

Insights and Limitations from Health-Related Quality-of-Life Research

In this issue of *JGIM*, Mangione and colleagues report a study designed to examine how well the 36-item Short Form of the Medical Outcomes Study (SF-36) performs and simultaneously how quality-of-life evolves after three different surgical procedures.¹ The competing nature of these goals illustrates the evolution of health status measurement and its limitations and potential for providing important information to clinicians.

The appropriateness of the term health-related quality of life (HRQL), which Mangione and colleagues use to describe what they are measuring, remains controversial. The key instrument they have chosen, the SF-36, focuses to a large extent on how patients are functioning, including their ability to take care of themselves and carry out their usual roles in life. Although this pragmatic view of HRQL seems to have gained ascendancy, there remain those who argue that unless investigators tap into individual patient values they are measuring only health status—they are not measuring HRQL.²

This issue can be clarified by thinking of a woman with posttraumatic quadriplegia who, despite her limitations, is happy and fulfilled and values her life highly (more, for instance, than most people, and more than she did before she developed quadriplegia). Most of this woman's SF-36 results would suggest a poor HRQL, despite the high value she places on her health state. My own view is that HRQL is an appropriate label for what Mangione and her colleagues have measured because there is a consensus of values both within and between cultures for the basic human functions included in the SF-36.

It is more difficult, however, to determine whether the SF-36 really measures HRQL. Because there is no criterion or "gold standard" for HRQL, it is a challenge to determine whether any HRQL measure is tapping into the intended aspect of peoples' experience. The most convincing approach to establishing whether an instrument is really measuring what it is designed to measure (the technical term is "validity") is for investigators to make predictions about the results they expect before they collect the data. Without such predictions, it is easy for investigators to rationalize their findings, whatever the results.

Although some of the predictions made by Mangione and colleagues proved accurate, others did not. For instance, their results did not show the deterioration in emotional function or health perceptions that the investigators anticipated 1 month postoperatively in lung cancer patients. Also, correlations between change in the Specific Activity Scale, which is a measure of cardiovascular physical function, and change in several SF-36 domains related to physical functioning were lower than predicted. We might interpret these findings as reflecting limitations in the validity of the SF-36. Alternatively, we might interpret them as limitations in the investigators' understanding of

the course of emotional function in lung cancer and problems with the validity of the Specific Activity Scale. To put the dilemma more vividly: Have the investigators discovered something we didn't know about how lung cancer patients feel after surgery, or have they discovered a limitation in the SF-36's ability to measure emotional function?

Because most of the investigators' predictions proved accurate, I am inclined to share their view that the study's results provide strong support for the validity of the SF-36. Nevertheless, the discrepancy between predictions and findings highlights the challenges of measurement in an area without a criterion standard for HRQL.

In their investigation, the authors focused on the ability of the SF-36 to measure change—what we have called its evaluative function.³ This contrasts with the discriminative function of the instrument, which is its ability to differentiate between those with a better and those with a worse quality of life at a point in time. Therefore, in studying the instrument's validity, the investigators have correlated changes in the SF-36 with changes in other measures but have not calculated correlations between different measures at a single point in time.

The second key property of an evaluative measure is its ability to pick up important changes in HRQL, even if those changes are small. Although Mangione and colleagues report the ability of the SF-36 to detect change (the technical term is "relative responsiveness"), they provide little information about the size of the changes they observed. Thus, whether the SF-36 can pick up small but important changes in physical or emotional function in these populations remains uncertain.

The responsiveness to small but important changes would be important if investigators used HRQL measures in the context suggested by the final paragraph of Mangione's discussion: randomized trials of surgery versus alternative treatments. The SF-36 is an example of a generic measure that tries to cover all important areas of HRQL. In contrast, specific measures focus on groups of patients with similar issues and explore areas of particular relevance in more detail.⁴ For example, a specific instrument for patients with hip osteoarthritis would focus on pain and mobility, while one for lung cancer patients might focus on dyspnea and fatigue.

Theoretically, specific HRQL measures are more responsive than generic HRQL measures, and accumulating data from head-to-head comparisons in randomized trials suggest that this is the case.⁵⁻¹⁰ Responsiveness to small but important changes is likely to matter if we compare different types of hardware for hip arthroplasty or different resection strategies in lung cancer. Indeed, investigators have chosen a disease-specific measure of HRQL as the primary outcome in several ongoing randomized trials of lung volume reduction surgery.

Because the area of HRQL measurement is relatively new, clinicians may find it difficult to interpret studies in which HRQL is an important outcome. To help them, we have suggested a set of guidelines for evaluating HRQL studies.¹¹ Mangione's report meets most of our criteria. The investigators measured aspects of patients' lives that patients consider important, and their instruments worked as intended. They have shown that the SF-36 is able to detect changes as patients go through surgery (whether the SF-36 can detect the change patients would experience with alternative management remains uncertain). The investigators have not omitted any important areas of HRQL. They did not attempt an economic assessment, which would have required other measurement instruments.

We are left, however, with some uncertainty about the size of the HRQL changes that patients experienced. Mangione and colleagues correctly point out that population norms help us understand the impact of surgery on HRQL as measured by the SF-36. Still, the meaning of the deterioration in, for instance, role physical function at 1 month after surgery is not evident. Full understanding of whether the observed changes in SF-36 scores represent trivial changes in HRQL, small but important changes in HRQL, or large changes in HRQL (the technical term is "interpretability")⁴ remains a research challenge.—

GORDON GUYATT, MD, *McMaster University, Hamilton, Ontario, Canada.*

REFERENCES

1. Mangione CM, Goldman L, Orav EJ, et al. Health-related quality of life after elective surgery. *J Gen Intern Med.* 1997;12:686-697.
2. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA.* 1994;272(8):619-26.
3. Kirschner B, Guyatt GH. A methodologic framework for assessing health indices. *J Chron Dis.* 1985;38:27-36.
4. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life: basic sciences review. *Ann Intern Med.* 1993;70:225-30.
5. Tandon PK, Stander H, Schwarz RP Jr. Analysis of quality of life data from a randomized, placebo controlled heart-failure trial. *J Clin Epidemiol.* 1989;42:955-62.
6. Smith D, Baker G, Davies G, Dewey M, Chadwick DW. Outcomes of add-on treatment with Lamotrigine in partial epilepsy. *Epilepsia.* 1993;34:312-22.
7. Chang SW, Fine R, Siegel D, Chesney M, Black D, Hulley SB. The impact of diuretic therapy on reported sexual function. *Arch Intern Med.* 1991;151:2402-8.
8. Tugwell P, Bombardier C, Buchanan WW, et al. Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. *Arch Intern Med.* 1990;150:59-62.
9. Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. *Control Clin Trials.* 1991;12:168-79S.
10. Goldstein RS, Gort EH, Guyatt GH, Stubbing D, Avendano MA. Prospective randomized controlled trial of respiratory rehabilitation. *Lancet.* 1994;344:1394-7.
11. Guyatt GH, Naylor CD, Juniper EL, Heyland DK, Jaeschke RZ, Cook DJ. Users' guides to the medical literature, XII: how to use articles about health-related quality of life. *JAMA.* 1997;277:1232-7.