Some Considerations for the Interpretation of Health-Related Quality of Life Data

To the Editor—Marquis et al. [1] recently proposed a stepwise strategy for improving the understanding and interpretation of health-related quality of life (HRQL) measures. The strategy consists of two mandatory steps: 1) understanding the content of the scale and its psychometric properties (e.g., reliability and construct validity), and 2) evaluating the magnitude of the change, or of the treatment difference, and its statistical significance through calculation of the effect size or standardized response mean to demonstrate responsiveness. The authors further suggest a supplemental step of selecting at least one of three approaches to understand the clinical significance of changes in HRQL scores: 1) determining the minimally important difference (MID), standard error of measurement (SEM) or score calibration using a global rating of change or a clinical anchor; 2) comparing baseline or follow-up scores with norms or available known group references; or 3) collecting data to understand the practical value of scores, based on the relationship to outcomes such as morbidity and death, patient behavior (compliance and resource utilization) and consequences on work (loss of productivity or working days). The above methods are intended to help clinicians and researchers understand the clinical relevance of a change in score for a given HRQL measure.

While this approach is clearly presented in relevant steps that appear to depict the state of the field and the general concepts follow those proposed by Lydick and Epstein [2], caution should be taken in proposing these as standards without other important considerations.

For example, the observed effect size obviously depends on the level of change found in a particular study, and hence, the inherent effects of specific interventions according to the characteristics of the population being studied (for example, disease severity or age). In addition, the interpretation of effect sizes may vary widely according to condition; small effect sizes may be meaningful to patients with a certain disease, whereas in other conditions, only moderate to large effect sizes may be important to patients.

When using a global rating of change or a clinical anchor, it should measure a related construct and correlate to some degree with the questionnaire of interest. The strength of the relationship of the HRQL scores with the global rating of change or clinical anchor should be characterized before calibrating the HRQL against such measures, and in subsequent interpretation. Intuitively, one might have greater confidence in the calibration where a stronger relationship is demonstrated between the anchor and the questionnaire of interest compared to a weaker relationship. This criterion is similar to that used by Guyatt et al. [3] in which HRQL was anchored to global transition ratings for change in disease state.

The method of calibrating measures has also been proposed to help understand the change in HRQL scale scores in terms of life event changes (for example, loss of a job or divorce), which are presumably more understandable than changes anchored to other measures. The example cited in Marquis et al. calibrated changes in HRQL scores to a Life Event scale [4]. Based on the calibration method, “a change of 0.15 units was associated with a 55-point change on the Life Events Index, corresponding to the impact on quality of life that single events, such as a major personal injury or illness (53 points), the death of a close family member (63 points), or a dismissal from work (47 points), might have on a person.” Nevertheless, there were no differences on the Life Events scale reported during the conduct of the study, nor were differences found on efficacy or safety parameters measured. While cross-sectional correlations were discussed (highest at week 8; range 0.36–0.57), there were no longitudinal correlations provided to examine the strength of the relationship for change in the HRQL and the Life Events scale used for the calibration (correlating changes in both scales over the same period of time). Therefore, it is difficult to know how strongly the changes in the HRQL and Life Events scales were correlated. Apart from using such a Life Events scale, linking to outcomes such as mortality or morbidity would necessitate large, long-term trials which are not typically conducted in traditional clinical trial programs.

As for MID, in cases where the correlations between the anchor and the questionnaire of interest may be low, there may be a great deal of variability in the distribution of the categories of the anchor. This results in a misleading estimation of MID. An additional step may be needed to ensure
that the anchor is sufficiently related to the questionnaire to warrant such an approach. As others have expressed, the estimated magnitude of any MID will inherently depend on the distribution and measurement properties of the external anchor [3]. The authors appropriately note the paradox of using a single, nonvalidated item to calibrate a validated multi-item questionnaire; the pitfalls of such an approach are well-known [1].

Intuitively, the difference in the MID may vary across a wide range of severity and tends to be greater for populations with more severe disease states. As Hays and Woolley suggest, the meaning of change may depend on where you start [5]. Thus, the MID may vary by the population being studied [6]. Likewise, the MID would be dependent on the distribution of the questionnaire domain scores as well. Finally, the degree of change may depend on the direction of change, and may not be of consistent magnitude for those who deteriorate and those who improve. In addition, insufficient consideration has been given to approaches aimed at documenting the lack of disease progression in cases where the goal of treatment is to maintain function while the untreated group is expected to deteriorate.

For the supplemental step in determining clinical significance, collecting data in relevant clinical severity groups and in nonsymptomatic groups may be useful, but the well-known limitations of using cross-sectional data as a basis for interpreting longitudinal change still apply. Likewise, one must be cautious about the representativeness of the groups selected on which the interpretation is being based, and as the authors note, it may be difficult to generate reference data for the general population. Finally, the understanding of the practical value of scores is certainly a laudable goal, but determining the practical value in the manner in which the authors propose may not be feasible for a new disease-specific measure being included in a clinical trial program, particularly in new therapeutic areas.—Josephine M. Norquist, Cindy J. Girman, and Nancy C. Santanello, Department of Epidemiology, Merck Research Laboratories, West Point, PA, USA.

References