SPECIAL ARTICLE

Precision of Health-Related Quality-of-Life Data Compared With Other Clinical Measures

ELIZABETH A. HAHN, MA; DAVID CELLA, PHD; OLIVIER CHASSANY, MD, PHD; DIANE L. FAIRCLOUGH, DRPH; GILBERT Y. WONG, MD; RON D. HAYS, PHD; AND THE CLINICAL SIGNIFICANCE CONSENSUS MEETING GROUP

To many clinicians, the assessment of health-related quality of life (HRQL) seems more art than science. This belief is due in part to the lack of formal training available to clinicians regarding HRQL measurement and interpretation. When HRQL is used systematically, it has been shown to improve patient-physician communication, clinical decision making, and satisfaction with care. Nevertheless, clinicians rarely use formal HRQL data in their practices. One major reason is unfamiliarity with the interpretation and potential utility of the data. This unfamiliarity causes a lack of appreciation for the reliability of data generated by formal HRQL assessment and a tendency to regard HRQL data as having insufficient precision for individual use. This article discusses HRQL in the larger context of health indicators and health outcome measurement and is targeted to the practicing clinician who has not had the opportunity to understand and use HRQL data. The concept and measurement of reliability are explained and applied to HRQL and common clinical measures simultaneously, and these results are compared with one another. By offering a juxtaposition of common medical measurements and their associated error with HRQL measurement error, we note that HRQL instruments are comparable with commonly used clinical data. We further discuss the necessary requirements for clinicians to adopt formal, routine HRQL assessment into their practices.

***Mayo Clin Proc.* 2007;82(10):1244-1254**

HIV = human immunodeficiency virus; HRQL = health-related quality of life; ICC = intraclass correlation coefficient; SEM = standard error of measurement

Measurement in medicine is not new; in fact, it has been an integral component of medical diagnosis and treatment since the beginning of the clinical practice of medicine. From the start of their schooling and training, clinicians are taught the utility of measurement, such as height, weight, vital signs, pathology reports, and laboratory chemistry values. Clinicians are also trained in obtaining the qualitative aspects of a medical history but are not routinely taught the quantitative measurement of patient-reported outcomes such as health-related quality of life (HRQL). As a result, HRQL measurement may seem more like an art than science to many clinicians. However, increasing evidence suggests that routine, formal assessment of HRQL improves care on multiple levels. For example, adding HRQL assessment to clinical practice has led to improved problem identification[2-11] and improved patient-physician communication.[12,13] Some studies have found significant increases in patient management activities designed to address the problems identified.[5,6,10,14-16] A few controlled studies of HRQL assessment vs standard care have demonstrated a positive impact on patient satisfaction or HRQL.[10,13,17-20]

Assessment of HRQL has been successfully used to change and influence patient and physician communication, resulting in improved patient satisfaction in a community practice setting.[21] The mechanisms by which routine assessment of HRQL might improve clinical practice include (1) aiding detection of physical or psychosocial problems that otherwise might be overlooked, (2) monitoring disease and treatment, (3) allowing precisely timed alterations in therapeutic plans, (4) facilitating patient-physician communication, and (5) improving the delivery of care.[18,19,22-29] It is also possible to routinely use HRQL instruments in clinical practice to evaluate the efficacy of interventions designed to prevent or treat common problems experienced by patients.[30] Several critical elements for the success of routine HRQL assessments have been identified.[15,26,31,32] The first is the availability of an acceptable set of measures from which to choose. These HRQL measures must be brief and simple to administer, complete, score, and interpret. The second critical factor involves clinical relevance and ease of use, ideally with results presented in a structured format that includes comparison (reference) data for these assessments.[33] The results and interpretation of HRQL information must be delivered in a manner that facilitates and guides interventions. Finally, buy-in from both clinic staff and patients is essential so that routine HRQL assessment can be effectively implemented.[31]

From the Department of Preventive Medicine (E.A.H.) and Department of Psychiatry and Behavioral Sciences (D.C.), Center on Outcomes, Research and Education, Evanston Northwestern Healthcare, Institute for Healthcare Studies, and Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine, Northwestern University, Evanston, IL; Département de la Recherche Clinique et du Développement de l'Assistance Publique-Hôpitaux de Paris, Hôpital Saint-Louis, Paris, France (O.C.); Colorado Health Outcomes, University of Colorado Health Sciences Center, Denver (D.L.F.); Division of Pain Medicine, Department of Anesthesiology, Mayo Clinic, Rochester, MN (G.Y.W.); and Departments of Medicine and Health Services Research, UCLA, Los Angeles. CA (R.D.H.). Dr Wong is now with Anesiva, Inc, South San Francisco, CA. A list of the members of the Clinical Significance Consensus Meeting Group appears in Frost et al.[1]

Individual reprints of this article are not available. Address correspondence to Elizabeth A. Hahn, MA, Center on Outcomes, Research and Education, Evanston Northwestern Healthcare, 1001 University Pl, Suite 100, Evanston, IL 60201 (e-hahn@northwestern.edu).

This article discusses HRQL in the larger context of health indicators and health outcome measurement. The discussion is targeted to the practicing clinician who has not had the opportunity to understand and use HRQL data. If one extends to HRQL assessment the observation that health care professionals not only tolerate but also depend on measurements inherently associated with error, whether from examination or laboratory findings, the future of HRQL assessment in clinical practice is bright. To support this statement, we offer a juxtaposition of common medical measurements and their associated error (when it has been studied), with HRQL measurement error and discuss what will likely be necessary for clinicians to adopt formal, routine HRQL assessment into their practices. Interested readers can find more detailed information about incorporating HRQL into practice in 2 other articles published in this issue.[1,34]

ORGANIZING MODEL OF PATIENT OUTCOMES

Although the quality of cancer care has traditionally been measured with such clinical outcomes as survival and tumor response, recognition of the importance of patient-reported outcomes is increasing.[35] During the course of disease and/or treatment, patients may experience many symptoms, including weight loss, fever, fatigue, and pain; treatment adverse effects, such as shortness of breath, fatigue, dizziness, hair loss, nausea, and pain; and challenges to their ability to cope with physical and emotional changes.[36,37] After completion of treatment, patients must contend with physical, emotional, and social problems related to the direct effects of the disease, consequences of treatment, and individual or family factors.[38,39] Physical problems may include organ dysfunction, infertility, second malignancies, and recurrence; social problems may include employability and insurability; emotional problems may stem from fears of recurrence, adjustment to physical limitations, loss of job flexibility, and posttreatment mood and stress disorders. Systematic attention should be directed to the full range of patient concerns to better address patients' needs both during and after treatment.[39-41]

Wilson and Cleary[42] proposed a conceptual model that linked clinical variables with HRQL. Another useful and complementary model has been proposed by Patrick,[43] which clarifies the source of data (eg, patient, observer) and its relationship to the HRQL outcome. Figure 1, an adaptation of these 2 models, displays at the top an essentially unidirectional (causal) pathway from biological and physiological processes, such as a disease or chronic condition, which often results in symptoms. Symptoms in turn can produce limitations in functional status and ability to engage in normal everyday activities. Over time, such an impact can have detrimental effects on patients' general views of their own health and even self-esteem or sense of personal value. All this can contribute to a decline in one's overall evaluation of quality of life. When evaluation of quality of life is limited to the context of health and illness, it is usually referred to as *health-related quality of life*.

Moving from left to right across the top of Figure 1, the strength of the association weakens. Thus, for example, biological and physiological variables, measured in numerous ways across medicine, can be expected to correlate most strongly with measures of symptoms and most weakly with overall HRQL (although there is a relationship across the expanse of the model). On this general model linking clinical and HRQL variables,[42] the source of the data can be overlaid (depicted in the middle of Figure 1), as has been suggested by Patrick.[43]

To be useful, instruments must be both reliable and valid. Reliability refers to an instrument's dependability, expressed as the extent to which it either measures something accurately or produces the same score on repeated applications. Validity refers to the extent to which an instrument measures what it proposes to measure. This article focuses on reliability of measurement. Using multiple clinical examples, we summarize relevant statistics, methods, and standards for assessing reliability and then describe the reliability of common physiological and self-report instruments.

METHODS

We conducted a literature review based on sources known to the authors, supplemented by a review of familiar sources in support of the effectiveness of formal practice-based HRQL assessment. We focused attention on clinical measurement studies that offered sufficient information about measurement error to allow comparison with error in HRQL measurement. To enable these comparisons, we recorded literature-based standards for high, moderate, and low reliability across commonly reported reliability (precision) statistics. We then categorized each of the selected clinical measurements and the reliability information into 1 or more of these predetermined classifications, according to the reviewed literature.

Definition of Measurement Error

Any measurement has some degree of error because of imperfect calibration of the measuring device, misunderstanding by the patient of a question, or the inherent lability of the characteristic.[44] The classic linear model for an observed value is $X = T + e$, in which X represents the observed value for a patient on some variable, T is the patient's true score, and e represents the difference between the true score and the observed score.[44,45]

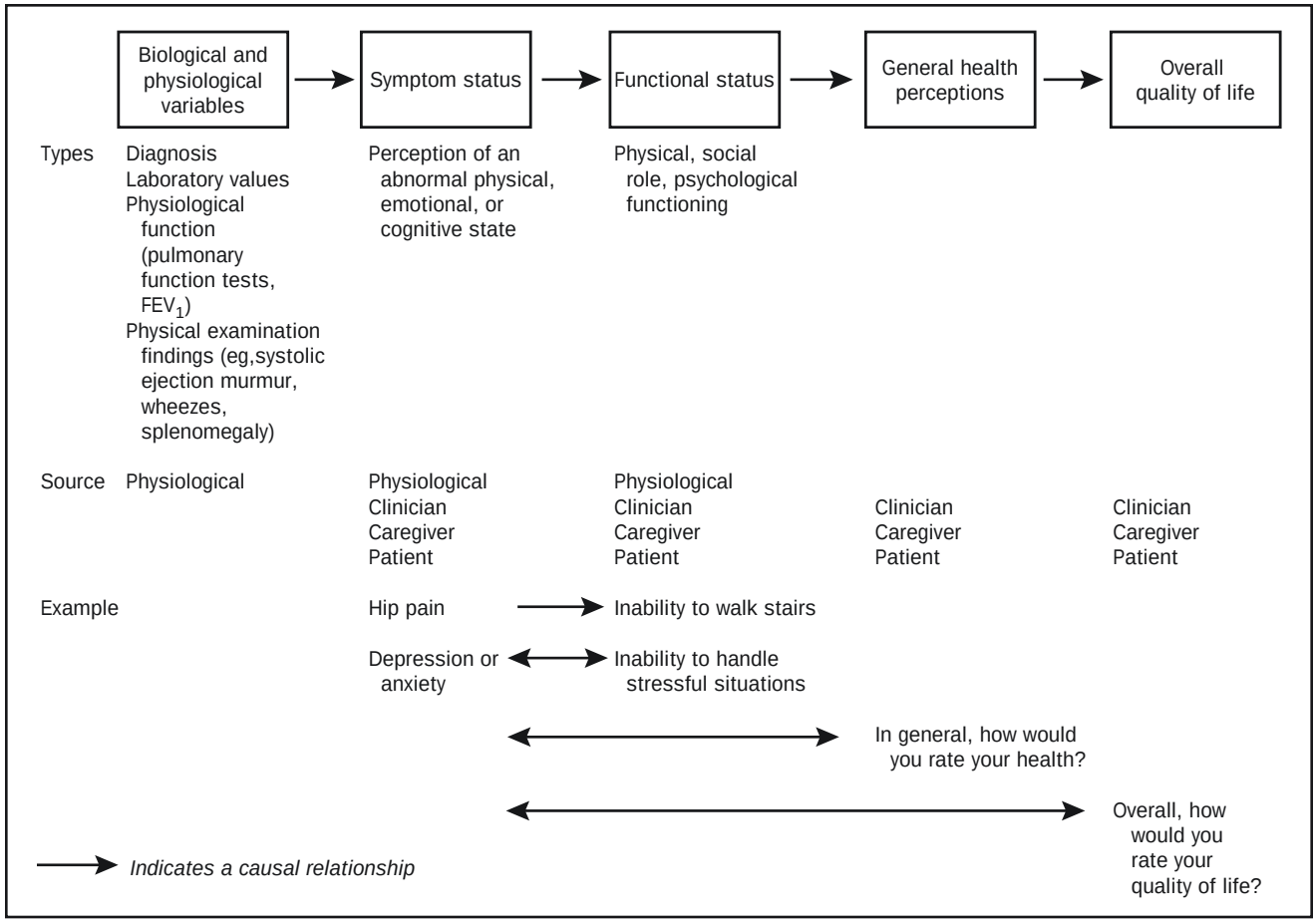

FIGURE 1. Measures of patient outcomes as organized by current models.[42,43] FEV_1 = forced expiratory volume in 1 second.

Measurement error can be due to systematic and/or random processes.[46] Systematic error might affect all observations equally or it might affect certain types of observations differently than others and be considered a type of bias. A miscalibrated thermometer that always records a temperature 3° higher than the actual measurement is an example of systematic error because it affects all observations. A thermometer reading affected by an object's color or density reflects a biased thermometer because characteristics not directly related to temperature are influencing it. Random temperature measurement error might be present if a person reading the thermometer occasionally transposed the digits. Biases and random errors are also present in outcome measurement.

The term *reliability* is sometimes defined as "freedom from random error."[46] It is used generically for 2 different characteristics of a measure: repeatability and internal consistency. The characteristic of repeatability can be measured over time (test-retest reliability), over observers (interrater reliability), or across different versions of an instrument (alternate forms reliability). A clinical analogy is the measurement of blood pressure, in which one might want to know the reliability of measures taken during a 24-hour period (test-retest) or by different health care professionals (interrater reliability). Internal consistency refers to the extent to which a set of questions measures a single underlying dimension, such as fatigue, depression, or physical functioning. It is analogous to measures of reliability of laboratory tests of replicate samples. The formulas used to estimate reliability for repeatability or internal consistency are equivalent. Table 1 provides 3 classification categories for reliability statistics (high, moderate, low) based on the type of data (nominal, ordinal, interval/ratio) and suggests ranges that we have found useful for classifying a measure or test in terms of reliability.

Creating a Common Ground: Reliability and Precision Statistics

Nominal and Ordinal Data. Recording the presence or absence of a symptom is an example of a nominal (named) variable; recording a symptom as none, moderate, or severe is an example of an ordinal (ordered) classification. The

TABLE 1. **Guidelines for Instrument Reliability and Precision***

Data type	Relevant statistic	Reliability: High or excellent (minimal or no error)	Reliability: Moderate or good (acceptable error)	Reliability: Low (high error)
Nominal	κ	>0.74	0.40-0.74	<0.40
Ordinal	Weighted κ	>0.74	0.40-0.74	<0.40
Interval/ratio	Intraclass correlation coefficient	>0.90	0.70-0.90	<0.70
	Pearson correlation coefficient (correlation with gold standard)	>0.90	0.70-0.90	<0.70
	Internal consistency reliability	>0.90	0.70-0.90	<0.70
	SEM	Intraindividual change <1 SEM suggests stability		

*SEM = standard error of measurement.

relevant statistic for estimating the reliability of a value on a nominal or ordinal scale is the κ or weighted κ, respectively.[47-50] A κ statistic quantifies the amount of agreement between 2 or more measurements that is greater than the amount expected by chance alone. The Kendall coefficient of concordance can also be used with ordinal data.[50] The repeated measurements may be over time, over observers, or over different forms of a test. If κ=0, there is just chance agreement; if κ<0, there is even less than chance agreement (a rare occurrence); if κ>0, there is greater than chance agreement; and if κ=1, there is perfect agreement. Table 1 lists some recommended criterion values for good (κ=0.40-0.74) and excellent (κ>0.74) agreement. Some authors have proposed even finer distinctions among levels of agreement, for example, 0.41 to 0.60 for moderate agreement, 0.61 to 0.80 for substantial agreement, and 0.81 to 1.00 for almost perfect agreement.[51]

Interval and Ratio Data. The most common units of measurement are based on interval or ratio scales, and several useful reliability statistics exist. Just as κ can be used to assess the 3 types of repeatability (test-retest, interrater, alternate forms) for nominal or ordinal data, the intraclass correlation coefficient (ICC)[52-54] assesses the 3 types for continuous measures. For nominal data, κ is mathematically equivalent to the ICC. For ordinal and interval data, weighted κ and the ICC are equivalent in certain conditions.[55] Numerous versions of ICCs are available; choosing the most appropriate one depends on several factors, including the types of raters and the types of patients.[56,57] In a simple example in which there is interest in assessing the test-retest reliability of an HRQL measure, a 1-way random-effects analysis of variance technique would be used. Recall the model for an observed value: $X = T + e$, in which T represents the patient's true score (also termed the *error-free score*, *steady-state value*, or *signal*[44]). In a population of patients, the T will vary around a mean value μ with a variance of σ_S^2, and the random error e has a variance of σ_E^2. The total variation in the scores can be partitioned into 2 parts: (1) variability among patients (σ_S^2) and (2) variability of the random errors σ_E^2. For this example, the ICC is defined as the ratio of the between-subject variance to the total variance: $r_{ICC} = \sigma_S^2 / (\sigma_S^2 + \sigma_E^2)$. Another way of thinking about this is the ratio of the variance of the true scores (σ_S^2) to the variance of the observed scores ($\sigma_S^2 + \sigma_E^2$). This ratio can range from 0 to 1. Values near 0 indicate that almost all the variation in score is due to measurement error and that the measure is unreliable. Values near 1 (>0.90 in Table 1) indicate that there is minimal measurement error and that the measure is very reliable.

An alternative to ICCs was proposed by Altman and Bland.[58-60] This approach involves plotting the differences of observed pairs of measurements against their mean values, creating limits of agreement (mean ± 2s, in which s is the SD of the differences), and examining trends using linear regression analysis. Although this more visual approach is often more easily understood by nonstatisticians, it relies on statistical significance tests of differences rather than on tests of consistency or equivalence, and it lacks a single measure that would be preferable, especially when more than 2 methods are compared.[56]

The Pearson correlation coefficient (r) is an estimate of the linear association between 2 interval/ratio variables. It is calculated using the SDs of the 2 variables (s_z and s_x) and their covariance (s_{zx}): $r = s_{zx}/s_z s_x$. The correlation coefficient can range from –1 to +1. Values near 0 indicate almost no linear association between the 2 variables, and values near –1 or +1 indicate that 1 variable can be almost perfectly predicted from the value of the other. Some investigators use r as a substitute for r_{ICC}, but these 2 coefficients yield different types of information and are not generally interchangeable.

Internal consistency reliability[46,61] has the same conceptual basis as the aforementioned stability (or repeatability) measures of reliability. Internal consistency can be interpreted as the ratio of the variance of the true values among patients to the variance of the observed values. If each patient completes a multi-item instrument or answers the same question on several occasions, the average of these observations should have higher reliability than a score based on a single answer. This occurs because the measure-

ment error is presumably random; when the values are averaged, the error is averaged out and thus decreases. Given that the error of the mean of k random values is $Var(\bar{e}) = \sigma_E^2/k$ then for a multi-item instrument, the reliability (r_α) of a k-item score is $\sigma_S^2 / \sigma_S^2 + \sigma_E^2/k)$. In laboratory studies and for multi-item instruments, the implication is that as the number of assessments is increased, the reliability will increase. Increasing the number of assessments (or questions) will have the greatest impact on the reliability of a test when each question has a large measurement error relative to the variation of the true values. One sees diminishing returns with increasing questions. Internal consistency reliability is most commonly assessed using the Cronbach coefficient α, and values greater than 0.90 are considered the standard for individual-level applications (Table 1).

The standard error of measurement (SEM)[44] is expressed on the same scale as the quantity being measured. The SEM is defined in terms of SD (σ_S) reliability (either r_α or r_{ICC}): $SEM = \sigma_S\sqrt{1 - R}$. If a measure has a reliability of 0.80 (common for many HRQL scales), the error of measurement associated with any individual score is 45% of the SD. If the reliability decreases to 0.50 (uncommon for most HRQL scales), then the SEM is 70% of the SD. One way to interpret this statistic is to note that we would expect a person's observed score to fall within the interval of ±1 SEM around a person's true score 68% of the time and in the interval of ±2 SEM 95% of the time. If the reliability is 0.80 and the SD is 10, the SEM is $10\sqrt{1 - 0.80}$ or 4.5. Thus, if the estimated true score was 60, we would expect that 68% of the time the observed score would fall within the interval of 60±4.5 or between 55.5 and 64.5. In a clinical setting, it is also of interest to know how big a difference one might expect if the person takes the same test on 2 occasions when his or her true value actually does not change. The SD of the difference of 2 scores is $\sqrt{2}\,\text{SEM}$. This can be used to estimate a confidence interval for the estimated true score. For example, if the true score for a patient was 62 on the first occasion and the patient is tested a second time without a change in true score, the probability is 68% that the second score will be in the interval $62 \pm \sqrt{2} \times 4.5$ or between 55.7 and 68.3. The SEM can be used to help interpret the meaningfulness of intrapatient change. Recent research has suggested that a change less than 1 SEM is rarely clinically meaningful (Table 1).[62,63]

Use of Reliability Statistics

A number of issues influence the evaluation of these reliability statistics. The most critical is how the information is to be used. If the measure is to be used in a patient management decision at the individual level, higher levels of reliability are required than for comparisons among groups of patients.[64] If the measure is being used as a screening tool to identify patients in need of additional assessment, the criteria for adequate reliability can be lowered. Another important feature is that this measure of reliability is closely linked to the population in which one wants to use the measure. Clinicians and researchers need to be aware of the characteristics of the sample used to assess the reliability of a test or measure. The more heterogeneous the population, the larger the differences between patients and thus σ_S^2. Thus, the reliability estimate will tend to be higher when there is a mixture of patients who would be expected to have values across the entire range of the measure as would occur when there are patients both with and without a condition. Finally, we need to be aware of the type of reliability that was measured and whether it is appropriate to our study setting. Are we interested in whether a measurement at one point in time will agree with one taken later or whether the patient's assessment will agree with the caretaker's assessment? The use of item response theory models is contributing to additional advances in the reliability of HRQL measurement.[65]

RESULTS AND DISCUSSION

Figure 1 illustrates an organizing model for this report, wherein it is hypothesized that the link between clinical and HRQL variables is stronger for self-reported disease symptoms than for more general health perceptions. The overlay of the source of data onto this model helps to clarify that patient-rated data, although paramount given the definition of HRQL as patient focused, are not the only sources of information regarding patient status. An example of the model at work can be seen in the case of cystic fibrosis, as depicted in Figure 2. In cystic fibrosis, varying degrees of association are observed across related measurements, from physiological to clinician- and caregiver-reported patient status to various types of patient-reported outcomes. The strongest association with patient HRQL is self-reported dyspnea, whereas the weakest (but still significant) association is with physiological variables (Figure 2).[66]

Table 2 reports data from selected studies of the reliability (degree to which error is reduced) of common clinical and HRQL measurements, according to the criteria outlined in Table 1.[67-78] Designed to be representative of common health measurements, rather than comprehensive, Table 2 provides a range of reliability estimates within each measurement category (see also Figure 1). For example, the reproducibility of vital sign measurements spans all 3 columns in Table 2: from high reproducibility for the classification of tachycardia, bradycardia, and systolic hypertension to low reproducibility for systolic hypotension.[67] Similarly, the reliability of commonly used HRQL mea-

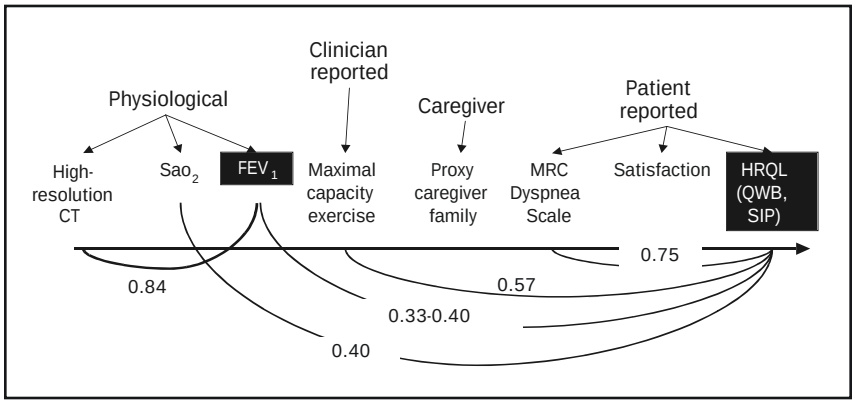

FIGURE 2. Proximal vs distal associations of clinical and health-related quality-of-life (HRQL) variables in cystic fibrosis based on a review of the literature.[66] Values correspond to correlation coefficients (*r*), which have been collected using different publications. CT = computed tomography; FEV_1 = forced expiratory volume in 1 second; Sao_2 = arterial oxygen saturation; MRC = Medical Research Council; QWB = quality of well-being (generic quality-of-life questionnaire); SIP = sickness impact profile (generic quality-of-life questionnaire).

sures varies across questionnaires and across subscales within a questionnaire (eg, 36-Item Short-Form Health Survey physical functioning vs role-physical).[78] As mentioned previously, awareness of the characteristics of the sample used to assess the reliability of each measure is important. The information in Table 2 should be used only as an overall summary of the possible range of reliability.

Association Between Patient-Reported HRQL and Biological and Physiological Measurements

The criteria generally used to measure the activity of a disease, such as a biological value, a physiological performance, or a radiographic image, do not by themselves reflect the perceptions and subjective state of the patient. Two patients with an identical biological value or physiological score may experience a different impact on their perceptions of symptoms or HRQL. For the same patient, a physical performance objectively assessed in a laboratory is not necessarily similar to the physical ability of the patient in everyday life.[79,80] For many conditions, correlation levels reported in the literature between a physical measurement of performance or a functional capacity (eg, forced expiratory volume in asthma or chronic obstructive pulmonary disease) and the measurement of the severity of symptoms or the physical dimension of HRQL seldom exceed 0.40 and are generally lower than 0.20.[81-83] Hemoglobin level, although related directly to oxygenation and therefore energy, also rarely correlates with self-reported fatigue and function beyond r=0.40, suggesting less than 15% shared variability in these conceptually linked measurements.[84,85] Another example is among patients with peripheral arterial occlusive claudication, in which the correlation between hemodynamic parameters and angiogram score vs self-reported functional disability and HRQL is low.[86,87] Still another example in osteoarthritis showed substantial discordance among radiographic osteoarthritis, physician-based diagnosis, and patient-reported pain.[88] The growing body of evidence linking patient-reported outcomes to clinical indicators suggests that although there is some common ground, there is even more uniqueness to the 2 types of information, and both have value. Across self-report and clinical measurements alike (Table 2), some of the lack of agreement is due to measurement error. Since there is clearly error in both clinical and self-report data and they converge only modestly in most cases, we suggest that self-report information is necessary to complete an accurate understanding of a patient's current HRQL.

Association Between Patient- and Physician-Assessed HRQL

Patient-reported outcomes provide additional information on treatment effects and patient perceptions that are not adequately captured by objective criteria and clinician-reported outcomes. By definition, HRQL is subjective. Therefore, patients are the best source to rate their own HRQL or perceived health and well-being. Patients' ratings of their experiences of disease or treatment often differ in both degree and type from those of health care professionals.[89-91] Furthermore, in some conditions, such as cancer, chronic heart failure, chronic obstructive pulmonary disease, or rheumatoid arthritis, baseline HRQL scores (especially physical domains) predict survival.[92-100] This predictive value has been recently extended to show that, in addition to the prognostic value of baseline patient self-report data, change over time forecasts outcome in advanced lung cancer.[101] The predictive validity of psychological and emotional function is less clear.[102,103]

TABLE 2. **Degree of Error in Common Health Measurements***

Measures		
High reliability (minimal error)	Good reliability (acceptable error)	Low reliability (high error)
Classification of vital signs[67] Tachycardia, κ=0.85 Bradycardia, κ=0.87 Systolic hypertension, κ=0.75	Classification of vital signs[67] Tachypnea, κ=0.60 Diastolic hypertension, κ=0.74	Classification of vital signs[67] Systolic hypotension, κ=0.27
Anthropometric and muscle performance factors[68] Height, r_{ICC}=1.00 Weight, r_{ICC}=0.99 Neck length, r_{ICC}=0.98-0.99 Neck circumference, r_{ICC}=0.99 Lateral flexion, extension of cervical spine, cervical short flexor endurance, r_{ICC}=0.96 each Extensor strength, κ=0.78 Flexor strength, κ=0.86	Frequency of headaches, questionnaire vs diary, intrasubject κ=0.66[68]	
Survival time	Tumor response classification using chest radiograph, mean interobserver κ=0.74 (range, 0.51-1.00)[69]	Tumor mass diameter, mean interrater difference of 17%[70]
Computed tomography for advanced ovarian cancer[71] Intrarater κ=0.77-0.84 Interrater κ=0.76-0.79	Computed tomography for advanced ovarian cancer[71] Intrarater κ=0.52-0.71 Interrater κ=0.47-0.61	Computed tomography for advanced ovarian cancer[71] Interrater κ=0.36
		Tumor mass diameter, 1.3%-18.4% misclassification of patient response[72]
		Tumor size measurements over time (response and progression)[73] Intraobserver misclassification rates of 3%-14% of tumors Interobserver misclassification rates of 10%-43% of tumors
		Time to tumor progression
Bedside blood glucose screening (One Touch II method) and laboratory serum glucose, r=0.92[74]	Bedside blood glucose screening (Chemstrip bG method) and laboratory serum glucose, r=0.87[74]	
Motor output: spinal mechanisms for gating motoneuron excitability[75] Peak-to-peak amplitude of the soleus H-reflex, overall trial-to-trial reliability, r_{ICC}=0.93-0.97	Motor output: spinal mechanisms for gating motoneuron excitability[75] Peak-to-peak amplitude of the soleus H-reflex, selected conditions, r_{ICC}=0.70-0.90	
	Systolic blood pressure, test-retest r=0.81[76]	Diastolic blood pressure, test-retest r=0.63 Heart rate, test-retest r=0.68[76]
Postbronchodilator FEV_1, r_{ICC}=0.95 at 2 wk[77]	Postbronchodilator $FEF_{25\%-75\%}$, r_{ICC}=0.89 at 2 wk Bronchodilator response, r_{ICC}=0.70[77]	
SF-36[78] Physical functioning, r_α=0.93	SF-36[78] Physical functioning, r_{ICC}=0.81 Role-physical, r_α=0.84 Pain, r_α=0.82, r_{ICC}=0.78 General health perceptions, r_α=0.78, r_{ICC}=0.80 Vitality, r_α=0.87, r_{ICC}=0.80 Social functioning, r_α=0.85 Role-emotional, r_α=0.83 Mental health, r_α=0.90, r_{ICC}=0.75	SF-36[78] Role-physical, r_{ICC}=0.69 Social functioning, r_{ICC}=0.60 Role-emotional, r_{ICC}=0.63
	DUKE[78] General health, r_α=0.78	DUKE[78] Physical health, r_α=0.67 Social health, r_α=0.55 Mental health, r_α=0.68 Self-esteem, r_α=0.64 Anxiety, r_α=0.60 Depression, r_α=0.65
	FSQ[78] ADLs, r_α=0.79 IADLs, r_α=0.82 Mental health, r_α=0.81	FSQ[78] Social activity, r_α=0.65 Work performance, r_α=0.65 Quality of interaction, r_α=0.64
	NHP[78] Pain, r_α=0.72 Emotional reactions, r_α=0.81	NHP[78] Physical mobility, r_α=0.39 Energy, r_α=0.57 Social isolation, r_α=0.34 Sleep, r_α=0.68

*ADLs = activities of daily living; DUKE = Duke Health Profile; $FEF_{25\%-75\%}$ = forced expiratory flow between 25% and 75%; FEV_1 = forced expiratory volume in 1 second; FSQ = Functional Status Questionnaire; IADLs = instrumental activities of daily living; NHP = Nottingham Health Profile; r = Pearson correlation coefficient; r_α = internal consistency reliability; r_{ICC} = intraclass correlation coefficient; SF-36 = 36-Item Short-Form Health Survey.

Association Among Different Patient-Reported Outcomes

Even though in a given disease a logical association exists between the severity of the symptoms and a worsening perception of HRQL by the patient (Figure 1), there are situations in which the measurement of the symptoms does not reflect the subjective real life of the patient. For example, irritable bowel syndrome is a functional and benign disease, but the long-term course is composed of symptomatic flares that significantly affect health perceptions.[104] The absence of pain or abdominal discomfort at a given time (eg, during a consultation with the physician) is not synonymous with a good HRQL score. The patient may be anxious to know when the next symptomatic bout will occur, may be limited in social activities, or may be constrained by having to take drugs and pay attention to food. The fear or the forecast of the crisis is possibly a handicap more significant than the crisis itself. Thus, the clinician cannot infer all aspects of HRQL.

Reliability Classification of Physiological and Self-report Measurements

Both HRQL and other patient-reported outcomes are sometimes labeled as subjective by clinicians because they are based on individual perceptions. However, we argue that the distinction between subjective and objective should not depend on *who* makes the rating; in other words, a measurement is not considered objective just because it is made by a clinician.[105] In fact, ratings of patient performance or other aspects of well-being by clinicians are often discordant with the self-ratings provided by patients, leading one to question the objectivity of the clinician rater. Even so-called objective morphological measures, such as tumor size or change, can lack reliability when a subjective observer must interpret results. More than a quarter century ago, Moertel and Hanley[70] pointed out the high unreliability of tumor measurements in an experiment using simulated tumors and palpation. Warde et al[71] illustrated the same unreliability of tumor measurement using computed tomography in ovarian cancer. More recently, Erasmus et al[73] evaluated the readings of 40 radiographs of lung tumors by 5 radiologists. They showed intraobserver misclassification rates of 3% to 14% of tumors using Response Evaluation Criteria in Solid Tumors and World Health Organization criteria for response and interobserver misclassification of 10% to 43% of tumors using these criteria for progression. This degree of misclassification places tumor response, long held to be objective, in the low reliability column of Table 2.

The considerably higher misclassification rate of progression compared with response raises concern regarding the recent Food and Drug Administration[106] shift to time to progression as a primary surrogate end point, unless such an end point can be shown to be associated with improved HRQL or survival.

Studies not dealing with oncology, such as in classification of fractures, have also reported poor to moderate intraobserver and interobserver agreement.[107,108] Similarly, the knowledge of treatment assignment may influence decisions regarding drug dose adjustments even when objective rules exist for those adjustments.[109]

Examples in Which HRQL Data Utilization Is (or Could Be) Common Practice

Several examples across medical practice indicate situations in which patient-reported information is used as the primary source of decision making.

Pain Management. Chest pain is the key symptom used to initiate work-up and diagnosis of acute myocardial infarction.[110] Self-reported pain is used to titrate analgesic medication and to determine the potential utility of intraspinal opioids among patients with cancer and others with acute and chronic pain.[111-113] These studies have subsequently led to the widespread use of both intrathecal and epidural opioids for pain management and relief of pain in numerous settings, including cancer, obstetrical labor and delivery, postsurgical management, and other acute and chronic pain syndromes.

Asthma. In asthma treatment, patient-reported outcomes are primary indicators of disease status and progress. Therefore, in a recent asthma clinical trial that compared combination fluticasone propionate and salmeterol with placebo, the Asthma Quality of Life Questionnaire was the primary trial end point. Improvement from baseline to end point (12 weeks) was greater in the combination group than in the placebo group. The differences between groups exceeded the prespecified minimally important difference (0.5 on a scale ranging from 1-7) for the 4 dimensions and for the global score.[114]

Self-report of Symptoms and Adverse Events in Human Immunodeficiency Virus Disease. The perception of patients with human immunodeficiency virus (HIV) regarding the symptoms related to multidrug antiretroviral therapies may differ from the perception of clinicians. In a recent validation study, when compared with open-ended clinician interviews, the 20-item self-reported HIV symptom index captured more frequent and bothersome symptoms.[115] Similarly, in a previous study of more than 800 patients with HIV, patient and clinician agreement was poor (mean κ=0.14; range, 0.07-0.25).[116] Compared with self-report, clinicians underreported the presence and severity of symptoms. Reports by clinicians demonstrated greater variability by site and poorer test-retest reliability. Clinician-reported severity scores were less strongly associated than were self-reports of functional status, global

quality of life, and survival. Thus, the perception of patients with HIV about their symptoms may be more informative than that of clinicians. Finally, a discrepancy exists between clinician-based diagnosis and self-report of depression among patients with AIDS.[117]

Self-report in Functional Gastrointestinal Disorders. Discrepancies between physician and patient responses, by either standardized information (eg, Rome II criteria) or self-report,[118] have been described in diagnosing functional gastrointestinal disorders such as irritable bowel syndrome, diarrhea, and constipation. For example, the rate of constipation estimated across a population may differ, depending on whether the estimate is based on a definition of frequency (<3 stools per week) or self-perception.[119] Some patients who routinely have fewer than 3 stools per week may not feel constipated, and conversely, some patients may feel very constipated if they do not have 1 stool per day, although they may be considered healthy by a clinician.[120] This discrepancy in criteria raises the question of which is the more appropriate source, especially for these functional disorders: the physician diagnosis based on norms or the perception of the patient and the impact on his or her satisfaction, well-being, and HRQL.

Self-report in Rheumatoid Arthritis. For rheumatoid arthritis, accepted disease severity indicators are based on clinical examination of joints and functional assessment, as well as patient self-report of symptom severity and impact on functioning. Patient-reported outcomes tend to be highly correlated with findings on clinical examination and with the familiar American College of Rheumatology criteria of 20%, 50%, and 80% improvement.[121,122] Thus, in nonresearch settings it may be more efficient (and more relevant to the patient) to use self-report in place of examination results.

CONCLUSION

The practice of medicine is art as much as science. Clinicians in daily practice depend on physical examination–based, laboratory, radiographic, and other measurements to assess and care for patients. Rarely do they use formal assessment of patient-reported outcomes as part of routine clinical practice. Given the importance of HRQL assessments in the lives of patients with chronic conditions such as cancer, one questions this underuse. One major barrier to routine use of HRQL instruments in clinical practice is the perception that they are not sufficiently reliable or trustworthy to make individual diagnosis and treatment decisions. This perception has been perpetuated by test developers themselves, many of whom are measurement scientists trained to focus on error and precision at times over meaning and usefulness.

The purpose of this article is to discuss the larger context of health indicators, including routine clinical measurements used every day in practice. Through this exercise, it can be seen that reliability of measurement varies for patient-reported outcomes and clinical measurements, such as blood pressure, heart rate, tumor measurement, or carotid wall thickness using ultrasonography. If one were to relax the requirements placed on use of patient-reported HRQL for use in clinical practice to a level comparable to other measurements used in clinical care, the fidelity of HRQL assessments would compare favorably.

Given the importance of HRQL to people with chronic diseases, the advent of computer-assisted assessment, and the emergence of electronic patient records, we suggest it is time to convert practice behavior to routine HRQL monitoring as a way to promote excellence in patient health care. Future research should focus on overcoming technical and system barriers to such a conversion, determining optimal ways to complement clinical and physiological data with self-report data, evaluating the efficacy of routine monitoring in clinical practice, and determining the cost-effectiveness of routine monitoring in chronic illness care.

We thank Amy Eisenstein for her assistance in researching details in Table 2.

REFERENCES

1. Frost MH, Bonomi AE, Cappelleri JC, Schünemann HJ, Moynihan TJ, Aaronson NK, Clinical Significance Consensus Meeting Group. Applying quality-of-life data formally and systematically into clinical practice. *Mayo Clin Proc*. 2007; 82(10):1214-1228.

2. German PS, Shapiro SH, Skinner EA, et al. Detection and management of mental health problems of older patients by primary care providers. *JAMA*. 1987;257(4):489-493.

3. Goldsmith G, Brodwick M. Assessing the functional status of older patients with chronic illness. *Fam Med*. 1989;21(1):38-41.

4. Linn LS, Yager J. The effect of screening, sensitization, and feedback on notation of depression. *J Med Ed*. 1980;55(11):942-949.

5. Magruder-Habib K, Zung WW, Feussner JR. Improving physicians' recognition and treatment of depression in general medical care: results from a randomized clinical trial. *Med Care*. 1990;28(3):239-250.

6. Mazonson PD, Mathias SD, Fifer SK, Buesching DP, Malek P, Patrick DL. The mental health patient profile: does it change primary care physicians' practice patterns? *J Am Board Fam Pract*. 1996;9(5):336-345.

7. McLachlan SA, Allenby A, Matthews J, et al. Randomized trial of coordinated psychosocial interventions based on patient self-assessments versus standard care to improve the psychosocial functioning of patients with cancer. *J Clin Oncol*. 2001;19(21):4117-4125.

8. Moore AA, Siu A, Partridge JM, Hays RD, Adams J. A randomized trial of office-based screening for common problems in older persons. *Am J Med*. 1997;102(4):371-378.

9. Rand EH, Badger LW, Coggins DR. Toward a resolution of contradictions: utility of feedback from the GHQ. *Gen Hosp Psychiatry*. 1988; 10(3):189-196.

10. Rubenstein LV, McCoy JM, Cope DW, et al. Improving patient quality of life with feedback to physicians about functional status. *J Gen Intern Med*. 1995;10(11):607-614.

11. Zung WW, Magill M, Moore JT, George DT. Recognition and treatment of depression in a family medicine practice. *J Clin Psychiatry*. 1983; 44(1):3-6.

12. Detmar SB, Muller MJ, Schornagel JH, Wever LD, Aaronson NK. Health-related quality-of-life assessments and patient-physician communication: a randomized controlled trial [published correction appears in *JAMA* 2003; 289(13):987]. *JAMA*. 2002;288(23):3027-3034.

13. Wright EP, Selby PJ, Crawford M, et al. Feasibility and compliance of automated measurement of quality of life in oncology practice. *J Clin Oncol*. 2003;21(2):374-382.

14. Reifler DR, Kessler HS, Bernhard EJ, Leon AC, Martin GJ. Impact of screening for mental health concerns on health service utilization and functional status in primary care patients. *Arch Intern Med.* 1996;156(22):2593-2599.

15. Wasson J, Keller A, Rubenstein L, Hays R, Nelson E, Johnson D. Benefits and obstacles of health status assessment in ambulatory settings: the clinician's point of view: the Dartmouth Primary Care COOP Project. *Med Care*. 1992;30(5, suppl):MS42-MS49.

16. Detmar SB, Muller MJ, Schornagel JH, Wever LD, Aaronson NK. Role of health-related quality of life in palliative chemotherapy treatment decisions. *J Clin Oncol*. 2002;20(4):1056-1062.

17. Brody DS, Lerman C, Wolfson HG, Caputo GC. Improvement in physicians' counseling of patients with mental health problems. *Arch Intern Med*. 1990;150(5):993-998.

18. Greenhalgh J, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *J Eval Clin Pract*. 1999;5(4):401-416.

19. Espallargues M, Valderas JM, Alonso J. Provision of feedback on perceived health status to health care professionals: a systematic review of its impact. *Med Care*. 2000;38(2):175-186.

20. Velikova G, Booth L, Smith AB, et al. Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial. *J Clin Oncol*. 2004;22(4):714-724.

21. Wasson JH, Stukel TA, Weiss JE, Hays RD, Jette AM, Nelson EC. A randomized trial of the use of patient-self assessment data to improve community practices. *Eff Clin Pract*. 1999;2(1):1-10.

22. Albrecht G. Using subjective health assessments in practice and policy-making. *Health Care Anal*. 1996;4(4):284-292.

23. Carlson LE, Speca M, Hagen N, et al. Computerized quality-of-life screening in a cancer pain clinic. *J Palliat Care*. 2001 Spring;17(1):46-52.

24. Kazis LE, Callahan LF, Meenan RF, Pincus T. Health status reports in the care of patients with rheumatoid arthritis. *J Clin Epidemiol*. 1990;43(11):1243-1253.

25. Kerr J, Engle J, Schlesinger-Raab A, Sauer H, Holzel D. Communication, quality of life, and age: results of a 5-year prospective study in breast cancer patients [published correction appears in *Ann Oncol.* 2003;14(6):967]. *Ann Oncol*. 2003;14(3):421-427.

26. Lansky D, Butler JB, Waller FT. Using health status measures in the hospital setting: from acute care to outcomes management. *Med Care*. 1992;30(5, suppl):MS57-MS73.

27. Lohr KN. Applications of health status assessment measures in clinical practice: overview of the third conference on advances in health status assessment. *Med Care*. 1992;30(5, suppl):MS1-MS14.

28. McHorney CA. Health status assessment methods for adults: past accomplishments and future challenges. *Ann Rev Public Health*. 1999;20:309-335.

29. Till JE. Measuring quality of life: apparent benefits, potential concerns. *Can J Oncol*. 1994;4(1):243-248.

30. Jacobsen PB, Davis K, Cella D. Assessing quality of life in research and clinical practice. *Oncology (Williston Park)*. 2002;16(9, suppl 10):133-139.

31. Davis KM, Cella D. Assessing quality of life in oncology clinical practice: a review of barriers and critical success factors. *J Clin Outcomes Manage*. 2002;9(6):327-332.

32. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. *Med Care*. 1992; 30(5, suppl):MS23-MS41.

33. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care*. 2000;38(9, suppl):II84-II90.

34. Guyatt GH, Ferrans CE, Halyard MY, et al, Clinical Significance Consensus Meeting Group. An exploration of the value of health-related quality-of-life information from clinical research and into clinical practice. *Mayo Clin Proc*. 2007;82(10):1229-1239.

35. Apolone G. Clinical and outcome research in oncology: the need for integration. *Health Qual Life Outcomes*. 2003;1:1-6.

36. Cella DF, Bonomi AE. Measuring quality of life:1995 update. *Oncology (Williston Park)*. 1995;9(11, suppl):47-60.

37. Tranmer JE, Heyland D, Dudgeon D, Groll D, Squires-Graham M, Couslon K. Measuring the symptom experience of seriously ill cancer and non-cancer hospitalized patients near the end of life with the memorial symptom assessment scale. *J Pain Symptom Manage*. 2003;25(5):420-429.

38. Cordoba CS, Fobair P, Callan DB. Common issues facing adults with cancer. In: Stearns NM, Lauria MM, Hermann JF, Fogelberg PR, eds. *Oncology Social Work: A Clinician's Guide.* Atlanta, GA: American Cancer Society; 1993:43-77.

39. Skeel RT. Quality of life dimensions that are most important to cancer patients. *Oncology (Williston Park)*. 1993;7(12):55-61.

40. Loescher LJ, Welch-McCaffrey D, Leigh SA, Hoffman B, Meyerson FL Jr. Surviving adult cancers: part 1: physiologic effects. *Ann Intern Med*. 1989;111(6):411-432.

41. Welch-McCaffrey D, Hoffman B, Leigh SA, Loescher LJ, Meyskens FL Jr. Surviving adult cancers: part 2: psychosocial implications. *Ann Intern Med*. 1989;111(6):517-524.

42. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA*. 1995; 273(1):59-65.

43. Patrick D. Concept of health-related quality of life and of patient-reported outcomes. In: Chassany O, Caulin C, eds. *Health-Related Quality of Life and Patient-Reported Outcomes: Scientific and Useful Outcome Criteria.* Paris, France: Springer-Verlag; 2003:23-34.

44. Fleiss JL. *Design and Analysis of Clinical Experiments.* New York, NY: John Wiley & Sons; 1999.

45. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley; 1968.

46. Nunnally JC, Bernstein IH. *Psychometric Theory.* 3rd ed. New York, NY: McGraw-Hill; 1994.

47. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(3):37-46.

48. Spitzer RL, Cohen J, Fleiss JL, Endicott J. Quantification of agreement in psychiatric diagnosis: a new approach. *Arch Gen Psychiatry*. 1967;17(1):83-87.

49. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70:213-220.

50. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York, NY: Wiley; 1981.

51. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

52. Ebel RL. Estimation of the reliability of ratings. *Psychometrika*. 1951; 16:407-424.

53. Haggard EA. *Intraclass Correlation and the Analysis of Variance.* New York, NY: Dryden Press; 1958.

54. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.

55. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Measurement*. 1973;33:613-619.

56. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. *Stat Med*. 1994;13(23-24):2465-2476.

57. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use.* New York, NY: Oxford University Press; 1995.

58. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician*. 1983;32:307-317.

59. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet*. 1986;1(8476):307-310.

60. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med*. 1990;20(5):337-340.

61. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.

62. Wyrwich KW, Neinaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care*. 1999;37(5):469-478.

63. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol*. 1999;52(9):861-873.

64. Cella D, Bullinger M, Scott C, Barofsky I, Clinical Significance Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc*. 2002;77(4):384-392.

65. Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. *Med Care*. 2000;38(9 suppl):II66-II72.

66. Chassany O. De la maladie chronique à la qualité de vie. Méthodes d'évaluation. *Rev Mal Respir*. 2003;20:S34-S37.

67. Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. *Ann Emerg Med*. 2002;39(3):233-237.

68. Blizzard L, Grimmer KA, Dwyer T. Validity of a measure of the frequency of headaches with overt neck involvement, and reliability of measurement of cervical spine anthropometric and muscle performance factors. *Arch Phys Med Rehabil*. 2000;81(9):1204-1210.

69. Langendijk HA, Lamers RJS, ten Velde GPM, et al. Is the chest radiograph a reliable tool in the assessment of tumor response after radiotherapy in nonsmall cell lung carcinoma? *Int J Radiat Oncol Biol Phys*. 1998;41(5):1037-1045.

70. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38(1):388-394.

71. Warde P, Rideout DF, Herman S, et al. Computed tomography in advanced ovarian cancer: inter- and intraobserver reliability. *Invest Radiol*. 1986;21(1):31-33.

72. Warr D, McKinney S, Tannock I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. *J Clin Oncol*. 1984;2(9):1040-1046.

73. Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol*. 2003;21(13):2574-2582.

74. Martin S, Jensen R, Daly L, Jergenson C, Johnson MB, Buell T. Comparison of two methods of bedside blood glucose screening in the NICU: evaluation of accuracy and reliability. *Neonatal Netw*. 1997;16(2):39-43.

75. Earles DR, Morris HH, Peng CYJ, Koceja DM. Assessment of motoneuron excitability using recurrent inhibition and paired reflex depression protocols: a test of reliability. *Electromyogr Clin Neurophysiol*. 2002;42(3): 159-166.

76. Gerin W, Christenfeld N, Pieper C, et al. The generalizability of cardiovascular responses across settings. *J Psychosom Res*. 1998;44(2):209-218.

77. Faul JL, Demers EA, Burke CM, Poulter LW. The reproducibility of repeat measures of airway inflammation in stable atopic asthma. *Am J Respir Crit Care Med*. 1999;160(5, pt 1):1457-1461.

78. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res*. 1995;4(4):293-307.

79. Wiklund I, Comerford MB, Dimenas E. The relationship between exercise tolerance and quality of life in angina pectoris. *Clin Cardiol*. 1991;14(3):204-208.

80. Reuben DB, Valle LA, Hays RD, Siu AL. Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. *J Am Geriatr Soc*. 1995;43(1):17-23.

81. Yohannes AM, Roomi J, Waters K, Connolly MJ. Quality of life in elderly patients with COPD: measurement and predictive factors. *Respir Med*. 1998;92(10):1231-1236.

82. Shingo S, Zhang J, Reiss TF. Correlation of airway obstruction and patient-reported endpoints in clinical studies. *Eur Respir J*. 2001;17(2):220-224.

83. Leidy NK, Schmier JK, Jones MK, Lloyd J, Rocchiccioli K. Evaluating symptoms in chronic obstructive pulmonary disease: validation of the Breathlessness, Cough and Sputum Scale. *Respir Med*. 2003;97(suppl A):S59-S70.

84. Cella D. The effects of anemia and anemia treatment on the quality of life of people with cancer. *Oncology (Williston Park)*. 2002;16(9, suppl 10):125-132.

85. Cella D, Dobrez D, Glaspy J. Control of cancer-related anemia with erythropoietic agents: a review of evidence for improved quality of life and clinical outcomes. *Ann Oncol*. 2003;14(4):511-519.

86. Muller-Buhl U, Kirchberger I, Wiesemann A. Relevance of claudication pain distance in patients with peripheral arterial occlusive disease. *Vasa*. 1999;28(1):25-29.

87. Muller-Buhl U, Engeser P, Klimm HD, Weisemann A. Quality of life and objective disease criteria in patients with intermittent claudication in general practice. *Fam Pract*. 2003;20(1):36-40.

88. Hannan MT, Felson DT, Pincus T. Analysis of the discordance between radiographic changes and knee pain in osteoarthritis of the knee. *J Rheumatol*. 2000;27(6):1513-1517.

89. Kwoh CK, Ibrahim SA. Rheumatology patient and physician concordance with respect to important health and symptom status outcomes. *Arthritis Rheum*. 2001;45(4):372-377.

90. Edwards SGM, Playford ED, Hobart JC, Thompson AJ. Comparison of physician outcome measures and patients' perception of benefits of inpatient neurorehabilitation. *BMJ*. 2002;324(7352):1493.

91. Campbell R, Quilty B, Dieppe P. Discrepancies between patients' assessments of outcome: qualitative study nested within a randomised controlled trial. *BMJ*. 2003;326(7383):252-253.

92. Langendijk H, Aaronson NK, de Jong JM, ten Velde GP, Muller MJ, Wouters M. The prognostic impact of quality of life assessed with the EORTC QLQ-C30 in inoperable non-small cell lung carcinoma treated with radiotherapy. *Radiother Oncol*. 2000;55(1):19-25.

93. Montazeri A, Milroy R, Hole D, McEwen J, Gillis CR. Quality of life in lung cancer patients: as an important prognostic factor. *Lung Cancer*. 2001;31(2-3):233-240.

94. Shadbolt B, Barresi J, Craft P. Self-rated health as a predictor of survival among patients with advanced cancer. *J Clin Oncol*. 2002;20(10):2514-2519.

95. Maisey NR, Norman A, Watson M, Allen MJ, Hill ME, Cunningham D. Baseline quality of life predicts survival in patients with advanced colorectal cancer. *Eur J Cancer*. 2002;38(10):1351-1357.

96. Lakusta CM, Atkinson MJ, Robinson JW, Nation J, Taenzer PA, Campo MG. Quality of life in ovarian cancer patients receiving chemotherapy. *Gynecol Oncol*. 2001;81(3):490-495.

97. Herndon JE II, Fleishman S, Kornblith AB, Kosty M, Green MR, Holland J. Is quality of life predictive of the survival of patients with advanced nonsmall cell lung carcinoma? *Cancer*. 1999;85(2):333-340.

98. Konstam V, Salem D, Pouleur H, et al, SOLVD Investigators. Baseline quality of life as a predictor of mortality and hospitalization in 5,025 patients with congestive heart failure. *Am J Cardiol*. 1996;78(8):890-985.

99. Fan VS, Curtis JR, Tu SP, McDonell MB, Fihn SD, Ambulatory Care Quality Improvement Project Investigators. Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases. *Chest*. 2002;122(2):429-436.

100. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. *Arthritis Rheum*. 2003;48(6):1530-1542.

101. Eton D, Fairclough D, Cella D, Yount SE, Bonomi P, Johnson DH. Early change in patient-reported health during lung cancer chemotherapy predicts clinical outcomes beyond those predicted by baseline report: results from Eastern Cooperative Oncology Group Study 5592. *J Clin Oncol*. 2003; 21(8):1536-1543.

102. Brown JE, Butow PN, Culjak G, Coates AS, Dunn SM. Psychosocial predictors of outcome: time to relapse and survival in patients with early stage melanoma. *Br J Cancer*. 2000;83(11):1448-1453.

103. Butow PN, Coates AS, Dunn SM. Psychosocial predictors of survival: metastatic breast cancer. *Ann Oncol*. 2000;11(4):469-474.

104. Chassany O, Marquis P, Scherrer B, et al. Validation of a specific quality of life questionnaire in functional digestive disorders. *Gut*. 1999; 44(4):527-533.

105. Mor V, Guadagnoli E. Quality of life measurement: a psychometric tower of Babel. *J Clin Epidemiol*. 1988;41(11):1055-1058.

106. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. May 2007. Available at: www.fda.gov /cder/guidance/7478fnl.pdf. Accessed September 10, 2007.

107. Flikkila T, Nikkola-Sihto A, Kaarela O, Paakko E, Raatikainen T. Poor interobserver reliability of AO classification of fractures of the distal radius: additional computed tomography is of minor value. *J Bone Joint Surg Br*. 1998;80(4):670-672.

108. Sjoden GO, Movin T, Guntner P, et al. Poor reproducibility of classification of proximal humeral fractures: additional CT of minor value. *Acta Orthop Scand*. 1997;68(3):239-242.

109. Sahlroot JT, Pledger GW. Dosage adjustments in response to monitored plasma concentrations: can unblinded staff adhere to objective criteria? *J Biopharm Stat*. 1994;4(1):91-100.

110. Grech ED, Ramsdale DR. Acute coronary syndrome: unstable angina and non-ST segment elevation myocardial infarction. *BMJ*. 2003;326(7401):1259-1261.

111. Wang JK, Nauss LA, Thomas JE. Pain relief by intrathecally applied morphine in man. *Anesthesiology*. 1979;50(2):149-151.

112. Behar M, Magora F, Olshwang D, Davidson JT. Epidural morphine in treatment of pain. *Lancet*. 1979;1(8115):527-529.

113. Smith TJ, Staats PS, Deer T, et al, Implantable Drug Delivery Systems Study Group. Randomized clinical trial of an implantable drug delivery system compared with comprehensive medical management for refractory cancer pain: impact on pain, drug-related toxicity, and survival. *J Clin Oncol*. 2002; 20(19):4040-4049.

114. GlaxoSmithKline. Prescribing information, Advair Diskus.® Available at: http://www.fda.gov/cder/foi/label/2006/021077s028lbl.pdf. Accessed August 8, 2007.

115. Justice AC, Holmes W, Gifford AL, et al, Adult AIDS Clinical Trials Unit Outcomes Committee. Development and validation of a self-completed HIV symptom index. *J Clin Epidemiol*. 2001;(suppl 1):S77-S90.

116. Justice AC, Rabeneck L, Hays RD, Wu AW, Bozzette SA, Outcomes Committee of the AIDS Clinical Trials Group. Sensitivity, specificity, reliability, and clinical validity of provider-reported symptoms: a comparison with self-reported symptoms.*J Acquir Immune Defic Syndr*. 1999;21(2):126-133.

117. Komiti A, Judd F, Grech P, et al. Depression in people living with HIV/AIDS attending primary care and outpatient clinics. *Aust N Z J Psychiatry*. 2003;37(1):70-77.

118. Talley NJ, Weaver AL, Zinsmeister AR, Melton LJ III. Self-reported diarrhea: what does it mean? *Am J Gastroenterol*. 1994;89(8):1160-1164.

119. Irvine EJ, Ferrazzi S, Pare P, Thompson WG, Rance L. Health-related quality of life in functional GI disorders: focus on constipation and resource utilization. *Am J Gastroenterol*. 2002;97(8):1986-1993.

120. Mertz H, Beck K, Dixon W, Esquivel MA, Hays RD, Shapiro MF. Validation of a new measure of diarrhea. *Dig Dis Sci*. 1995;40(9):1873-1882.

121. Dwyer KA, Coty MB, Smith CA, Dulemba S, Wallston KA. A comparison of two methods of assessing disease activity in the joints. *Nurs Res*. 2001;50(4):214-221.

122. Pincus T, Strand V, Koch G, et al. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. *Arthritis Rheum*. 2003;48(3):625-630.