

Methods to Explain the Clinical Significance of Health Status Measures

GORDON H. GUYATT, MD; DAVID OSOBA, MD; ALBERT W. WU, MD; KATHLEEN W. WYRWICH, PhD;
GEOFFREY R. NORMAN, PhD; AND THE CLINICAL SIGNIFICANCE CONSENSUS MEETING GROUP

One can classify ways to establish the interpretability of quality-of-life measures as anchor based or distribution based. Anchor-based measures require an independent standard or anchor that is itself interpretable and at least moderately correlated with the instrument being explored. One can further classify anchor-based approaches into population-focused and individual-focused measures. Population-focused approaches are analogous to construct validation and rely on multiple anchors that frame an individual's response in terms of the entire population (eg, a group of patients with a score of 40 has a mortality of 20%). Anchors for population-based approaches include status on a single item, diagnosis, symptoms, disease severity, and response to treatment. Individual-focused approaches are analogous to criterion validation. These methods, which rely on a single anchor and establish a minimum important difference in change in score, require 2 steps. The first step establishes the smallest change in score that patients consider, on average, to be important (the minimum important difference). The second step estimates the proportion of patients who have achieved that

minimum important difference. Anchors for the individual-focused approach include global ratings of change within patients and global ratings of differences between patients. Distribution-based methods rely on expressing an effect in terms of the underlying distribution of results. Investigators may express effects in terms of between-person standard deviation units, within-person standard deviation units, and the standard error of measurement. No single approach to interpretability is perfect. Use of multiple strategies is likely to enhance the interpretability of any particular instrument.

Mayo Clin Proc. 2002;77:371-383

CHQ = Chronic Heart Failure Questionnaire; CRQ = Chronic Respiratory Questionnaire; EORTC QLQ-C30 = European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire core 30 items; FEV₁ = forced expiratory volume in 1 second; MID = minimum important difference; NNT = number needed to treat; NYHA = New York Heart Association; QOL = quality of life; SF-36 = 36-Item Short-Form Health Survey

Measures to detect important effects of treatment must be valid (measure what is intended), responsive (able to detect an important change, even if that change is small), and interpretable (the intended audience must understand the magnitude of effect). This series of articles deals with interpretation, and this article summarizes the approaches that investigators have used thus far. None of these approaches is without its limitations, but all contribute important information. The intent of this article is pri-

marily to describe the available strategies, but also to point out some strengths, weaknesses, and directions for subsequent investigation.

THE PROBLEM OF MEANINGFULNESS

Those responsible for making treatment recommendations, such as clinicians for individual patients or experts and health policymakers for groups of patients, must weigh the expected benefits of a treatment against its adverse effects, toxic effects, inconvenience, and cost. This process requires a reasonably accurate understanding of the benefits and risks of alternative treatments. Acquiring this understanding presents a significant problem even for dichotomous clinical outcomes, such as stroke, myocardial infarction, or death.^{1,2} For instance, how clinical trials present their results influences clinicians' inclination to treat. Typically, clinicians' enthusiasm for intervening decreases progressively as they see results presented in terms of relative risk reduction, absolute risk reduction, or the number needed to treat (NNT, the inverse of the absolute risk reduction).³⁻⁶ When presented as increases in life expectancy, mortality benefits of most life-prolonging treatments appear trivial and are likely to leave clinicians less enthusiastic about intervening.

From the Department of Clinical Epidemiology and Biostatistics, McMaster University and Health Sciences Center, Hamilton, Ontario (G.H.G., G.R.N.); QOL Consulting, West Vancouver, British Columbia (D.O.); Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Md (A.W.W.); and Department of Research Methodology, Saint Louis University, St Louis, Mo (K.W.W.). A complete list of other Clinical Significance Consensus Meeting Group contributors to this article appears at the end of the article.

This project was supported in part by Public Health Service grants CA25224, CA37404, CA15083, CA35269, CA35113, CA35272, CA52352, CA35103, CA37417, CA63849, CA35448, CA35101, CA35195, CA35415, and CA35103.

Individual reprints of this article are not available. The entire Symposium on the Clinical Significance of Quality-of-Life Measures in Cancer Patients will be available for purchase as a bound booklet from the *Proceedings* Editorial Office at a later date.

The problem becomes more complex when one considers that patients may vary in the value they place on a particular benefit. Furthermore, the same patient may place a different value on the same benefit, depending on the patient's circumstances. For instance, a 40-year-old patient may consider an increase in life expectancy of 3 months worth the inconvenience and potential adverse effects of lifelong antihypertensive therapy. Another patient may not. Both, however, may be willing to receive toxic chemotherapy to gain 3 months of life expectancy when they face a cancer that is likely to kill them within a year. In other words, the same 3-month gain in life expectancy may seem more important when it pertains to the short term rather than the long run.

These difficulties occur despite the ease with which one can conceptualize an event such as a stroke, myocardial infarction, or death. The problem becomes considerably more challenging when decisionmakers face quality-of-life (QOL) scores from unfamiliar ordinal or continuous scales. In this article, we discuss the challenges of making health status measures meaningful and the strategies that investigators have used to address these challenges. We discuss the possible target audience for information about intervention effects on health status, issues in understanding differences in scores between individuals and between groups, and a taxonomy of alternative approaches for making results more interpretable. Throughout, we present examples of applications of each of these approaches.

THE TARGET AUDIENCES FOR CLINICAL SIGNIFICANCE

The intended audience for our discussion on clinical significance includes patients, clinicians, and policymakers. Increasing awareness that value judgments are implicit in every clinical management decision⁷ has focused more attention on the role of the patient in the decision-making process.^{8,9} For patients who desire major involvement in decision making, one approach involves presenting patients with the options and eliciting their choice. Using this approach requires that patients understand the magnitude of benefits they may expect from treatment.

Widespread and effective integration of patients into clinical decision making remains an elusive goal. Whether the decision-making process remains primarily in the hands of the clinician or whether in the future the clinician's prime charge will evolve to communicating the expected consequences of management options to the patients, the clinician must understand the anticipated consequences of alternative management plans.

Constraints on health care spending demand difficult resource allocation choices. Hospital and public health administrators, bureaucrats, politicians, and, particularly in

the United States, the executives of private companies carry the burden of such choices. If they are to make wise decisions, these individuals must have a clear sense of the impact of alternative allocation of resources on patient well-being.

In our endeavor to make outcome measures meaningful, we must consider whether we are addressing patients, clinicians, health policymakers, or some combination. The helpfulness of presenting a particular expected treatment impact as an effect size, the NNT required to achieve a particular functional improvement in an individual patient, or an anticipated reduction in health service utilization is likely to differ among these audiences. The "clinical" in clinical significance must therefore be broadly and, depending on the context, variably defined.

THE PROBLEM OF MEANINGFULNESS IN QOL MEASURES

We have noted a problem in presenting results of studies using binary outcomes: the different meaning conveyed by relative and absolute risk reduction, NNT, and life-years gained. The complexity increases with the realization that no binary outcome is truly unambiguous. Deaths can be painful or painless, strokes can be mild or severe, and myocardial infarctions can be large and complicated or small and uncomplicated. In fact, severity of stroke and myocardial infarction are continuous in nature, although clinicians refer to them in categorical terms, such as mild, moderate, and severe. When we ask patients, clinicians, and policymakers to consider a decrease in risk of stroke or myocardial infarction (ignoring the distribution of size or severity of the strokes or infarcts they are preventing), we simplify the problem in hopes of making the information manageable. Our difficulties involve avoiding both the Scylla of misleading oversimplifications and the Charybdis of overly complex presentations that decisionmakers find confusing or useless.

Attempts to make the results of QOL studies meaningful are complicated by another challenge. Although they may choose to make decisions on the basis of strokes or myocardial infarctions prevented, rather than incorporating information about severity, clinicians have no difficulty in understanding the implications for their patients of strokes or myocardial infarctions, be they mild or severe. On the other hand, even those familiar with the concept of QOL assessment generally have no intuitive notion of the significance of a change in score of a particular magnitude.

Thus, before we can make the results of a clinical trial of treatment impact on QOL meaningful to a clinician, we must understand the significance of changes in individual patient scores. One can frame the problem as an issue of interpretability: what changes in score correspond to

trivial, small, moderate, or large patient benefit.¹⁰ If a person improves by 5 points in emotional function, will he be perceived as being happier by his family, will he miss less work, and will he no longer have to take antidepressant medication? Or will no such changes occur? If a patient with chronic lung disease improves by 5 points in physical function, will she now be able to climb a flight of stairs comfortably, keep up with her spouse when they go for a walk, and resume playing with her grandchildren? Or will she remain incapacitated by exertional dyspnea?

One can conceptualize 2 steps in the process of making results meaningful. One is understanding what changes in score mean to the individual, and the other is making results of clinical trials comprehensible to decisionmakers.¹¹ Presentation of mean changes in QOL (the treatment group improved by 5 points more than the control group) can be misleading. The proportion of patients achieving a particular degree of benefit and the corresponding NNT to ensure a single person obtains that benefit provide a more informative way of presenting results.

INFERENCES CONCERNING INDIVIDUALS AND INFERENCES CONCERNING GROUPS

Observers frequently distinguish between the significance of a particular change in score in an individual and a change of the same magnitude in the mean score of a group of patients.¹² A change in mean blood pressure in a population of a magnitude that would be trivial in an individual (eg, 2 mm Hg) may translate into a large number of reduced strokes in a population. Indeed, a mean change of 2 mm Hg in a population would reduce the number of strokes substantially. There are 2 reasons for the difference in interpretation.

The first reason that one might consider a change in blood pressure of 2 mm Hg in an individual trivial is that it is within the error of measurement. In this sense, the change is trivial because we do not believe it is real. However, one could specify (although it would be challenging to establish) that a treatment achieves a true change of 2 mm Hg in an individual. Then, we may not be as dismissive of such a finding.

The second reason for the difference between interpretation of individual and group differences is that every individual in the population does not experience the same change in outcome; rather, there is a distribution at the aggregate level. Some patients achieve a much greater reduction in blood pressure than the mean, whereas others achieve less or may even have a rise in blood pressure as a result of treatment.

Considering the variability in individual response highlights the fundamental deficiency of summarizing treatment effects as a difference in means. The clinician who

assumes that each individual experiences the mean effect is liable to make flawed clinical decisions. Depending on the distribution of the individual differences, the same mean difference can have different implications.¹³ For instance, assume there is a threshold below which any change in status has no important consequences for the patient, and mean change in a population is below that threshold. If the distribution of change with treatment is narrow, it is possible that no patient will achieve an important benefit with treatment. On the other hand, if the distribution of change is large, a substantial number of patients may achieve a benefit.

The difference between interpretation of a change in score in an individual and a group highlights the 2 steps in making clinical trial results meaningful.¹¹ For example, consider the effect of thrombolytic agents on long-term function in stroke patients. To quantify the impact of therapy, the Rankin scale, which measures functioning after a stroke, could be used. The scale is an ordinal one, grading patients as having no poststroke symptoms to being severely handicapped and totally dependent, with a higher score indicating poorer health.¹⁴ Scores of up to 2 of 5 indicate that patients are still able to look after themselves, so the investigators classified scores of 3 to 5 on this instrument as characterizing a poor outcome. The investigators were therefore able to present clinicians with a bottom line: thrombolytic therapy reduces the odds of the combined outcome of death and dependency after approximately 3 months by 17% (odds ratio, 0.83; 95% confidence interval, 0.70-0.95). In absolute terms, 22 patients need to be treated to prevent 1 patient from dying or becoming dependent after 3 months.¹⁵

Note that to use this result, the clinician does not have to know about the Rankin instrument. All that is required is an intuitive notion of what it means to be dependent and agreement that dependency is a central outcome for this patient population. However, without an ability to interpret the Rankin scale (the first step), the investigators could not have arrived at this transparent and useful way of presenting the results (the second step).

ANCHOR-BASED METHODS

Investigators have used 2 easily separable strategies to achieve an understanding of the meaning of scores on a given instrument.¹² The first relies on anchor-based methods and examines the relationship between scores on the instrument whose interpretation is under question (the target instrument) and some independent measure (an anchor). For instance, we might examine the relationship between scores on a QOL measure for heart failure and the New York Heart Association (NYHA) functional classification that categorizes patients into 1 of 4 groups from

Table 1. Two Anchor-Based Approaches to Establishing Interpretability

	Individual focused (single anchor)	Population focused (multiple anchor)
No. of anchors	1 (analogous to criterion validity)	Many (analogous to construct validity)
Specification of threshold	Specifies threshold between important and trivial change (minimum important difference [MID])	Instead of threshold, offers relationships between target measure and multiple anchors
Steps in application	2 steps: establishing MID and examining proportion who achieved MID	1 step: presenting population differences in status on anchors

those with no limitations to those whose dyspnea and fatigue limit them to minimal activity.

One can subclassify anchor-based approaches into those that solve the presentation problem in a single step, which we call a population-focused approach, and those, like the Rankin scale, that require 2 separate steps (Table 1), which we call an individual-focused approach. The former approach classifies patients in terms of the population to which they belong. For instance, mean scores on a hypothetical heart failure QOL measure are 50 in the intervention group and 40 in the control group. The 1-step approach would tell the clinician that, in a year, a group of patients with a mean score of 40 typically have a mortality of 20% and those with a mean score of 50 have a mortality of 10%.

The individual-focused approach would begin by specifying, for instance, that on the basis of long clinical experience with the NYHA functional classification, we know that the difference between class II and class III represents twice the minimum important change in patient function. Class III patients have a mean score of 40 and those in class II have mean scores of 50 (which happen to be the mean scores in the control and intervention group). This does not, however, mean that every patient changes from class III to class II. To avoid leaving clinicians with this impression, one would have to go further.

For instance, one might examine the proportion of patients in the intervention and control arms who improved by 1 class, remained the same, or deteriorated by 1 class. One could further calculate the proportion of treated patients who were 1 class better than they would have been had they not received treatment and calculate the associated NNT.¹³ The experience of one of the authors (G.H.G.) with medical trainees and practitioners, in his home institution and many other institutions, suggests that the NNT may be a particularly compelling way to present results. The measure appears to be capturing attention: the numbers of articles in MEDLINE mentioning NNT in the title or abstract for the years 1995 through 2000 were 13, 32, 42, 68, 116, and 159, respectively.

There are other differences between the population-focused and individual-focused approaches. The popula-

tion-focused approach is analogous to establishing construct validity, in that multiple anchors are generally required. For instance, the clinician told about the mortality associated with scores of 40 and 50 might rightly claim that she still has little idea of how to interpret the results in terms of patients' function. One could then add that 10% of a population of patients younger than 65 years with a mean score of 40 would be gainfully employed, whereas 40% of those with a score of 50 would have a job. The individual-focused strategy tends, in contrast, to focus on a single anchor.

Finally, those taking the population-based approach most commonly avoid identifying a threshold between a change in score that is trivial and a change that is important. Perhaps this hesitation implicitly acknowledges that the threshold may vary, depending on the population under study and the range and severity of the QOL problems. Those taking individual-based approaches usually attempt to define such a threshold.

APPROACHES FOR IDENTIFYING CLINICAL SIGNIFICANCE

We have not conducted a systematic search for approaches to clinical significance. Thus, our examples are neither comprehensive nor representative. Rather, we have attempted to provide a broad sample of approaches investigators have used, focusing on those we believe are both well done and instructive. However, we have surveyed the entire group of participants in this conference to ensure that we have not omitted any salient methods.

Similarly, we have not tried to be systematic in our critique. Rather, our comments reflect the particular, and perhaps idiosyncratic, perspectives of the authors. Nevertheless, our range of opinions before we started the exercise was somewhat diverse, and the article has been through the crucible of criticism from our colleagues at the conference.

ANCHOR-BASED METHODS OF ESTABLISHING INTERPRETABILITY: REQUIREMENTS

Whether relying on a single anchor or multiple anchors, anchor-based methods have 2 requirements. First, the an-

chor must be interpretable. It would be of little use to tell clinicians that a 2-point change per item in the fatigue scale (range, 1-7) in the Chronic Heart Failure Questionnaire (CHQ)¹⁶ is equivalent to a 30-point change in the Medical Outcome Study physical function scale if they had no idea how to interpret the Medical Outcome Study instrument. On the other hand, if they use the NYHA functional classification system in daily practice, knowing that a 2-point change in the CHQ corresponds to change of 1 NYHA class would be useful.

Second, there must be an appreciable association between the target and anchor. If there is no relationship between the target and anchor, there will be no difference between mortality in patients with differing QOL scores, nor will NYHA classification separate patients into groups with varying scores. Further experience will be required to determine just how small the association may be and still yield sensible and useful results. For the time being, we note that the stronger the association, the more secure the inferences about interpretation of the target measure, and weak associations are liable to yield misleading results. Finally, the single-anchor methods will require a higher degree of association than the multiple-anchor methods to generate convincing inferences.

CLINICIANS' TRADITIONAL APPROACHES AND INTERPRETABILITY

Experienced clinicians show little hesitation in acting on the clinical measures, yielding continuous scores, by which they judge their patients' status. Hemoglobin concentration, platelet count, creatinine level, and treadmill exercise capacity constitute a few examples. How does the process of establishing interpretability occur? How, for instance, do chest physicians decide that a change in forced expiratory volume in 1 second (FEV₁) of 15% approximates a minimum important change?

Chest physicians consider the degree of difference generally required to keep the patient out of the hospital, to increase the patient's likelihood of returning to work, or to make the patient less dyspneic when engaging in basic activities of daily living. In addition, they are able to observe directly the patients who receive testing and ask patients for their own impressions. Their experience with the relation between the target instrument and independent measures, holistically processed, informs their judgment.

In deciding whether they were ready to consider a 10%, 15%, or 20% change in FEV₁ important, clinicians used independent standards, which they valued highly, such as cutting the risk of hospitalization, doubling the probability of going back to work, and substantially decreasing distressing dyspnea in daily activities. There are candidate-independent standards that are easily interpretable that cli-

nicians would not use in making their judgment. They might, for instance, reject patients' general level of happiness. The reason is that they know that there is a weak correlation between this standard and FEV₁, making it an unsatisfactory comparator.

Because clinicians rarely use QOL measures in clinical practice, the intuitive approaches that investigators have used to establish the interpretability of clinical measures are an undeveloped option. As a result, investigators must choose independent standards and present data from QOL studies to facilitate interpretation of clinical trial results. At the same time, it is worth bearing in mind that clinicians' day-to-day use of QOL measures could ultimately enhance interpretability.¹⁷

MULTIPLE ANCHORS

Ware and Keller,¹⁸ with the 36-Item Short-Form Health Survey (SF-36), have accomplished extensive and comprehensive work using multiple anchors, and we rely to a large extent on their studies to provide examples of this approach. In our discussion, we deal initially with anchors that involve concurrent measurement of the target and anchor and subsequently discuss anchors that involve monitoring patient outcome over time (health care utilization, job loss, and death).

Level of Function From a Single Item

Status on an important, common, and easily understood measure of function can provide a useful anchor. For example, mobility understood as the difference between using a wheelchair and walking with an aid has provided 1 such standard.¹⁹ In 1 way of applying this approach, for instance, one picks a difference in score of 40 and 50 on the SF-36 physical function scale.¹⁸ One then chooses a dichotomy on a single item that is part of the instrument, such as whether people are limited in walking a single block. Although 32.1% of those who score 40 on the SF-36 physical function scale can walk a block without difficulty, this is true of 49.7% of those who score 50 on the physical function scale, a difference of 17.6% in absolute terms.¹⁸

Symptoms

Investigators can use the presence of symptoms to aid interpretability in exactly the same way that they can use functional levels. For example, as scores on the 5-item SF-36 mental health scale increase from 20 to 40, the percentage of people reporting suicidal ideation decreases from 29.6% to 14.4%, and the percentage reporting dissatisfaction with life decreases from 42.4% to 29.1%.¹⁸ An additional challenge in interpretability is that the same absolute difference in score may have a different meaning across different portions of the scale (20 to 40 may mean some-

thing different from 80 to 100). As scores on the SF-36 mental health scale increase from 80 to 100, the percentage of people reporting suicidal ideation decreases from 1.9% to 0.5%, and the percentage reporting dissatisfaction with life decreases from 4.5% to 1.0%.¹⁸

Diagnosis

Mean scores of patients with a particular diagnosis have provided another anchor for adding meaning to health status measures. Investigators have reported Sickness Impact Profile mean scores (range, 0-100; higher scores indicate worse health status) for a number of diagnoses, including severe rheumatoid arthritis (15.6²⁰ and 25.8²¹), chronic airflow limitation severe enough to require a home oxygen tank (24²²), chronic, stable angina (approximately 11.5²³), and amyotrophic lateral sclerosis.²⁰ These anchors may be particularly useful for clinicians who treat such patients. A related method, which evaluates patients according to whether their scores fall in the normal range or in the range associated with those with psychiatric diagnoses, has proved popular in mental health research.^{24,25}

Disease Severity

King²⁶ has provided an impressive example of how severity of illness can provide anchors for interpretation of the European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire core 30 items (EORTC QLQ-C30) by reviewing results from 14 published studies. In general, as one would anticipate, sicker patients showed lower (worse) scores in variables such as the presence or absence of metastatic disease or prognosis.

Response to Treatment

In the same article, King also provides data on EORTC QLQ-C30 scores of patients before and after 7 days of chemotherapy for breast cancer, before and after chemotherapy or radiotherapy for carcinoma of the lung, and before and after radiotherapy for gynecological cancer. Clinicians treating these patients on a regular basis are likely to find these results helpful in enhancing the interpretability of the EORTC QLQ-C30. In another example of this approach, investigators found that mean Sickness Impact Profile scores of 30 in patients shortly after hip replacement decreased to less than 5 after full convalescence.²⁷

Job Loss

Typically, chronic diseases affect older persons, many of whom may have retired for reasons other than their health problems. If one is dealing with a younger population, health problems may be a major cause of loss of work. Thus, job loss can be used to help interpret score differ-

ences.²⁸ In 1 study,¹⁸ the proportion of people who had lost their jobs 1 year later increased progressively from 3.1% for those with SF-36 physical component summary scores of 55 to 72, to 10% of those with scores of 45 to 54, to 17.8% of those with scores of 35 to 44, and finally to 32.2% of those with scores of 8 to 34.

Health Care Utilization

Executives working in the managed care health industry in the United States may be particularly interested in health care utilization as an anchor for interpreting QOL measures. Typically, those with poorer self-reported health will make more visits to the physician, use more drugs, and require more hospitalizations.²⁹ In 1 study,¹⁸ the proportion of patients using mental health outpatient services increased from 37.4% to 45.8% as SF-36 mental health scores decreased from 40 to 20 and increased from 4.6% to 13.2% as scores decreased from 100 to 80. We note that in terms of the anchor of health care utilization, changes in different parts of this scale had almost the same meaning.

Mortality

We have already shown, in a hypothetical example, how mortality can be used as an anchor to enhance the meaningfulness of QOL measures. As self-reported health deteriorates, mortality tends to rise.³⁰ Using the same cutpoints for SF-36 physical component summary scores as for job loss previously stated (but dividing the 8- to 34-year age group into those with scores of 25 to 34 and those with scores of 8 to 24), the percentage of people dying during the next 5 years increased in 1 study¹⁸ from 1.8% to 4.7%, 6.2%, 15.1%, and finally 21.5%.

Comments on the Multiple-Anchor Approach

We have already mentioned the trade-off between approaches of enhancing interpretability whose goal is simplicity and those that retain more of the inherent complexity of the QOL measurement. The former strategies risk oversimplification, whereas the latter may provide more information than patients and clinicians can readily interpret. In addition, availability of different anchors may allow individuals to find examples more salient to them. To the extent that these methods fail to expose human suffering that characterizes the clinical arena, it may be because of the inherent complexity of the presentation. The multiple-anchor approach acknowledges that changes in score that represent small, medium, and large effects may differ with varying diseases and disease severity and even across the range of scores of a single instrument. This is both its strength (it avoids misleading simplification) and its weakness (it may impose an excessive cognitive burden on

patients and clinicians). Finally, these methods do not answer the question of the meaning of scores smaller than those observed in the different groups that the investigators are comparing.

SINGLE-ANCHOR METHODS

The Minimum Important Difference

Single-anchor methods generally aim to establish differences in score on the target instrument that constitute trivial, small but important, moderate, and large changes in QOL. However, they generally put great emphasis on a threshold that demarcates trivial from small but important differences: the minimum important difference (MID). One popular definition of the MID is “the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s (health care) management.”³¹

Several factors have made the concept of MID useful. First, it ties the magnitude of change to treatment decisions in clinical practice. Second, the smallest important difference one wishes to detect helps with the study design and choice of sample size; this definition also links to a crucial decision in trial design. Third, it emphasizes the primacy of the patient’s perspective and implicitly links that perspective to that of the physician. Since discussions of the ethics of clinical care increasingly emphasize shared decision making, this link is useful. Finally, the concept appears easily understood by clinicians and investigators (although there is little experience with patients).

A limitation of this definition of MID is that it does not explicitly address deterioration. One way to address this problem would be to modify the definition as follows: “the MID is the smallest difference in score in the domain of interest that patients perceive as important, either beneficial or harmful, and which would lead the clinician to consider a change in the patient’s management.”

An alternative to labeling a change as being of “minimum importance” is to think of it as “subjectively significant.”³² This latter term emphasizes that one can have an important deterioration and an important improvement. It also makes explicit that the meaningfulness of change over time is based entirely on the patient’s self-assessment of the magnitude of change. Thus, the term *subjectively significant* is congruent with the concept that QOL is a subjective construct and that the prime assessor of QOL status and change in that status is not an observer, but the patient. Others may, however, interpret the patient’s experience in light of their particular roles in the care of the patient. For example, the physician may or may not decide to introduce or alter an existing intervention, depending on a consideration of the balance in benefits and risks. A health care

system may characterize its goal as improving QOL at a reasonable cost. Alternatively, treatments that do not improve QOL status by at least a subjectively significant degree in at least some patients may be deemed worthless, and educational programs may be launched to discourage their use.

It is certain that the MID varies across patients and possibly patient groups. Santanello et al,³³ for instance, found differing MIDs in younger and older patients. Furthermore, any estimate of the MID will be associated with a degree of uncertainty. Thus, a more accurate statement than “the MID is 10” would be that “the average MID is likely to be in the vicinity of 7 to 13” or some other appropriate range for the variation among individuals. Nevertheless, using the best estimate of the average MID is likely to facilitate communication of results in a succinct and comprehensible fashion.

Within-Patient Health Transition Global Ratings

This strategy assumes that patients can judge whether, during a specified period, they are fundamentally unchanged, better, or worse. In the first major study using this approach, investigators attempted to enhance the interpretability (including establishing the MID) of the Chronic Respiratory Questionnaire (CRQ)³⁴ and an almost identical instrument for patients with heart failure, the CHQ. Both these disease-specific questionnaires include domains of dyspnea, fatigue, and emotional function and presented response options for each item as 7-point scales.

Data from 3 clinical trials contributed to the exploration of the interpretability of the CRQ and CHQ. In each of these studies, patients completed the CRQ or CHQ at each clinic visit. On all visits but the first, they also completed global ratings of change in their shortness of breath on day-to-day activities, their fatigue level, and how they were feeling emotionally. Patients specified whether they were worse, about the same, or better. If worse or better, they quantified the magnitude of the change using the following global rating scale: 1 indicates almost the same, hardly any worse (better) at all; 2, a little worse (better); 3, somewhat worse (better); 4, moderately worse (better); 5, a good deal worse (better); 6, a great deal worse (better); and 7, a very great deal worse (better).

The investigators classified ratings of 1 to 3 as small changes in function representing the MID, 4 or 5 as moderate changes, and 6 or 7 as large changes. They noted the corresponding change in the appropriate CRQ or CHQ domain (dyspnea with dyspnea, fatigue with fatigue, emotional function with emotional function) from the previous visit. For all 3 domains, the mean change per question associated with a global rating of “unchanged” was approximately 0. The larger the change as assessed by

global ratings, the larger the change in CHQ or CRQ score. The mean change in score per question corresponding to a small difference (MID) was consistently around 0.5; a mean change of 0.81 (emotional function) to 0.96 (fatigue dimension) represented a moderate effect, whereas a change from 0.86 (emotional function) to 1.47 (dyspnea) represented a large effect.³¹

Subsequently, the investigators used essentially the same technique to enhance the interpretability of questionnaires for measuring QOL in childhood and adult asthma and rhinoconjunctivitis, focusing on domains of symptoms, activity limitations, and emotional function. These instruments all had a similar structure and presented response options as 7-point scales. The results showed that the MID was consistently represented by a difference of approximately 0.5 per question (0.75 to 1.25 represents moderate differences, and differences of greater than 1.5 constituted large differences).³⁵⁻³⁷

Other investigators have used similar approaches. Osoba et al³² generated results that suggested that, in patients with breast or lung cancer, the MID of the EORTC QLQ-C30 was in the range of 5 to 10, with moderate and large effects represented by, respectively, changes of 10 to 20 and more than 20. Osoba et al^{38,39} have used this approach in the interpretation of several clinical trials. In the report of one of these studies,³⁸ the investigators emphasize that an improvement in QOL must continue for a minimum period before patients will consider it important.

Barber et al⁴⁰ compared use of a similar anchor to previous studies with an anchor that used an absolute rating ("How well is your asthma controlled?") rather than a transition question ("How much have you changed in the past *x* weeks?"). They found that the alternative anchors lead to different estimates of small, moderate, and large differences in questionnaire score.

One strength of this approach is its relative simplicity. Both clinicians and patients can relate easily to the concept of small, moderate, and large changes in QOL (although the extent to which individual patients and clinicians have similar effects in mind when they think of small, medium, and large changes remains to be established). The similar estimates of the MID generated when applying the same anchor to similarly structured questionnaires are reassuring.

This method is critically dependent on the validity of patients' ratings of change. The results are therefore considerably strengthened if the association between the target and anchor is appreciable. Even if the relationship is appreciable, research in some areas suggests that transition scores may simply reflect patients' current health state, rather than the extent to which they have improved or

deteriorated. In particular, patients' ability to recall their initial state may be critically dependent on the interval between assessments and the salience of the baseline time point.⁴¹ To date, few reported studies examine these relationships or other psychometric properties of transition ratings.

Although investigators have offered the global rating of change as a single target, inferences about the MID based on its results must be corroborated by other methods. These might include data from one or more of the multiple-anchor approaches. For example, in a study of patients with anaplastic astrocytoma who were receiving chemotherapy, those whose EORTC QLQ-C30 scores improved by more than 10 (on a 0-100 scale) in 3 or more domains were much more likely to have had a complete or partial tumor response (82% of 132 patients) than stable disease (59%) or progressive disease (20%). Thus, the tumor response provided support to the MID estimate.⁴² Data from alternative single anchors and intuitive estimates of the MID from clinicians who have used the target instrument in clinical practice constitute other sources of potential corroboration.⁴³

Between-Patient Global Ratings

Redelmeier et al⁴⁴⁻⁴⁷ have introduced a method that parallels the within-patient global ratings but relies on between-patient ratings. With this approach, patients who have completed the target instrument pair off and discuss their problems. Following the discussion, they rate their problems as the same, worse (to varying degrees), or less severe (to varying degrees) than the individual with whom they have spoken. The approach assumes that the difference in score between patients who rate themselves "a little better" or "a little worse" constitutes the MID.

Investigators have used the approach in patients with arthritis,⁴⁴ respiratory disease,⁴⁵⁻⁴⁷ and inflammatory bowel disease.⁴⁸ After discarding the dyspnea dimension's data and averaging the remaining dimensions' MIDs, Redelmeier's CRQ MID finding was also in the 0.5 range, consistent with those findings generated by within-patient global ratings of change. Because it involves seeing patients only once and without any particular requirements for their clinical status, the approach is practical.

The approach is limited in that patients may have difficulties describing health status to one another, particularly in areas related to emotional function. Theoretically, ratings will be compromised by the noise generated by patients' perception of their own QOL and that of their paired partner. Finally, in treating patients, clinicians are interested in the within-patient change over time, and it is plausible that the within- and between-patient, or longitudinal and cross-sectional, MIDs may differ.

Life Events

Testa et al^{49,50} have pioneered an approach to interpretability that relies on an instrument that quantifies the stress of life events, the Holmes-Rahe stressful life events scale, and have applied the approach to the results of several clinical trials. The problems with this approach are considerable, in that both fundamental requirements of a suitable anchor are open to serious question. First, few would find life-change units themselves interpretable. For instance, most clinicians are likely to find Testa's statement that "a change of 0.1 responsiveness units over a 2-month period was associated with a change of 27 [life-change units] over the same period"¹¹ unenlightening. Second, the association between life change and many QOL measures is likely to be too low to be useful.

Established Functional Rating Systems

Two functional rating systems, the American Rheumatism Association functional classification and the NYHA functional classification, are in sufficiently wide use by clinicians, and thus sufficiently well understood, that they offer possible anchors for QOL measures.⁵¹ For instance, overall Sickness Impact Profile scores associated with American Rheumatism Association classes I to IV have been estimated as 8.2, 15.1, 20.3, and 25.8.⁵² The approach is limited to functional classifications that are familiar to clinicians. Furthermore, although clinicians may believe they are communicating effectively when they describe a patient as "NYHA class III," the interobserver reliability of the classification system is limited.⁵³ Finally, in a rating system with 4 categories, the difference between one category and the next is almost certain to be considerably greater than the MID.

ANALYTIC STRATEGIES FOR SINGLE-ANCHOR APPROACHES

Having chosen a single-anchor approach, investigators may use alternative analytic strategies that will lead to different estimates of the MID.⁵⁴ The simplest and so far most widely used approach is to specify a result or range of anchor instrument results that corresponds to the MID and calculate the target score corresponding to that value. For example, investigators have examined the mean change in QOL score corresponding to global ratings of change that included "hardly any better," "a little better," and "somewhat better." Investigators can also choose anchor categories that represent moderate and large differences and calculate corresponding mean scores.

The most widely used alternative is an approach borrowed from diagnostic testing, the use of receiver operating characteristic curves.⁵⁵⁻⁵⁷ In this strategy, each patient is classified according to the anchor instrument as experienc-

ing an important change or not experiencing a change. Investigators then test a series of cutpoints to determine the number of misclassifications. These misclassifications will include false-positive results (patients mistakenly categorized as changed) and false-negative results (patients mistakenly categorized as unchanged). The optimal cutpoint will minimize the number of misclassifications.

Whichever single anchor or whatever method of analysis one uses, one faces the possibility that the interpretation of change in scores differs across the range of possible scores. A difference of 0.5 point may mean something different in the portion of the scale that corresponds to mild QOL impairment and the portion of the scale that corresponds to severe QOL impairment. Furthermore, it is possible that the same change in QOL score on a target instrument warrants different interpretation if it is an improvement, rather than a deterioration. These issues have, up to now, received limited attention in work with single-anchor methods.

SINGLE-ANCHOR APPROACHES AND CLINICAL TRIALS INTERPRETATION

Once one has established the MID for a patient, one must decide how to use this information in clinical trials. A naive approach would assume that if the mean difference between treatment and control was less than the MID, the treatment effect would be trivial, and if greater than the MID, the treatment effect would be important. This ignores the distribution of the results. For example, assume a MID of 0.5. A mean difference of 0.25 (trivial in a naive interpretation) could be achieved if 25% of the patients experience a benefit of 1.0 and 75% experience no benefit. This would result in an absolute difference of 25% in the proportion of patients achieving improvement and an NNT of 4.

The proportion of patients achieving a particular benefit, be it a small, moderate, or large difference, is therefore much more relevant than a mean difference from the clinician's point of view and less likely to mislead. To calculate the proportion who achieve a MID, one must consider not only the difference between groups in those who achieve that improvement but also the difference between groups in those who deteriorate by the same amount. One must therefore classify patients as improved, unchanged, or deteriorated. In a parallel group trial, the subsequent calculation is not altogether straightforward, and 1 approach involves assumptions about the joint distribution of responses in the 2 groups.¹³ Statisticians are developing alternative approaches to this problem, several of which are likely to prove reasonable.⁵⁸ What is not reasonable is simply to present mean values without taking the second step that is necessary for clinicians to interpret clinical trial results effectively.

The proportion who benefit is in itself a passably interpretable statistic for clinicians. Because of its intuitive appeal to clinicians, the NNT is even better. Another advantage of calculating the proportion who benefit is that it facilitates economic analysis. One can calculate the cost needed to achieve benefit in an individual.⁵⁹ However, it is important to note that simulation studies have demonstrated a strong relationship between the average treatment effect and the proportion benefiting from treatment (and the NNT), which is only weakly dependent on the choice of the MID. That is, for any treatment effect, the overall proportion benefiting and the NNT can be directly estimated from the effect size without the use of an established MID.⁶⁰

Distribution-based methods differ from anchor-based methods in that they interpret results in terms of the relation between the magnitude of effect and some measure or measures of variability in results. The magnitude of effect may be the difference in an individual patient's score before and after treatment, a single group's score before or after treatment, or, most germane to the current discussion, the difference in score between treatment and control groups. As a measure of variability, investigators may choose between-patient variability (the standard deviation of patients at baseline, for instance) or within-patient variability (the standard deviation of change that patients experienced during a study). If an investigator used the distribution-based approach, the clinician would see a treatment effect reported as, for instance, 0.3 standard deviation unit.

The enormous advantage of distribution-based methods is that the values are easy to generate. Whatever the study, there will always be one or more measures of variability available. This contrasts with the work needed to generate an anchor-based interpretation, evident from the prior discussion.

Distribution-based methods have, in general, 2 fundamental limitations. First, estimates of variability will differ from study to study. For instance, if one chooses the between-patient standard deviation, one has to confront its dependence on the heterogeneity of the population under study. If a trial enrolls an extremely heterogeneous population, an important effect may be small in terms of the between-person standard deviation and thus judged trivial. The same effect size, in a trial that enrolls an extremely homogeneous population, may be large in terms of the between-person standard deviation, and thus judged extremely important. The true impact of the change remains the same, but the interpretation differs radically.

There are at least 2 ways to deal with this problem. One is to choose the variability from a particular population, such as the standard deviation of a measure when applied to the general population at a point in time, and always refer

to that same measure of variability. The second is to choose the standard error of measurement (which we will discuss subsequently), which is theoretically sample independent.

Neither solution addresses the second fundamental problem of distribution-based methods. In deciding whether the magnitude of a treatment effect is worth the risks and costs, a clinician who knows that the effect is 0.3 standard deviation unit will be no further ahead. The units do not have intuitive meaning to clinicians. It is possible, however, that clinicians could gain experience with standard deviation units in the same way they learn to understand QOL scores.

In an enormously influential work, Cohen⁶¹ addressed this problem by suggesting that changes in the range of 0.2 standard deviation unit represent small changes, those in the range of 0.5 standard deviation unit represent moderate changes, and those in the range of 0.8 standard deviation unit represent large changes. Thus, one would tell a clinician that if trial results show a 0.3 standard deviation difference between treatment and control, then his/her patient can anticipate a small improvement in QOL with treatment.

The problem with this approach is its arbitrariness. Do 0.2, 0.5, and 0.8 standard deviation units always represent small, medium, and large effects? In response to this problem, recent investigations have attempted to provide empirical evidence about the relationship between distribution-based and anchor-based results. These studies address the question, "What is the appropriate interpretation of a particular magnitude of effect, in distribution-based units, as judged by the results of anchor-based studies?" In the remainder of this discussion, we will review some of the work of investigators who have focused on distribution-based methods.

BETWEEN-PERSON STANDARD DEVIATION UNITS

The most widely used distribution-based method to date is the between-person standard deviation. The group from which this is drawn is typically the control group of a particular study at baseline or the pooled standard deviation of the treatment and control groups at baseline. As we have mentioned herein, an alternative is to choose the standard deviation for a sample of the general population or some particular population of special interest, rather than the population of the particular treatment study under consideration. An advantage of this approach is that it has been applied widely in areas of investigation other than QOL.

Kazis et al⁶² have provided examples of how effect sizes can be used to provide a benchmark for interpreting change by examining the effect sizes generated by different treatments. These examples, drawn from studies of arthritis treatments, all used a single measure, the Arthritis Impact

Measurement scales. More recently, Samsa et al⁶³ have presented data suggesting that Cohen's effect sizes may, in fact, be generally applicable. In situations in which a single scale is used in several studies, one might look at the absolute magnitude of change rather than the number of standard deviation units. This would provide the same information but be free of the problem of varying baseline standard deviations dependent on the heterogeneity of the population enrolled.

Not all investigators have found that the MID corresponds to 0.2 standard deviation unit in their studies. Osoba et al³² found the MID to be in the range of 0.2 to 0.5 standard deviation unit in their investigation involving cancer patients.

STANDARD ERROR OF MEASUREMENT

The standard error of measurement is defined as the variability between an individual's observed score and the true score and is computed as the baseline standard deviation multiplied by the square root of 1 minus the reliability of the QOL measure. Theoretically, a QOL measure's standard error of measurement is sample independent, whereas its component statistics, the standard deviation and the reliability estimate, are sample dependent and vary around the standard error of measurement.⁶⁴ For instance, as the between-person variability in the population increases, the standard deviation will increase (tending to raise the standard error of measurement), but the reliability will also increase (tending to lower the standard error of measurement). Thus, the standard error of measurement largely reflects within-person variability over time.

Wyrwich et al^{64,65} have presented data comparing the standard error of measurement to the MID for the CRQ and CHQ developed using the single-anchor method described previously. They found a close correspondence between the anchor-based approach and a criterion of 1 standard error of measurement. Other investigators⁴³ have suggested that larger estimates (up to 2.77 standard errors of measurement) represent important changes in psychometric and physiologic measures.

A critical choice in calculating the standard error of measurement is whether one uses internal consistency or test-retest methods to calculate the reliability for the standard error of measurement. Although Wyrwich and colleagues argue for internal consistency, other investigators^{66,67} favor test-retest reliability.

RECONCILIATION OF ANCHOR-BASED AND DISTRIBUTION-BASED METHODS

Investigators are adducing increasing evidence concerning the relationship between statistical measures of patient variability and anchor-based estimates of small, moderate,

and large differences in QOL. To the extent that standard deviations across QOL studies using the same instruments are consistent, one will see a consistent relationship between the standard deviation and the MID. If this relationship were also consistent across instruments, this area of investigation would become much easier. Although the relationship between measures of variability and the MID will differ (and there is certain to be appreciable variability), clinical conclusions from QOL studies may prove robust to the variability seen in most QOL studies. If this is so, distribution-based methods may prove extremely useful.

TOPICS FOR FURTHER STUDY

It is possible that presentation of the proportion of patients achieving varying degrees of treatment benefit relative to control and the associated NNTs can provide an informative way of introducing QOL results to clinicians. However, this remains to be demonstrated. It would be important to involve clinicians in studies comparing alternative methods to present QOL data (multiple anchors vs single anchors, NNT effect sizes, and so on). In addition, global rating or health transition scales are an essential component of many anchor-based approaches, yet little is known about their psychometric properties. Research should address some of these concerns: degree of bias, serial applications in longitudinal studies, standards for association with QOL difference scores, sufficient sample sizes, and qualitative studies to investigate the cognitive process that individuals use to retrospectively assess changes in their health over time. It may also be possible to calculate MID on a given scale score for multiple anchors. Although this would increase the complexity of interpretation, it might allow users to base their decision making on anchors most personally relevant. It is also important to determine the relationship of differences expressed in standard deviation units to other methods of interpretation.

CONCLUSIONS

This review reflects both the considerable work that has been done to establish the interpretability of QOL measures in the last 15 years and the enormous amount left to do. The field remains controversial, and there are many alternative approaches, each with its advocates. The following conclusions, however, may be relatively safe. First, distribution-based methods will not suffice on their own but will be useful to the extent that they bear a consistent relationship with anchor-based methods. Second, even the single-anchor methods will require validation with alternative anchors. Finally, much more work is required on the acceptability of the various approaches if these are to be useful to clinicians in their day-to-day practice.

Clinical Significance Consensus Meeting Group contributors:

Neil Aaronson, PhD, Division of Psychosocial Research and Epidemiology, Cancer Institute, Amsterdam, the Netherlands; Ivan Barofsky, PhD, Johns Hopkins University School of Medicine, Baltimore, Md; Rick Berzon, PhD, Boehringer Ingelheim Pharmaceuticals, Ridgefield, Conn; Amy Bonomi, MPH, Center for Health Studies, MacColl Institute for Healthcare Innovation, Seattle, Wash; Monika Bullinger, PhD, University of Hamburg, Hamburg, Germany; Joseph C. Cappelleri, PhD, MPH, Global Research and Development, Pfizer Inc, Groton, Conn; David Cella, PhD, Center on Outcomes, Research and Education, Evanston Northwestern Healthcare, Northwestern University, Evanston, Ill; Diane Fairclough, DrPH, Colorado Health Outcomes, University of Colorado Health Sciences Center, Denver; Carol Estwing Ferrans, PhD, RN, College of Nursing, University of Illinois at Chicago; Marlene Frost, PhD, RN, Women's Cancer Program, Mayo Clinic, Rochester, Minn; Ron D. Hays, PhD, Departments of Medicine and Health Services Research, UCLA, Los Angeles, Calif; Patrick Marquis, MD, MBA, Mapi Values, Boston, Mass; Carol M. Moinpour, PhD, Southwest Oncology Group Statistical Center, Seattle, Wash; Tim Moynihan, MD, Division of Medical Oncology, Mayo Clinic, Rochester, Minn; Donald Patrick, PhD, MSPH, Department of Health Services, University of Washington, Seattle; Dennis Revicki, PhD, MEDTAP International Inc, Bethesda, Md; Teresa Rummans, MD, Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minn; Charles Scott, PhD, American College of Radiology, Philadelphia, Pa; Jeff A. Sloan, PhD, Department of Health Sciences Research, Mayo Clinic, Rochester, Minn; Mirjam Sprangers, PhD, Department of Medical Psychology, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; Tara Symonds, PhD, Outcomes Research Department, Pfizer Ltd, Sandwich, Kent, United Kingdom; Claudette Varricchio, DSN, RN, Division of Cancer Prevention, National Cancer Institute, Bethesda, Md; Gilbert Wong, MD, Division of Pain Medicine, Department of Anesthesiology, Mayo Clinic, Rochester, Minn.

REFERENCES

- Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centered approach to determining clinically important effect sizes for randomized treatment trials? *J Clin Epidemiol.* 1994;47:787-795.
- Feinstein AR. Indexes of contrast and quantitative significance for comparisons of two groups. *Stat Med.* 1999;18:2557-2581.
- Naylor DC, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med.* 1992;117:916-921.
- Hux JE, Levinton CM, Naylor CD. Prescribing propensity: influence of life-expectancy gains and drug costs. *J Gen Intern Med.* 1994;9:195-201.
- Redelmeier DA, Tversky A. Discrepancy between medical decisions for individual patients and for groups. *N Engl J Med.* 1990;322:1162-1164.
- Bobbio M, Demichellis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet.* 1994;343:1209-1211.
- Guyatt GH, Sinclair J, Cook DJ, Glasziou P, Evidence-Based Medicine Working Group and Cochrane Applicability Methods Working Group. Users' guides to the medical literature, XVI: how to use a treatment recommendation. *JAMA.* 1999;281:1836-1843.
- O'Connor AM, Rostom A, Fiset V, et al. Decision aids for patients facing health treatment or screening decisions: systematic review. *BMJ.* 1999;319:731-734.
- Guyatt G, Straus S, McAlister F, et al. Moving from evidence to action: incorporating patient values. In: Guyatt G, Rennie D, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* Chicago, Ill: AMA Press; 2002:567-582.
- Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med.* 1993;118:622-629.
- Testa MA. Interpretation of quality-of-life outcomes: issues that affect magnitude and meaning. *Med Care.* 2000;38(9, suppl):II166-III174.
- Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res.* 1993;2:221-226.
- Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ.* 1998;316:690-693.
- De Haan R, Limburg M, Bossuyt P, van der Meulen J, Aaronson N. The clinical meaning of Rankin "handicap" grades after stroke. *Stroke.* 1995;26:2027-2030.
- Wardlaw JM, del Zoppo G, Yamaguchi T. Thrombolysis for acute ischaemic stroke. *Cochrane Database Syst Rev.* 2000;2:CD000213.
- Guyatt GH, Nogradi S, Halcrow S, Singer J, Sullivan MJ, Fallen EL. Development and testing of a new measure of health status for clinical trials in heart failure. *J Gen Intern Med.* 1989;4:101-107.
- Lydick E. Approaches to the interpretation of quality-of-life scales. *Med Care.* 2000;38(9, suppl):III180-III183.
- Ware JE Jr, Keller SD. Interpreting general health measures. In: Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials.* 2nd ed. Philadelphia, Pa: Lippincott-Raven Publishers; 1996:445-460.
- Thompson MS, Read JL, Hutchings HC, Paterson M, Harris ED Jr. The cost effectiveness of auranofin: results of a randomized clinical trial. *J Rheumatol.* 1988;15:35-42.
- Brooks WB, Jordan JS, Divine GW, Smith KS, Neelon FA. The impact of psychologic factors on measurement of functional status: assessment of the sickness impact profile. *Med Care.* 1990;28:793-804.
- Deyo RA, Inui TS, Leininger JD, Overman SS. Measuring functional outcomes in chronic disease: a comparison of traditional scales and a self-administered health status questionnaire in patients with rheumatoid arthritis. *Med Care.* 1983;21:180-192.
- McSweeney AJ, Grant I, Heaton RK, Adams KM, Timms RM. Life quality of patients with chronic obstructive pulmonary disease. *Arch Intern Med.* 1982;142:473-478.
- Fletcher A, McLoone P, Bulpitt C. Quality of life on angina therapy: a randomised controlled trial of transdermal glyceryl trinitrate against placebo. *Lancet.* 1988;2:4-8.
- Kendall PC, Marrs-Garcia A, Nath SR, Sheldrick RC. Normative comparisons for the evaluation of clinical significance. *J Consult Clin Psychol.* 1999;67:285-299.
- Jacobson NS, Roberts LJ, Berns SB, McGlinchey JB. Methods for defining and determining the clinical significance of treatment effects: description, application, and alternatives. *J Consult Clin Psychol.* 1999;67:300-307.
- King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res.* 1996;5:555-567.
- Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care.* 1981;19:787-805.
- Brook RH, Ware JE Jr, Rogers WH, et al. Does free care improve adults' health? results from a randomized controlled trial. *N Engl J Med.* 1983;309:1426-1434.
- Kravitz RL, Greenfield S, Rogers W, et al. Differences in the mix of patients among medical specialties and systems of care: results from the medical outcomes study. *JAMA.* 1992;267:1617-1623.
- Mossey JM, Shapiro E. Self-rated health: a predictor of mortality among the elderly. *Am J Public Health.* 1982;72:800-808.

31. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407-415.
32. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*. 1998;16:139-144.
33. Santanello NC, Zhang J, Seidenberg B, Reiss TF, Barber BL. What are minimal important changes for asthma measures in a clinical trial? *Eur Respir J*. 1999;14:23-27.
34. Guyatt GH, Berman LB, Townsend M, Pugsley SO, Chambers LW. A measure of quality of life for clinical trials in chronic lung disease. *Thorax*. 1987;42:773-778.
35. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol*. 1994;47:81-87.
36. Juniper EF, Guyatt GH, Griffith LE, Ferrie PJ. Interpretation of rhinoconjunctivitis quality of life questionnaire data. *J Allergy Clin Immunol*. 1996;98:843-845.
37. Juniper EF, Guyatt GH, Feeny DH, Ferrie PJ, Griffith LE, Townsend M. Measuring quality of life in children with asthma. *Qual Life Res*. 1996;5:35-46.
38. Osoba D, Brada M, Yung WK, Prados M. Health-related quality of life in patients treated with temozolomide versus procarbazine for recurrent glioblastoma multiforme. *J Clin Oncol*. 2000;18:1481-1491.
39. Osoba D, Tannock IF, Ernst DS, Neville AJ. Health-related quality of life in men with metastatic prostate cancer treated with prednisone alone or mitoxantrone and prednisone. *J Clin Oncol*. 1999;17:1654-1663.
40. Barber BL, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. *Qual Life Res*. 1996;5:117-122.
41. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol*. 1997;50:869-879.
42. Osoba D, Levin V, Yung WKA, et al. Health-related quality-of-life (HRQL) benefits in patients with recurrent anaplastic astrocytoma (AA) treated with temozolomide (TEM) [abstract]. *Proc Am Soc Clin Oncol*. 1998;17:3889. Abstract 1496.
43. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. *J Eval Clin Pract*. 2000;6:39-49.
44. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements: an illustration in rheumatology. *Arch Intern Med*. 1993;153:1337-1342.
45. Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the Six Minute Walk test in chronic lung disease patients. *Am J Respir Crit Care Med*. 1997;155:1278-1282.
46. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol*. 1996;49:1215-1219.
47. Redelmeier DA, Guyatt GH, Goldstein RS. On the debate over methods for estimating the clinically important difference. *J Clin Epidemiol*. 1996;49:1223-1224.
48. Best WR, Beckett JM. The Crohn's disease activity index as a clinical instrument. In: Pena AS, Weterman IT, Booth C, Strober W, eds. *Developments in Gastroenterology: Recent Advances in Crohn's Diseases*. Dordrecht, the Netherlands: Martinus Nijhoff; 1981:7-12.
49. Testa MA, Lenderking WR. Interpreting pharmacoeconomic and quality-of-life clinical trial data for use in therapeutics. *Pharmacoeconomics*. 1992;2:107-117.
50. Testa MA, Anderson RB, Nackley JF, Hollenberg NK. Quality-of-Life Hypertension Study Group. Quality of life and antihypertensive therapy in men: a comparison of captopril with enalapril. *N Engl J Med*. 1993;328:907-913.
51. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. *Med Care*. 1995;33(4, suppl):AS286-AS291.
52. Deyo RA, Inui TS, Leininger J, Overman S. Physical and psychosocial function in rheumatoid arthritis: clinical use of a self-administered health status instrument. *Arch Intern Med*. 1982;142:879-882.
53. Goldman L, Hashimoto B, Cook EF, Loscalzo A. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: advantages of a new specific activity scale. *Circulation*. 1981;64:1227-1234.
54. Brant R, Sutherland L, Hilsden R. Examining the minimum important difference. *Stat Med*. 1999;18:2593-2603.
55. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis*. 1986;39:897-906.
56. Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. *J Clin Epidemiol*. 2000;53:279-284.
57. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther*. 1998;78:1186-1196.
58. Walter SD, Irwig L. Estimating the number needed to treat (NNT) index when the data are subject to error. *Stat Med*. 2001;20:893-906.
59. Goldstein RS, Gort EH, Guyatt GH, Feeny D. Economic analysis of respiratory rehabilitation. *Chest*. 1997;112:370-379.
60. Norman GR, Sridhar FG, Guyatt GH, Walter SD. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Med Care*. 2001;39:1039-1047.
61. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
62. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care*. 1989;27(3, suppl):S178-S189.
63. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. *Pharmacoeconomics*. 1999;15:141-155.
64. Wyrwich KW, Nienaber 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;37:469-478.
65. 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;52:861-873.
66. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res*. 1995;4:293-307.
67. Hebert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. *Arch Phys Med Rehabil*. 1997;78:1305-1308.