, and no comprehensive widely applicable procedure exists. Determining whether differences are statistically significant is the most common method, but this yields conclusions that may be difficult to understand in a clinical context or which may be of no practical value. Consequently, there is a need for a comprehensive interpretation strategy that gives results that are meaningful to a variety of audiences, including patients, clinicians, and decision-makers.

Methods: The review of available interpretation strategies revealed that not all methods are applicable to all questionnaires, and some strategies may be difficult to implement for interpreting trial results. In addition, the issues decision-makers may have when assessing HRQL results have not really been addressed: what is measured and what is the meaning beyond statistical significance?

Results: A comprehensive stepwise strategy, based on the most effective methods available, has been developed to address the key interpretation issues of decision-makers. It is structured around several steps: understanding the content of the questionnaire; evaluating the magnitude of changes and their statistical significance; determining whether results are clinically significant, e.g., whether the observed changes crossed ranges of established threshold for meaningful differences; comparing pre- and post-treatment scores distribution with norms of references; and relating score changes to other outcomes end points such as morbidity, death, compliance, resource utilization, or productivity.

Conclusions: The proposed strategy should help to structure and successfully address interpretation issues and thus make HRQL results more convincing.

Keywords: quality of life, interpretation, clinical significance, comprehensive strategy.

Introduction

Health-related quality of life (HRQL) end points and other patient-reported outcomes (PRO) such as satisfaction with treatment or symptoms questionnaires are becoming increasingly popular in clinical trials [1,2] and are being integrated into regulatory submission packages with increasing frequency. To be accepted as a scientific measure, a PRO questionnaire must undergo a validation process to confirm that it is reliably measuring what it was intended to measure. To be useful to decision-makers, the results must also be interpreted by attaching clinical meaning to numerical data. A number of issues make this a complex task for the decision-maker and suggest that both researchers and decision-makers would benefit from supplementing data with information on interpretation.

Because most instruments have been developed in the past 10 years, there is a relative lack of data available for most questionnaires, and the wealth of experience used in interpreting clinical variables simply does not exist for HRQL. The Hospital Anxiety and Depression scale (HADS) [3] used in patients with depression and anxiety is a good example of a questionnaire where clinical experience supports the interpretation of subjective data and where score changes can be related to clinically meaningful differences. In many cases, however, decision makers will not have had direct experience with a particular HRQL questionnaire.

The processes for development and validation of HRQL questionnaires are well defined [4–6]. In contrast, the interpretation process is not, resulting in a misconception that HRQL data are less meaningful than traditional efficacy data. At this time,

Address correspondence to: Patrick Marquis, Mapi Values, 15 Court Square, Suite 620, Boston, MA 02108. E-mail: Patrick.Marquis@Mapivaluesusa.com

 1098-3015/04/$15.00/93 93–104

interpretation is based primarily on statistical results, for example, determining whether a difference between treatment groups is statistically significant. Nevertheless, because a trivial numerical difference can become statistically significant given a sufficiently large sample, statistical significance alone is insufficient for concluding that HRQL changed in a meaningful way. In contrast, a meaningful difference may not reach a statistical level of significance owing to insufficient sample size leading to an increase in the type II error.

Meaningful difference is a subjective concept that can be approached from several perspectives [7]. From a patient's perspective, it may be defined as the score change that corresponds to a change in his/her experience of daily life. Decision-makers might define a meaningful difference as the smallest difference in a score that would mandate a change in the management of the patient [8]. Physicians may look at individual patient well-being, treatment success, and prognostic value. Payers may be most interested in HRQL effects if they predict changes in medical costs, and health authorities will be interested in linking results to resources used, work productivity, and general health and well-being.

The assessment of HRQL is further complicated by specific problems that make the interpretation of scores more difficult. For example:

- Concepts vary from one questionnaire to another, even for those targeting the same condition [9]. For example, the Functional Assessment of Cancer Therapy (FACT) [10] and the European Organization for the Treatment of Cancer (EORTC) [11] quality-of-life questionnaires are two well-validated and widely used measures for assessing cancer care. Concepts covered by the FACT include physical, functional, social, and emotional well-being and relationship with doctor, whereas the EORTC QLQ-C30 addresses global quality of life, physical, role, emotional, social, and cognitive functions and symptoms including nausea and vomiting, pain, fatigue, dyspnea, sleep disturbance, appetite loss, and constipation.
- The scale label does not always reflect the exact content of items in the scale. Moreover, the content of two similarly named scales may be very different. For example, the EORTC social functioning scale measures interference with family life and social activities, whereas the FACT social well-being scale evaluates family and friends' support, family communication, and acceptance and satisfaction with sex life. Thus, interpretation cannot be made on the basis of the scale name alone.
- The number of items and their scaling vary, leading to differences in anchors, direction, and magnitude of change. For example, the Nottingham Health Profile (NHP) [12] and the SF-36 [13] have both been used in patients with diabetes. A score of 0 reflects the best HRQL on the NHP scores, whereas 0 means the worst HRQL in each dimension of the SF-36. Furthermore, a 10-point change is not the same for both instruments and, even across the SF-36 dimensions, a 10-point change should be interpreted in different ways.

Several strategies have been proposed for interpreting HRQL scores. These address the meaning of data in relation to standardized ratios, clinical parameters, or other external references. Nevertheless, they are not always applicable in the context of a clinical trial or are not relevant for a disease-specific questionnaire or are limited to only one particular aspect. More recently, the clinical significance of quality-of-life measures in cancer patients has been addressed using a number of specific perspectives [14]. Sprangers and colleagues [14] have proposed a check list to clinicians for assessing meaningful changes in HRQL overtime.

So far, however, there is no comprehensive approach suitable for the interpretation of HRQL questionnaires and HRQL results from a global perspective. Most importantly, the decision-maker's point of view, considering results of a study using a new questionnaire, has not been the central perspective for interpretation, and the issues he or she may encounter have not been properly addressed.

The purpose of this review is to identify existing strategies for the interpretation of generic and disease-specific instruments and to devise a strategy for interpreting HRQL data that can be used across instruments. In doing this, we have gathered the experience in interpretation gained to date and have also considered the issues decision-makers face in trying to understand and interpret HRQL data. We have also attempted to address the fundamental issue of the criteria for demonstrating clinical significance, or "meaningful difference," in clinical trials. Specific issues related to study design, study quality, or missing data have not been addressed in this article either because they were considered nonspecific to HRQL data or because specific publications are available on the subject [15–17].

General Strategies for Interpretation of HRQL Data

Mandatory Starting Point: A Well-Developed and Validated HRQL Instrument

The development and validation of HRQL questionnaires is fundamental to understanding the content of a questionnaire and provides a foundation for interpreting scores. Findings and decisions taken during the development and validation stages will impact the interpretation; for example, the number of hypothesized scales, the names given to scales, and the scaling of answers. Validation, which can be regarded as the first step toward interpretation [18], consists of analyzing the construct being measured and includes the relevance of the scoring system, the reliability of the scores or measurement error, their validity, or extent to which scales measure what they are supposed to measure and their responsiveness [19]. Questionnaires frequently need to be translated into several different languages for use in international studies and each translation must be culturally adapted and validated [20,21]. The linguistic validation should be complemented by a psychometric analysis aimed at demonstrating the structural and technical equivalence of the translated versions with the original questionnaire [22,23].

Although precise knowledge of the content, scoring, reliability, and validity of a questionnaire is a prerequisite for its interpretation, the information provided by validation studies is not generally sufficient for a comprehensive and convincing interpretation of HRQL changes [18].

Distribution- and Anchor-Based Interpretation

In their comprehensive review of interpretation strategies, Lydick and Epstein [7] grouped interpretation methods into two broad categories: distribution-based and anchor-based interpretation (Table 1).

Distribution-based strategies use the statistical distribution of results and are therefore based on means and standard deviations, statistical tests of differences or changes over time. Effect size has become a popular approach [24]. It can be used to create a picture of the magnitude of changes in HRQL data and can be used as a benchmark against which other instruments can be measured. Scores from the EORTC HRQL questionnaire QLQ-C30, which assesses HRQL in patients with cancer, have been interpreted using effect sizes. For example, an effect size of 0.51 has been reported in women with metastatic breast cancer feeling a little better, whereas the effect size was 0.86 in women feeling very much better. This provides a sense of the means and differences between different clinical groups, resulting in a clinical benchmark for interpretation [25]. Other potential forms of distribution-based interpretation [26] shown in Table 1 do not appear to be very popular in HRQL research.

Anchor-based strategies relate—or anchor—HRQL levels or changes to clinical status or to other meaningful criteria such as life events or a global rating. Retrospective patient or clinician assessment of change in status falls into this category. Other examples of anchor-based interpretation are shown in Table 1. Most relevant approaches of anchor-based interpretation are presented in the "specific interpretation strategies" section. Lydick and Epstein [7] concluded that anchor-based interpretations were more likely to be relevant to clinicians than distribution-based strategies.

Table 1 Distribution- and anchor-based interpretation strategies

Distribution-based interpretation	Anchor-based interpretation
Statistical significance	Disease conditions
Effect size	Global rating
Reliable change	Life events
Proximity to mean	Threshold effect
Unit of change	Changes with time
Normative level of functioning	Changes with therapy
	Predictive (receiver operating characteristic)
	Predictive (correlation)

Content-, Construct-, Criterion-, and Norm-Based Interpretation

In developing the Medical Outcomes Survey (MOS) 36-item short form (SF-36), a widely used generic instrument, Ware and colleagues [18,27] used content-, construct-, norm-, and criterion-based interpretation strategies (Table 2).

Content-based interpretation involves examining the content of general health measures, using qualitative and quantitative descriptions of scales and their anchors. The content of a question determines what a respondent thinks as he or she answers the question and therefore provides information about what the answer means. While interpretation of meaning based on the content of the question is straightforward for single-item measures, it becomes more complex in multi-item scales. One approach is to use the percentage of response to one item as an internal criteria, which is easy to understand in terms of health status, to interpret the scores across scale levels. For example, in the SF-36, a representative item within the physical function-

Table 2 Content-, construct-, norm- and criterion-based interpretation strategies

Strategy	Definition	Example
Content-based	Interpretation is based on characterization of the content of the measure, using a specific item of a scale as an internal criteria.	To interpret the 36-item short form (SF-36) general health scale, ranging from 0 to 100, the percentage of the general US population that evaluates their health as excellent, good, fair, or poor has been described for each category of scores ranging from 10 to 10 (0 to 10, 11 to 20, etc.).
Construct-based	Based on how scales relate to one another, to the dimensions they were intended to measure, and to other conceptually related variables.	*Between questionnaires:* comparison of the SF-36 and the Nottingham Health Profile (NHP) shows high correlation between the physical morbidity, pain, mental health, and vitality scales of both questionnaires, suggesting that these scales are measuring similar concepts. *Within questionnaires:* the physical functioning, bodily pain, and role physical dimensions of the SF-36 are highly correlated, suggesting that there are some underlying similarities in the concepts being measured (physical function).
Norm-based	Normative data make it possible to interpret scale scores by comparison with scores for other individuals.	SF-36 scores of patients with migraine have been compared with norm scores to interpret the impact of migraine. Migraine particularly affects the bodily pain and role physical scores.
Criterion-based	This strategy relates changes in health-related quality-of-life (HRQL) scores to external variables.	*Ability to work and physical functioning:* 10 levels of the physical functioning scale of the SF-36 have been correlated with the percentages of medical outcomes survey (MOS) panel participants who said that their health kept them from working. These percentages can be used to interpret the physical functioning scale.

ing scale—the ability to walk one block—was assessed in a group of respondents who had a physical function score of 75 and also in those with a score of 45. It was found that 90% of people with a score of 75 were able to walk one block without limitation, but only 32% of those with a score of 45 could walk the same distance. This gives the investigators the context in which to distinguish between scores of 75 and 45 [27].

In construct-based interpretation, the relationship between or among scales is examined. Constructs are defined as variables that cannot be directly measured and that are thought to be responsible for the relationship between measured variables [28]. Understanding how questionnaire scale scores relate to underlying health concepts, how scales relate to other scales within the questionnaire, and the relationship of scale scores with scores on other questionnaires can help in interpreting the meaning of HRQL scores. A strong relationship between two sets of scales in a questionnaire means that they are measuring similar concepts and that there are probably overlaps in interpretation. For example, a factor analysis of the SF-36 revealed two components: one that is strongly related to the physically oriented scales of the SF-36, that is, physical function, bodily pain, and role physical scales, and another that was related to the mental components [29]. High correlations were observed between the physical, mental, vitality, and pain scales of the SF-36 and the NHP, indicating that these scales measure similar concepts [30].

Criterion-based interpretation uses information about how scale scores relate to external variables to determine their meaning; these external variables include clinical or socially meaningful life events such as job loss, utilization of health-care services, ability to work, or prediction of mortality. For example, the physical functioning scale of the SF-36 has been correlated with a patient's ability to work; almost 70% of patients with physical functioning scores of 20 were unable to work, compared with 3% to 6% of patients with physical functioning scores ≥80 [28]. In light of the relationship between the scale score range and an external criterion (ability to work), the social relevance of the physical functioning scale can be more easily understood.

Norm-based interpretation and known–group interactions were originally classified as examples of criterion-based interpretation [18]. Norm-based analysis involves calculating scores for a large population-based sample and using these norms to examine how the particular group under study deviates from this expected behavior in this case self-reported HRQL. This method has been used in interpreting SIP scores [31], where scores from a population of cardiac-arrest survivors were compared with those for a healthy population.

In known–group interactions, the meaning of a score is understood by referencing scores obtained in populations differing for known clinical characteristics. Hunt and colleagues [32] used this method in the interpretation of results from the NHP to

compare patients who consulted their physician with those who did not consult their physician, based on the assumption that nonconsulters were healthier than consulters. Patients who consulted their physician had consistently higher scores for every scale in the questionnaire, indicating a worse HRQL than nonconsulters.

Specific Interpretation Strategies

Interpretation Based on Life Events

Changes in HRQL scores have been calibrated against recognized stressful life events to assess the change in health during clinical studies of antihypertensive drugs. This provides a means of relating change in HRQL score with easily understood concepts such as moving house, getting divorced, or the death of a spouse. Testa and colleagues [33] have used this method to interpret the effects of treatment on the HRQL of men with hypertension over a 24-week study period. Changes in HRQL were indexed in a calibration model. The difference in the overall HRQL scores of 0.22 unit between the treatment groups was considered clinically relevant with a positive change of 0.11 for one treatment and a negative change of 0.11 for the other treatment, corresponding to life events such as major change in work responsibilities, problems with in-laws, or mortgage foreclosure. This method has also been used to define the threshold for clinically important changes.

Interpretation Based on Global Rating of Change

Minimal important difference using a 15-point intensity scale. A substantial amount of research has led to the development of the concept of minimal important difference (MID). Jaeschke and colleagues [8,34] defined the MID in HRQL scores as the smallest change in score large enough to mandate a modification of treatment from the clinician's perspective and the smallest change considered important by patients. When clinical measures are interpreted as part of decision-making, this concept is often understood intuitively, as clinicians use their experience to decide what is an important change in their patients. This combination of intuition and experience is not yet available in the majority of HRQL studies or instruments; consequently, Jaeschke and colleagues have attempted to establish a framework around which this might develop. Two analyses were performed: one in patients with heart and lung disease [8] and the second in patients with a variety of illnesses [34]. Two very similar HRQL measures were used: the Chronic Respiratory Questionnaire (CRQ) [35] and the Chronic Heart Failure Questionnaire (CHFQ) [36] and, later on, the Asthma Quality of Life Questionnaire (AQLQ) [37,38]. In these questionnaires, each item had a 7-point answer scale; for example, the question that asked "Please indicate how much shortness of breath you have had during the last 2 weeks while climbing stairs" had a range of seven answers from "Extremely short of breath" to "Not at all short of breath." A global rating of change, from –7, a very great deal worse, to +7, a very great deal better, was used as an anchor to define small, medium, and large changes. These changes were related to the mean change in CRQ, CHFQ, AQLQ, or subscale scores. Jaeschke and colleagues [8,39] only considered a global rating of 0 as no change for their analysis of the CRQ and CHFQ, whereas Juniper and colleagues [37,38] lumped 0, –1, and 1 global ratings together to constitute the "no change" category in their analysis of the AQLQ. Patients felt that a mean change per item of approximately 0.5 on a 7-point Likert scale was the minimal change that they perceived as significant, whereas differences of 0.5–1 and >1 were considered moderate and large, respectively (Table 3).

Table 3 Changes in combined CRQ and CHFQ scores corresponding to different global ratings of change

Global rating of change	Dyspnea score	Fatigue score	Emotional function score
0 (no change)	0.10	0.12	0.02
1–3 (small change)	0.43	0.64	0.49
4–5 (moderate change)	0.96	0.87	0.81
6–7 (large change)	1.47	0.94	0.86

CRQ, Chronic Respiratory Questionnaire; CHFQ, Chronic Heart Failure Questionnaire

Anchoring changes using a 5-point distress scale or a 7-point intensity scale. Anderson [40] has proposed a system for anchoring HRQL scores on a symptom distress scale in patients with angina. This involved administering a symptom distress index, consisting of 73 symptoms that were rated on a scale of 0 to 5, to measure both how much the patient suffered from a symptom and how distressing it was. Patient responses were: "did not have the symptom," "had it but no distress," "some distress," "moderate distress," "very much distress," and "extreme distress." These six answers were simplified to three distress levels: none, some, and extreme. Anderson defined five types of change: a change from none to extreme (two steps worse);

none to some and some to extreme (one step worse); none to none, some to some, and extreme to extreme (no change); extreme to some and some to none (one step better); and extreme to none (two steps better). These changes in symptom distress could be correlated with changes in the scores of a coadministered HRQL questionnaire, to increase the interpretability of the HRQL questionnaire score. This method has been used to help interpret the meaning of changes, with 95% confidence intervals, in the Mental Health Index for patients suffering from hypertension and angina [40].

Osoba and colleagues [26] used effect size related to a 7-point change subjective significance questionnaire (SSQ) in interpreting scores obtained from the EORTC QLQ-C30 questionnaire. Women with breast cancer were asked to rate changes in physical, emotional, and social functioning and global quality of life after chemotherapy. Patients who indicated a small change on the SSQ had an increase or decrease in the range 5 to 10 on the QLQ-C30 questionnaire. The mean change in QLQ-C30 score for patients who had a moderate change on the SSQ was in the range 10 to 20 and large changes correlated with a QLQ-C30 score change >20. Effect sizes increased in agreement with increasing changes in SSQ and QLQ-C30 scores. In this way, results from the questionnaire were anchored to patients' perception of changes in health status and patients effectively acted as their own comparison.

Interpretation Based on the Standard Error of Measurement (SEM)

Wyrwich and colleagues [41,42] have recently proposed using the one-SEM criterion as a proxy to evaluate clinically meaningful change. The SEM is estimated by multiplying the standard deviation of the scale by the square root of 1 minus its reliability coefficient [41]. Although the SEM is a distribution-based statistic evaluating the true score change, the authors have shown that one-SEM change corresponded well with the patient-driven MID using the CRQ [8]. The MID and SEM for each scale of the CRQ (dyspnea, fatigue, and emotional function) were, respectively, 0.43, 0.64, and 0.49 (MID) and 0.48, 0.53, and 0.41 (SEM). Weighted kappa coefficients between MID and one-SEM of patients classified as improved, stable, or declined were for each scale, respectively, 1.00, 0.87, and 0.91 ($P < .001$ for the three scales). The authors concluded that the SEM could be a useful tool in establishing the link between clinically relevant and statistically meaningful intraindividual changes. Results should be confirmed with other instruments.

Determining Threshold Scores Using a 4-Point Bother Scale

The American Urological Association has developed a symptom index for use in patients with benign prostatic hyperplasia (BPH) [43,44]. The scaling of the 7 items of this index ranged from 0 to 5. The authors used a global distress question: "Overall, how bothersome has any of your trouble with urination been in the last month (not at all, a little, some, a lot)?" to determine threshold values in a cross-sectional design. This method permits categorization of patients who completed the questionnaire once, into three groups according to symptom severity. The summated scores generated by the 7 items of this index range from 0 to 35; patients with scores from 0 to 7 were classified as having mild symptoms, those with scores from 8 to 19 were classed as moderate, and patients with scores of 20 were considered to have severe symptoms. This index has subsequently been used in the validation of the USA–Spanish version of the SF-36 in a Cuban-American population with BPH [45].

Examples of Interpretation for Existing Instruments

To illustrate the issues in interpretation of HRQL results, a review of the literature was undertaken to identify the approaches that have been used in interpreting a range of key questionnaires. These include the SF-36, SIP, and NHP, all generic instruments widely used in clinical trials and the EORTC QLQ-C30, the FACT, and the AQLQ, a specific questionnaire. The SF-36 has undergone the most extensive interpretation, based on numerous studies of norms, including general populations in the United States [27,29,46] and extensive validation in many countries [47–49] and specific statistical techniques like receiver operating characteristics (ROC) [50,51]. Two user manuals are also available for this questionnaire [27,29].

The SIP has also undergone content-based interpretation and clinical interpretation based on known groups [52], as well as extensive statistical interpretation, including effect size, MID based on patient satisfaction, and ROC [53]. A user manual and interpretation guide is available [52]. The development and the international validation of the NHP has led to the development of a European guide [54] including norms and known groups comparison.

The cancer-specific EORTC QLQ-C30 questionnaire has undergone clinical interpretation based on 14 published studies [25]. Effect size interpretation

has also been used [25,26]. Norwegian general population norms [55] and norms for Danish women [56] have been generated for use in interpretation. Other strategies used have mainly centered on criteria, including the MID [26] and a comparison of patient and observer ratings [57].

The FACT-G questionnaire has been interpreted based on the initial validation data used as reference [58] and, more recently, using the global rating of change to determine meaningful change for improvement or worsening [59]. A user manual is available for this questionnaire [60]. The clinically meaningful change of the FACT-L (lung) has been determined using a range of criteria such as performance status rating, weight loss, presence of primary disease symptoms, change over time related to disease progression, SEM, and effect size [61].

Interpretation of the AQLQ has been somewhat more limited and has concentrated on MID in a limited number of patients [38], with instability of the results according to the wording of the global question used [62]. A user manual is available for this questionnaire.

Although several methods have been used in the interpretation of some of the above-mentioned questionnaires, there is a scarcity of information on the comparability of the results obtained using different methods. In a recent literature review, Samsa and colleagues [63] noted similarities between clinical significance benchmarks of HRQL scores as determined by Cohen's standard effect sizes, and results from explicit anchor-based methods. In their own pilot study, the clinical significance of Health Utilities Index (HUI) scores was determined using the general health domain of the SF-36 and the SIP-total as external anchors. Similar HUI benchmarks (MIDs) were obtained using the distribution-based (effect size) method and anchor-based (SF-36 general health/SIP-total) method. More recently, Cella and colleagues [61] found concordant results using clinical or distribution methods to determine the clinically meaningful change of the FACT-L.

Proposed Stepwise Interpretation Strategy

The overview of the issues associated with interpretation of HRQL data revealed not only the diversity of methods available, but also the absence of a standardized procedure suitable for a variety of instruments or adapted to the interpretation of clinical trial results. This research has led us to formulate recommendations for a comprehensive strategy useful across instruments and studies. The proposed interpretation strategy comprises five main steps, each of them addressing a specific point corresponding to major issues decision-makers may encounter when assessing HRQL results. Each step provides a specific set of information that is complementary to the others:

- Full description of the content of the scale and its psychometric properties.
- Evaluation of the magnitude of the change or difference and its statistical significance.
- Comparison of observed changes with established meaningful magnitudes of changes or score calibration.
- Comparison of the baseline/follow-up scores with norms or available known-group references.
- Determination of the practical value of scores.

The different steps that should ideally be undertaken to generate the relevant data for each questionnaire are presented below with the specific points related to the interpretation of the results of a particular study. Knowledge of the content of the scale and its psychometric properties can be seen as a basic background for interpretation. The second step, dealing with statistical significance, is also mandatory to begin interpretation. These two steps should be supplemented by at least one of the remaining three steps to determine the clinical significance, based on either comparison of scores with meaningful change, comparison with norms and known-group references, or other methods to determine the practical or predictive value of the scores. In practice, it may be difficult to generate the level of information required for each step. The proposed steps should be seen as a strategy to accumulate evidence for demonstrating the meaningfulness of results. The higher the level of evidence, the more convincing the results. In Table 4, we have summarized these steps and the strategy that can be followed to interpret study results.

Content of the Scale and Its Psychometric Properties

This step should provide a clear and precise description of what is measured by a scale at the item level, including the meaning of the lowest and the highest possible scores. General wording like "social functioning" is not sufficiently informative to describe the content of a scale. This step, very often missing in the interpretation, may encompass the content criterion proposed by Ware [27] for interpreting the SF-36, but at the very least requires detailed description of the meaning of items and the underlying constructs being measured. Reliability and validity data should be provided to under-

Table 4 Proposed stepwise comprehensive strategy for interpretation of a new specific questionnaire and study results

Steps	Initial development of instruments and further use in observational/epidemiologic studies	Interpretation of study results
Two mandatory steps:		
1. Understanding the content of the scale and its psychometric properties	• Describe the content of the scale at item level, including the meaning of the lowest and highest possible scores and range of well-being. • Provide reliability coefficient and construct validity evidence.	• Acquire knowledge of the scale content and psychometric properties.
2. Evaluation of the magnitude of the change or of the difference and its statistical significance	• Calculate ES or SRM related to meaningful clinical changes to demonstrate the responsiveness of the scale.	• Calculate ES or SRM related to treatment effect and interpret them using existing criteria.
Supplemented by at least one of the following steps to determine the clinical significance:		
3. Comparison of changes with established thresholds for meaningful magnitudes of changes or score calibration	• Determine MID, SEM, or score calibration using global rating of change or distress scale, for example, or using a clinical anchor if relevant.	• Compare within-group changes to ranges of MID, SEM, or score calibration.
4. Comparison of the baseline/follow-up scores with norms or available known-group references	• Collect data in different relevant clinical severity groups and in nonsymptomatic patients or general population if relevant. • Provide full distribution for each group, including confidence intervals and percentage at floor and ceiling.	• Interpret baseline distribution according to references or norms. • Determine percentage of patients already in the well-being range at baseline using distribution parameters; relevant clinical groups used as references chosen according to the condition and the availability of meaningful clinical indicators. • Interpret distribution over time and end point distribution according to references or norms.
5. Understanding the practical values of scores	• Collect data using relevant parameters retrospectively or prospectively. Structure around three sets of criteria: morbidity and death, patient behavior (compliance and resource utilization), and consequences on work (loss of productivity or working days). • Describe the association of scale score ranges with the proposed set of criteria (e.g., percentage able to walk, number of treatments, treatment failure, visits to clinicians, days out of work). • Describe the predictive value of scores and their changes for the same set of criteria.	• Interpret changes with available parameters such as morbidity, patient behavior in terms of compliance or resource utilization, and consequences on work (productivity, days out of work). • Estimate NNT, when relevant, according to ranges of meaningful changes (i.e., ranges of MID), or to the number of patients having reached the expected level of well-being or normative level of functioning.

ES, effect size; SRM, standardized response mean; NNT, number needed to treat; MID, minimally important difference; SEM, standard error of measurement.

stand the homogeneity of the scale, its stability over time, and its relationships with other related constructs.

Magnitude of Changes and Statistical Significance

This component looks at the statistical significance of differences and aims to produce a directly interpretable description of the magnitude of changes. This is based on standardized ratios generating a common reference for the interpretation of changes, thereby avoiding the problem of distribution, or standard deviation, which varies from one scale to another. Effect size and standardized response mean are typical examples of useful statistics that may be generated. These ratios are used to evaluate the responsiveness of the instruments during psychometric validation. Most of the distribution-based strategies described by Lydick and Epstein [7] are included in this category.

Meaningful Change and Score Calibration

The determination of the MID and—probably in the future—the SEM, or the development of a score calibration are essential. Nevertheless, no consensus has yet been reached concerning the best way to determine the MID. Conceptual and practical difficulties regarding the use of retrospective patient global assessment have been discussed [62,64–66]; in particular, the paradox resulting from the use of a single, nonvalidated item to calibrate a validated multi-item questionnaire. Currently, the global rating of change or the subjective perceived change scales are the most commonly used methods to integrate the patient perspective. Keeping in mind the

need for further research, the use of ranges of MID, with confidence intervals, for example, is recommended to express thresholds. Other methods should be considered using clinical anchors [66,67,68] or other types of patient-reported outcomes (satisfaction with care [53] or comparative rating, for example [69]). The usefulness of the SEM should be further considered as a criterion for evaluating minimum important change.

One must remember that the MID is determined within a group and not between groups; for example, its use is relevant to interpret changes from baseline in each treatment group of a trial, but it should be used with care when comparing differences in change between treatment groups or in groups with noticeable differences in disease or sociodemographic characteristics.

References or Norms

Use of references or norms provides a distribution of scale scores in reference groups, such as the general population, or a population without any known disease [70]. In the case of a specific questionnaire, reference scores in the general population may be difficult to obtain and it may be necessary to use data from cured patients or patients without symptoms as references. Other known groups may include patients with a recognized and relevant clinical status.

Practical Value

This final component is often of greatest interest to clinicians and other interested parties as it relates HRQL data to practical variables, leading to concrete implications. This is the method in which a researcher might predict other outcomes for patients with a given HRQL score. Different outcomes can be used, such as the association with or the prediction of 1) morbidity (clinical status, relapses) and death; 2) patient behavior in terms of compliance and resource utilization (need for treatment, level of treatments needed, adherence, treatment switches, physician contacts, hospital stays, need for assistance); and 3) loss of productivity and days out of work.

Ware's criterion-based interpretation strategy, the number of patients who benefit from treatment and the number needed to treat (NNT), all fall in this category. The NNT can be calculated using the number of patients having reached the expected level of well-being or functioning derived from a reference population or the number of patients whose HRQL score has increased by a certain size judged to be clinically significant [71,72]. Indeed, this part is the most difficult to achieve because the amount of data required is high. Furthermore, there is a risk that emphasizing this criterion for interpretation minimizes the perceptions of the patient, whose condition is really the matter of interest.

Conclusions

Potential users of HRQL questionnaires are at the beginning of a lengthy learning process, which can be accelerated if instrument developers provide thorough documentation on the instruments produced. This information should have educational, convincing, and demonstrative values. Therefore, it should be relevant to the audience, based on known parameters, and structured around a standardized interpretation protocol. When practical constraints make it impossible to address the perspectives of all decision-makers, the challenge is to identify the primary audience and then determine which interpretation strategies are both feasible and relevant for addressing the needs of that audience [73].

One must keep in mind that the issues faced in the interpretation of HRQL data are issues that have also been encountered historically in clinical assessment. Only with time, experience, and data are clinical assessments standardized, modified, or rejected.

Our recommendation is that a user manual, which would include interpretation guidelines composed of different approaches, should accompany the development and validation of a questionnaire. The objective is to accumulate evidence to help decision-makers assess the usefulness and the meaning of study results. We recommend that interpretation include content, psychometric, and statistical significance information as well as at least one of the final three steps—meaningful change, references, or practical value. This involves the planning of interpretation perspectives from the beginning of the development of the questionnaire and also during the analysis and reporting of studies.

This study was supported financially by Mapi Values.

References

1 Wood-Dauphinee S. Assessing quality of life in clinical research: from where have we come and where are we going? J Clin Epidemiol 1999; 52:355–63.
2 Saunders C, Egger M, Donovan J, et al. Reporting on quality of life in randomised controlled trials: bibliographic study. BMJ 1998;317:1191–4.

3 Chassany O, Sagnier P, Marquis P, et al. Patient-reported outcomes. the example of health-related quality of life—a European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. Drug Inf J 2002;36:1–10.
4 Herrmann C. International experiences with the Hospital Anxiety and Depression Scale: a review of validation data and clinical results. J Psychosom Res 1997;42:17–41.
5 Streiner DL, Norman GR. Health Measurement Scales. A Practical Guide to Their Development and Use. New York: Oxford University Press, 1989.
6 MOS Instrument Review Criteria. Med Outcomes Trust Bull, September, 1995;3:I–IV.
7 Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–6.
8 Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 1989;10: 407–15.
9 Erickson P, Scott J. The On-Line Guide to Quality-of-Life Assessment (OLGA): resource for selecting quality of life assessments. In: Walker S, Rosser RM, eds., Quality of Life Assessment: Key Issues in the 1990s. Dordrecht: Kluwer Acedemic, 1993.
10 Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11:570–9.
11 Aaronson NK, Ahmedzai S, Bergmen B, et al. The European Organisation for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85: 365–76.
12 Keinanen-Kiukaanniemi S, Ohinmaa A, Pajunpaa H, et al. Health-related quality of life in diabetic patients measured by the Nottingham Health Profile. Diabet Med 1996;13:382–8.
13 Anderson RM, Fitzgerald JT, Wisdom K, et al. A comparison of global versus disease-specific quality of life measures in patients with NIDDM. Diabetes Care 1997;20:299–305.
14 Sprangers MAG, Moinpour CM, Moynihan TJ, et al. Assessing meaningful change in quality of life over time: a users' guide for clinicians. Mayo Clin Proc 2002;77:561–71.
15 Fayers P, Machin D. Missing data. In: Fayers PM, ed., Quality of Life: Assessment, Analysis, and Interpretation. West Sussex: Wiley, 2000. p. 224–46.
16 Fayers P, Curran D, Machin D. Incomplete quality of life data in randomized trials: missing items. Stat Med 1998;17:679–96.
17 Fairclough D, Cella D. Functional assessment of cancer therapy (FACT-G): non-response to individual questions. Qual Life Res 1996;5:321–9.
18 Ware JE. Interpreting general health measures. In: Spilker B, ed. Quality of Life and Pharmacoeconomics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven, 1996.
19 Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures, statistics and strategies for evaluation. Control Clin Trials 1991;12(Suppl):S142–58.
20 Acquadro C, Jambon B, Ellis D, Marquis P. Language and translation issues. In: Spilker B, ed., Quality of Life and Pharmaco-Economics in Clinical Trials. 2nd ed. Philadelphia: Lippincott-Raven, 1995.
21 Guillemin F, Bombardier C, Beaton DE. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1994;47:1465–6.
22 Bullinger M, Alonso J, Apolone G, et al. Translating health status questionnaires and evaluating their quality: the IQOLA Project approach—international quality of life assessment. J Clin Epidemiol 1998;51:913–23.
23 Boyle P. Cultural and linguistic validation of questionnaires for use in international studies: the nine-item BPH-specific quality of life scale. Eur Urol 1997;32(Suppl 2):S50–2.
24 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:178–89.
25 King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res 1996;5:555–67.
26 Osoba D, Rodrigues G, Myles J, et al. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16:139–44.
27 Ware JE, Snow KK, Kosinski M, et al. SF-36 Health Survey: manual and interpretation guide. Boston: New England Medical Center, The Health Institute, 1993.
28 Nunnally JC, Bernstein IR. Psychometric Theory Boston. 3rd ed. New York: McGraw-Hill, 1994.
29 Ware JE, Kosinski M, Keller SD. SF-36 physical and mental component summary measures—a user's manual. Boston: The Health Institute, 1994.
30 Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992;305: 160–4.
31 Bergner L, Bergner M, Hallstrom AP, et al. Health status of survivors of out-of-hospital cardiac arrest six months later. Am J Public Health 1984;74: 508–10.
32 Hunt SM, McKenna SP, Williams J, et al. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med 1981;15a: 221–9.
33 Testa MA, Anderson RB, Nackley JF, et al. Quality of life and antihypertensive therapy in men: a comparison of captopril with enalapril. The Quality of

Life Hypertension Study Group. N Engl J Med 1993;328:907–13.
34 Jaeschke R, Guyatt GH, Keller J, et al. Ascertaining the meaning of change in quality of life questionnaire score: data from N of 1 randomized control trials. Control Clin Trials 1991;12(Suppl): S226–33.
35 Guyatt GH, Berman LB, Townsend M, et al. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987;42:773–8.
36 Guyatt GH, Nogradi S, Halcrow S, et al. Development and testing of a new measure of health status for clinical trials in heart failure. J Gen Intern Med 1989;4:101–7.
37 Juniper EF, Guyatt GH, Epstein RS, et al. Evaluation of impairment of health related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax 1992;47:76–83.
38 Juniper EF, Guyatt GH, Willan A, et al. Determining a minimal important change in a disease-specific quality of life questionnaire. J Clin Epidemiol 1994;47:81–7.
39 Jaeschke R, Guyatt GH, Keller J, et al. Ascertaining the meaning of change in quality of life questionnaire score: data from N of 1 randomized control trials. Control Clin Trials 1991;12(Suppl): S226–33.
40 Anderson RB. What does it mean? Anchoring psychosocial quality of life scale score changes with reference to concurrent changes on reported symptom distress. Drug Inf J 1999;33:445–53.
41 Wyrwich KW, Nienaber NA, Tierney W, et al. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 1999;37: 469–78.
42 Wyrwich KW, Tierney WM, Wolinsky FE. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health related quality of life. J Clin Epidemiol 1999;52:861–73.
43 Barry MJ, Fowler FJ, O'Leary MP, et al. The American Urological Association symptom index for benign prostatic hyperplasia. J Urol 1992;148: 1549–57.
44 Barry MJ, Williford WO, Chang Y, et al. Benign prostatic hyperplasia specific health status measures in clinical research: how much change in the American Urological Association symptom and the benign prostatic hyperplasia impact index is perceptible to patients? J Urol 1995;154:1770–4.
45 Arocho R, McMillan CA, Sutton-Wallace P. Construct validation of the USA–Spanish version of the SF-36 Health Survey in a Cuban-American population with benign prostatic hyperplasia. Qual Life Res 1998;7:121–6.
46 McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. Med Care 1994;32:551–67.
47 Gandek B, Ware JE. Translating functional health and well-being: international quality of life assessment (IQOLA) project studies of the SF-36 health survey. J Clin Epidemiol 1998;51:891–1214.
48 Lewin-Epstein N, Sagiv-Schifter T, Shabtai EL, Shmueli A. Validation of the 36-item short-form health survey (Hebrew version) in the adult population of Israel. Med Care 1998;36:1361–70.
49 Loge JH, Kaasa S. Short form 36 (SF-36) health survey: normative data from the general Norwegian population. Scand J Soc Med 1998;26:250–8.
50 Prieto L, Alonso J, Ferrer M, et al. Are results of the SF-36 health survey and the Nottingham Health Profile similar? A comparison in COPD patients. Quality of Life in COPD Study Group. J Clin Epidemiol 1997;50:463–73.
51 Bronfort G, Bouter LM. Responsiveness of general health status in chronic low pain: a comparison of the COOP charts and the SF-36. Pain 1999;83: 201–9.
52 Damiano AM. The Sickness Impact Profile: User's Manual and Interpretation Guide. Medical Outcomes Trust, Boston MA. Baltimore (MD): The Johns Hopkins University, 1996.
53 Stucki G, Liang MH, Fossel AH, et al. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369–78.
54 The European Group for Quality of Life and Health Measurement. European Guide to the Nottingham Health Profile. The European Group for Quality of Life and Health Measurement (eds). 1992.
55 Hjermstad MJ, Fayers PM, Bjordal K, et al. Health related quality of life in the Norwegian general population assessed by the European Organisation for Research and Treatment of Cancer core quality of life questionnaire: the QLQ-C30(+3). J Clin Oncol 1998;16:1188–96.
56 Klee M, Groenvold M, Machin D. Quality of life of Danish women: population-based norms of the EORTC QLQ-C30. Qual Life Res 1997;6: 27–34.
57 Groenvold M, Klee MC, Sprangers MA, et al. Validation of the EORTC QLQ-C30 quality of life questionnaire through combined qualitative and quantitative assessment of patient–observer agreement. J Clin Epidemiol 1997;50:441–50.
58 Overcash J, Extermann M, Parr J, et al. Validity and reliability of the FACT-G scale for use in the older person with cancer. Am J Clin Oncol 2001;24:591–6.
59 Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–21.

60 Cella D. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System. Evanston (IL): Evanston Northwestern Healthcare and Northwestern University, 1997.
61 Cella D, Eton DT, Fairclough DL, et al. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol 2002;55:285–95.
62 Barber BL, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QoL questionnaire. Qual Life Res 1996;5:117–22.
63 Samsa G, Edelman D, Rothman M, et al. Determining clinically important differences in health status measures. Pharmacoeconomics 1999;15: 141–55.
64 Wright JG. The minimal important difference: who's to say what is important? J Clin Epidemiol 1996;49:1221–2.
65 Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lessons of Cronbach. J Clin Epidemiol 1997;50:869.
66 Hays RD, Woolley JM. The concepts of clinically meaningful difference in health-related quality-of-life research: how meaningful is it? Pharmacoeconomics 2000;18:419–23.
67 Donald P, Gagnon D, Zagari M. Assessing the clinical significance of changes in health-related quality of life (HRQOL) scores. In: Abstract Issue 7th Annual Conference of the International Society of Qual Life Res 2000;9:275.
68 Jones PW, Quirk FH, Baveystock CM. The St George's Respiratory Questionnaire. Respir Med 1991;85(Suppl B):S25–31.
69 Wells GA, Tugwell P, Kraag GR, et al. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557–60.
70 Rose MS, Koshman ML, Spreng S, Sheldon R. Statistical issues encountered in the comparison of health-related quality of life in diseased patients to published general population norms: problems and solutions. J Clin Epidemiol 1999;52:405–12.
71 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318:1728–33.
72 Guyatt GH, Juniper EF, Walter SD, et al. Interpreting treatment effects in randomised clinical trials. BMJ 1998;316:690–3.
73 Symonds T, Berzon R, Marquis P, et al. The clinical significance of quality-of-life results: practical considerations for specific audiences. Mayo Clinic Proc 2002;77:572–83.