Patient-reported Outcomes in Cancer: A Review of Recent Research and Policy Initiatives

Joseph Lipscomb, PhD; Carolyn C. Gotay, PhD; Claire F. Snyder, PhD

ABSTRACT There is growing recognition that patient-reported outcome (PRO) measures—encompassing, for example, health-related quality of life—can complement traditional biomedical outcome measures (e.g., survival, disease-free survival) in conveying important information for cancer care decision making. This paper provides an integrated review and interpretation of how PROs have been defined, measured, and used in a range of recent cancer research and policy initiatives. We focus, in turn, on the role of PRO measurement in the evaluation and approval of cancer therapies, the assessment of cancer care in the community, patient-provider decision making in clinical oncology practice, and population surveillance of cancer patients and survivors. The paper concludes with a discussion of future challenges and opportunities in PRO measure development and application, given the advancing state of the science in cancer outcomes measurement and the evolving needs of cancer decision makers at all levels. (CA Cancer J Clin 2007;57:278–300. © American Cancer Society, Inc., 2007.)

INTRODUCTION

For diseases that are often chronic and sometimes incurable, with interventions that can have toxic and long-term consequences, it is especially important that decisions influencing patient outcomes reflect the patient’s own perspective. Cancer provides a compelling case in point.

The principal means for treating cancer—surgery, chemotherapy, radiation, and hormonal therapy—are frequently very effective in stopping tumor progression, reducing cancer-attributable pain and discomfort, extending life, and in many instances, curing the disease. However, all such therapies come with the risk of substantial side effects. Some are short-term and time-limited, others are long-term and persistent, and still others arise only years after the initial cancer treatment. Traditional biomedical outcome measures, particularly survival and disease-free survival, remain indisputably of central importance in cancer decision making. But there has been growing recognition that patient-reported outcome (PRO) measures—including, in particular, measures of health-related quality of life (HRQOL)—can convey important additional information for assessing the overall burden of cancer and the effectiveness of interventions.

Informal affirmation that PROs “matter” in cancer decision making is registered whenever a provider asks patients how they have been feeling, whether they have been fatigued or in pain, whether they have been able to carry on with the usual activities of life, whether they have required caregiver support, whether they have been well enough to stick with their prescribed therapy, and so on.

Concrete indicators that support the importance of PROs in the cancer sphere can be found in a variety of recent US research and policy-related developments. See Table 1 for a summary of some leading public, private, and mixed public–private sector initiatives that revolve around the application of PROs in cancer. The efforts noted in Table 1 have virtually all come into being since 2000, suggesting a growing interest in bringing “the patient’s perspective” to cancer decision making. Taken together, such initiatives are leading to a clearer understanding of the current strengths and limitations of using PRO measures in cancer.

Disclosure: Dr. Lipscomb’s work on this article was partially supported by a grant from the Georgia Cancer Coalition’s Distinguished Cancer Clinicians and Scientists program.
The purpose of this paper is to provide an integrated review and interpretation of how PROs have been defined, measured, and used in these recent cancer research and policy initiatives, with an eye also to future decision-relevant applications.

In the next section, we examine in some detail how PROs may be defined and measured for cancer outcomes assessment. In a given patient-provider encounter, the patient’s response to a simple inquiry of “how do you feel today?” will likely provide valuable qualitative information for treatment evaluation and planning. However, for research studies to evaluate the impact of specific interventions on PROs, for population surveillance of progress against the cancer burden, and for systematic and interpretable PRO data to inform patient-provider decision making, one needs high-quality quantitative measures of PROs. Thus, this section looks briefly at how HRQOL-oriented PRO measurement instruments are developed, applied in practice, and evaluated in terms of performance in the field.

The succeeding 4 sections of the paper examine the actual and potential role of PROs in, respectively, the evaluation and approval of cancer therapies, the assessment of cancer care in the community, patient-provider decision making in clinical oncology practice, and population surveillance of cancer patients and survivors. The final section identifies future challenges and opportunities in PRO measure development and application in light of advances in the state of the science in cancer outcome measurement and the evolving needs of decision makers.

**DEFINING AND MEASURING PROS IN CANCER**

**What is a PRO?**

As discussed in Acquadro et al., which summarizes the findings of an ad hoc task force formed by several professional organizations (see Note 1) to encourage harmonization of health outcomes review criteria within and across US and European regulatory agencies, the term PRO itself seems to have come into frequent usage only in the post-2000 period. Although the task force began with a focus on HRQOL, the discussion expanded over time “to include any outcome based on data provided by patients or patient proxy as opposed to data provided by other sources (including providers and caregivers)…The FDA proposed the term ‘patient-reported outcome’ (PRO) to represent these types of outcomes in the regulatory review
process... The working group subsequently became known as the PRO Harmonization Group."¹⁵ The task force identified several types of measures that fall under the PRO umbrella, including HRQOL, functional status, symptom status, overall well-being, satisfaction with care, and treatment adherence.

**What is HRQOL?**

Over the past 40 years, the published literature dealing with “quality of life” in “cancer or neoplasms” has grown substantially. A Medline search crossing these 2 terms yields over 12,500 English-language citations over the period of 1966 to 2006, with about 92% occurring from 1990 onward (1,382 over 1990 to 1994; 2,866 over 1995 to 1999; 5,236 over 2000 to 2004; and 2,063 over 2005 to 2006). In line with this growth in the literature, there has been increasing discussion about how HRQOL in cancer is defined and how to measure it. In 1993, Aaronson et al¹⁶ noted the "growing interest in broadening the evaluation criteria employed in cancer clinical trials beyond traditional biologic markers of therapeutic outcome—tumor response, time to progression, and disease-free and overall survival—to include an assessment of the impact of the disease and its treatment on the physical, psychological, and social functioning of the patient." In a comprehensive assessment of conceptual models of HRQOL, Ferrans¹⁷ identified a range of HRQOL definitions, including this one by Cella¹⁸: "the extent to which one's usual or expected physical, emotional, and social well-being is affected by a medical condition and/or its treatment.”

After evaluating hundreds of published applications of quality-of-life measures in cancer, the National Cancer Institute (NCI)’s Cancer Outcomes Measurement Working Group (COMWG) concluded that the distinguishing features of an HRQOL measure are that it is patient-reported and that it involves the patient’s subjective assessment or evaluation of important aspects of his or her well-being.¹⁹ An implication is that all HRQOL measures are PRO measures, but there are PRO measures that have little or no evaluative component and, thus, would not qualify as HRQOL measures. For example, a simple patient report on the presence or absence of a symptom such as nausea may require some subjective interpretation on the respondent’s part, but it conveys little or no information about the impact of the symptoms on functioning or other aspects of well-being. Specifically, the COMWG defined HRQOL measurement to include patient assessments of symptom impact, functional status, and/or global well-being.

**Symptom measures** that would qualify as HRQOL measures thus report not only the existence or frequency, but also the severity, bother, or other impacts of symptoms, including both disease-related and treatment-induced. Well-known examples include the Rotterdam Symptom Checklist,²⁰ which encompasses multiple aspects of symptom effects (psychological distress, physical distress, and disease-specific symptoms), and the Brief Pain Inventory,²¹ which focuses expressly on one prominent symptom area. The widely used Common Terminology Criteria for Adverse Events,²² which are based on patient interview and/or laboratory data, but are physician-recorded and reported, are distinct from PROs and would not qualify as HRQOL measures.

**Functional status measures** that are intended to capture only one dimension (that is, one definable aspect or domain) of HRQOL are termed unidimensional; an example is the Beck Depression Inventory.²³ In point of fact, most functional status measures of HRQOL are multidimensional, designed to reflect multiple domains of impact. The specific domains of focus vary by instrument, but often include physical, psychological, and social components of outcome. Prominent examples include the European Organization for Research and Treatment of Cancer (EORTC) QLQ-C30,¹⁶ the Functional Assessment of Cancer Therapy General (FACT G),²⁴ the Health Utilities Index (HUI),²⁵ and the EQ-5D Health Questionnaire.²⁶

Despite the fact that all of these questionnaires purport to measure HRQOL, there are sharp distinctions in the conceptualization, construction, and intended application among different multidimensional instruments. One basic difference is between (1) measures based on psychometric science, in which an individual indicates his or her HRQOL response along a
subjective scale of well-being (eg, the EORTC QLQ C30, the FACT G, and indeed, the majority of HRQOL functional status measures applied in cancer to date); and (2) measures based on the science of economic evaluation in health care in which respondents supply a relative value (or utility) rating for a given HRQOL state in comparison with designated anchor states (often the “best” and “worst” states the individual can imagine) (eg, the HUI and EQ-5D; see Feeny27). The latter type of measure is generally termed “preference-based,” while the psychometrically oriented measures are sometimes referred to as “nonpreference-based.” These distinctions give rise to practical differences in how HRQOL summary scores are derived and used.

For the psychometric-based measures, an individual’s HRQOL score is based on the specific survey items endorsed, which are transformed to scale scores indicating the relative degree of functioning or well-being the individual reports along each posited HRQOL dimension. That is, one is attempting to pinpoint the individual’s “location” along the measurement scale corresponding to each HRQOL dimension. In this sense, the score is nonpreference-based since there is no explicit attempt otherwise to compute a utility value for the individual’s scale location. Summary scores for nonpreference-based multidimensional HRQOL measures may be reported in terms of a profile of (unidimensional) scale scores (eg, the EORTC QLQ-C3016) or, in addition, an overall summary score (eg, FACT G24), as will be illustrated below.

For the economic-based measures, an individual’s HRQOL score basically reflects 2 categories of information: specific levels of functioning along each of the posited dimensions, as indicated by which survey items are endorsed; and the relative value or utility weight assigned to each of these levels of functioning. These utility weights may be assigned by the individual directly, or (as is more commonly the case) they may be imputed to the individual based on community health state preference surveys. By combining both categories of information, an overall preference-based HRQOL score is assigned to the individual (as will be illustrated below). Global well-being measures of HRQOL capture the individual’s overall assessment of well-being or happiness in a summary score or indicator typically based on a single (global) question. A frequently used psychometric global HRQOL measure asks the individual to indicate whether his or her health is Excellent, Very Good, Good, Fair, or Poor (the “E-VG-G-F-P” scale). A common preference-based measure is obtained by asking the individual simply to rate his or her overall health or well-being numerically on a scale that includes an explicit comparative standard (eg, a 0 to 100 “visual analog scale” where 100 represents the best possible health level and 0 the worst). Note that global approaches do not deny that HRQOL may be multidimensional. Rather, such measures require the individual to engage in a holistic evaluation that effectively aggregates across whatever dimensions are (implicitly) important to him or her.

Additionally, HRQOL measures in cancer may be classified as either generic, general cancer, or cancer site-specific or cancer problem-specific. A generic HRQOL measure can be applied to a range of diseases and conditions that may be, but need not be, cancer related. Examples include the Medical Outcomes Study Short Form (SF)-3628,29 and a number of other psychometrically based measures (see Erickson30), as well as the EQ-5D, HUI, and indeed, most all preference-based measures (see Feeny27). A general cancer measure is intended for application across the full range of cancer-related events, regardless of the patient’s tumor type; among the many examples reviewed by the COMWG are the FACT G and the EORTC QLQ-C30. A cancer site-specific or cancer problem-specific HRQOL measure is tailored, respectively, to a particular tumor type (eg, EORTC-QLQ-BR2331 for breast cancer), problem area (eg, FACT N32 for febrile neutropenia associated with adjuvant chemotherapy), or treatment modality (eg, FACT BRM33 for treatment with biologic response modifiers such as interferon).

That PRO measures generally, and HRQOL measures in particular, can play multiple important roles in cancer intervention assessment and decision making was emphasized in a recent Journal of the National Cancer Institute Monograph34 and is a recurring theme in this paper. In particular, multidimensional (nonpreference-based) HRQOL measures have been employed...
extensively to evaluate cancer interventions in clinical trials and in community settings. Preference-based HRQOL measures have been less frequently embraced in these contexts, but are often used in cost-effectiveness analyses of cancer interventions. As discussed here later, applications of PRO measures in clinical oncology practice and in population surveillance of the cancer burden are at early stages of development, and we are just beginning to understand what types of PRO formulations might be most appropriate and feasible in these settings.

Finally, to provide a concrete feel for how HRQOL scores are derived, we illustrate the application of a multidimensional, psychometrically based, cancer site-specific instrument—namely, the FACT Colorectal (C)—in Figure 1 and a multidimensional preference-based instrument—namely, the EQ-5D—in Figure 2. In each case, we compute the HRQOL score for a hypothetical patient undergoing chemotherapy following surgery for Stage III colorectal cancer.


The sequence of tasks for developing a non-preference-based HRQOL instrument (the kind most frequently used in cancer applications) has been summarized by Juniper et al. In the development stages, one (a) identifies the goals of measurement; (b) generates candidate instrument items (often using patient interviews, focus groups, consultation with providers, and/or reviews of the literature); (c) pares down the items to a parsimonious set (based on further expert consultation and/or on statistical approaches such as factor analysis); and (d) designs and constructs the instrument, with attention not only to item content, but to additional matters such as specification of the response options (e.g., whether to use a simple “yes-no” format or a multilevel Likert scale), the time period encompassed by the questions (e.g., “over the past 7 days”), and the mode of administration (e.g., self-completed versus interviewer-administered). To launch the testing stage, one often conducts small pretests to determine whether respondents correctly interpret the items, whether the questionnaire’s format is suitable, and if the instructions are clear. Further investigations using larger samples seek to confirm and improve as needed the instrument’s reliability, validity, responsiveness, interpretability, and feasibility of administration.

The process for developing a preference-based HRQOL instrument includes essentially the steps above for ensuring appropriate dimension-specific item content, but it generally also includes the challenging task of deriving utility weights associated with each item on each dimension scale. For discussions of how this has been accomplished for the major preference-based HRQOL measurement systems, see Feeny et al. regarding the HUI, Kaplan et al. regarding the Quality of Well-Being (QWB) index, and Shaw et al. regarding the new, US-based utility weights for the EQ-5D.

**Evaluating the Performance of PRO Instruments: The Medical Outcomes Trust (MOT) Framework**

How might one evaluate the technical quality and appropriateness of an instrument for measuring PROs? To guide its assessment of PRO measures, NCI’s COMWG adopted the framework of attributes and criteria developed by the Scientific Advisory Committee of the nonprofit MOT. The MOT attributes for judging the psychometric performance of health status and quality-of-life measures generally are shown in Table 2 (as adapted from the MOT and Lipscomb et al.). The important point here is that standardized criteria to evaluate PRO questionnaires have been developed, are widely accepted, and can be used to assess the merit of the measurement tools used in cancer.

**PROS IN THE EVALUATION AND APPROVAL OF CANCER INTERVENTIONS**

**Application in Randomized Clinical Trials**

While there are multiple potential uses of PRO measures in cancer, the primary area of application has been in randomized clinical trial (RCT)-investigated interventions to treat, screen for, or prevent cancer, or to manage disease symptoms. Correspondingly, recent efforts to evaluate the state of the science in cancer outcomes measurement have focused principally on the use of
Functional Assessment of Cancer Therapy—Colorectal (FACT C) HRQOL Instrument—Completed and Scored for a Hypothetical Patient with Stage III Disease. This Version 4 of the FACT C courtesy of David Cella, PhD, Northwestern University, the instrument's developer.

\[ \text{FIGURE 1} \]
Instructions to respondent: “By placing a checkmark in one box in each group below, please indicate which statements best describe your own health state today.”

<table>
<thead>
<tr>
<th>Mobility (M)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems in walking about (Level 1)</td>
<td>☐</td>
</tr>
<tr>
<td>I have some problems in walking about (Level 2)</td>
<td>☐</td>
</tr>
<tr>
<td>I am confined to bed (Level 3)</td>
<td>☐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self Care (SC)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems with self-care (Level 1)</td>
<td>☐</td>
</tr>
<tr>
<td>I have some problems washing or dressing myself (Level 2)</td>
<td>☐</td>
</tr>
<tr>
<td>I am unable to wash or dress myself (Level 3)</td>
<td>☐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usual Activities (UA)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems with performing my usual activities</td>
<td>☐</td>
</tr>
<tr>
<td>I have some problems with performing my usual activities</td>
<td>☐</td>
</tr>
<tr>
<td>I am unable to perform my usual activities</td>
<td>☐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pain/Discomfort (PD)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no pain or discomfort</td>
<td>☐</td>
</tr>
<tr>
<td>I have moderate pain or discomfort</td>
<td>☐</td>
</tr>
<tr>
<td>I have extreme pain or discomfort</td>
<td>☐</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anxiety/Depression (AD)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I am not anxious or depressed</td>
<td>☐</td>
</tr>
<tr>
<td>I am moderately anxious or depressed</td>
<td>☐</td>
</tr>
<tr>
<td>I am extremely anxious or depressed</td>
<td>☐</td>
</tr>
</tbody>
</table>

Shaw et al. summarize the derivation of the recently developed US population-based preference weights for the EQ-5D (all previous analyses involving the EQ-5D were based on weights derived from European population samples). A stratified random sample of over 4,000 US adults were surveyed in 2002 to obtain individual preference scores for a total of 45 such states, selected by factorial design methods that would ultimately permit statistically imputing preference weights to any of the possible 243 (3^5) EQ-5D states. Each respondent assigned utility values to a strategically chosen sample of the 45 states using the frequently employed “time-tradeoff” technique. (Under this approach, the respondent’s utility weight for a given state is proportionally related to the length of stay in that state judged to be equivalent in preference to some fixed amount of time in “excellent health,” defined here as the state [1, 1, 1, 1, 1]). From these survey data, Shaw et al estimated linearly additive multiple regression models wherein the utility value assigned to a given state was assumed to be a function of the state’s dimension-level composition, with inclusion of both direct-effect and interaction terms (see Shaw et al, Appendix 2). Specifically, their “final” statistical model (selected on the basis of predictive validity) takes the following specific form, reported here because it is integral to computing an EQ-5D preference score for our colorectal cancer patient: Score = 1.000 – 0.146M2 – 0.558M3 – 0.175SC2 – 0.471SC3 – 0.374UA3 – 0.173PD2 – 0.537PD3 – 0.156AD2 – 0.450AD3 – 0.140PD – 0.011X + 0.122Y – 0.015Z, where

\[ W = \text{number of dimensions at level 2 or 3 beyond the first one at level 2} \]
\[ X = \text{square of the additional number of dimensions at level 2 beyond the first one} \]
\[ Y = \text{number of additional dimensions at level 3 beyond the first one} \]
\[ Z = Y^2 \]

For the EQ-5D state (2, 1, 2, 2, 2), the independent variables in the equation take on the following values: M2=A2=PD2=AD2=1, with W=3 and X=9; the remaining variables are equal to 0. It follows that for this colorectal cancer patient, Score=0.706.

Interpretation: Using these US-based weights, Luo et al. computed mean EQ-5D scores for a number of demographic subgroups within the original study sample. In particular, mean EQ-5D declined by age category: 0.91 for age 18 to 44 years, 0.84 for age 45 to 64 years, and 0.81 for age 65+ years. The great majority of individuals with Stage III colorectal cancer will be in the latter 2 categories. As one would expect, our computed EQ-5D score for Stage III disease, 0.706, is notably lower than for the average score for the age-comparable general population.

Deriving EQ-5D score for this hypothetical patient. In EQ-5D terminology, our colorectal cancer patient is in health state (2, 1, 2, 2, 2), indicating her levels on the instrument’s 5 dimensions (ordered as above). We now show how a preference score is assigned to this health state.

\[ \text{Score} = 1.000 - 0.146M^2 - 0.558M^3 - 0.175SC^2 - 0.471SC^3 - 0.374UA^3 - 0.173PD^2 - 0.537PD^3 - 0.156AD^2 - 0.450AD^3 - 0.140PD - 0.011X + 0.122Y - 0.015Z \]

PRO measures, particularly HRQOL, in randomized trials.

Perhaps the most comprehensive such evaluation to date was carried out by NCI’s COMWG. Created in 2001, the COMWG comprised 35 experts in outcomes assessment drawn from academia, government, industry, and the cancer patient and survivorship communities. For comprehensive discussions concerning the functioning of the working group, key findings, and lessons learned, see Lipscomb et al., Lipscomb et al., Snyder et al., and Gotay et al. Among the main points are the following:

- A variety of generic, general cancer, and cancer site-specific HRQOL instruments have demonstrated adequate reliability, validity, responsiveness, feasibility, and attention to the demands of cultural and language adaptation. However, important challenges remain in understanding and improving the interpretability of instruments, though considerable progress has been made in defining what might constitute a “minimum important difference” (MID) in HRQOL scale scores. Much work is still needed in strengthening the theoretical foundations of HRQOL measurement, especially the link between the “conceptual model,” which shows the hypothesized cause-effect relationships among outcome variables and their determinants, and the “measurement model” describing the statistical analyses to be conducted.
- Not only the quantity but also the technical quality of HRQOL applications has increased...
TABLE 2  Attributes of Health Status and Quality-of-Life Instruments as Identified by the Medical Outcomes Trust (MOT)*

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptual and measurement model</td>
<td>The conceptual model provides the rationale for the specific concepts, or domains, of importance and also their interrelationships in measuring the outcome of interest (eg, HRQOL) in a particular population (eg, breast cancer patients). The measurement model specifies the way that the domains are quantified through scores on specific questions and also whether and how domain scores are aggregated to derive a summary health status score.</td>
</tr>
<tr>
<td>Reliability</td>
<td>The degree to which an instrument is free from random error. The focus is on internal consistency (whether the items on a scale are reliably measuring the same construct) and reproducibility, either test-retest (comparable results in the same person over time) or inter-rater (comparable results across raters) reliability.</td>
</tr>
</tbody>
</table>
| Validity                           | The degree to which an instrument measures what it claims to measure, specifically:  
  - content validity (the degree to which each domain of an instrument includes a full range of questions that matches the intended use);  
  - criterion validity (the degree to which the scores from an instrument relate to a gold-standard measure of the concept);  
  - construct validity (the degree to which the measure reflects the theory behind it, ie, does the instrument actually measure what it purports to assess? In principle, construct validity examines the relationship between the conceptual model, the corresponding measurement model, and the data obtained. In practice, the construct validity of an instrument is typically examined by measuring hypothesized relationships between instrument scores and other variables thought to be related, eg, whether there is a positive association between the scores on a PRO scale measuring nausea and the receipt of chemotherapy). |
| Responsiveness                     | Refers to the ability of an instrument to detect outcome changes over time. By convention, sensitivity refers to the ability of an instrument to detect differences in a cross-section of respondents at a given point in time. |
| Interpretability                   | The degree to which readily understood meaning can be attached to the quantitative scores from an instrument for either an individual or a group.                                                                 |
| Burden and alternative modes of administration† | Burden refers to the time, effort, and other demands on those to whom the instrument is administered or on those who administer it; modes of instrument administration include patient self-report through traditional communication channels (questionnaires, telephone); interviewer administered; and computer-assisted approaches, such as computer-adaptive testing using item banks. |
| Cultural and language adaptations  | The degree to which an instrument that has been translated into another language or cultural setting is conceptually and linguistically equivalent to the original; such equivalence is generally evaluated by comparing the instrument’s various measurement properties. |

*Adapted from the original presentation by the Scientific Advisory Committee of the Medical Outcomes Trust39 with permission from Quality of Life Research.
†The distinct MOT attributes of burden and alternative modes are discussed together here because they frequently are interrelated.

in recent years, with newer investigations more likely to have used well-validated instruments and study designs better suited for PRO assessment.

- The statistical analysis of PRO data poses no greater problems fundamentally than arise with more traditional biomedical endpoints.
- Missing data in studies with HRQOL endpoints tend to bias treatment comparisons unless a convincing statistical corrective can be identified, which is generally difficult. Thus, the best “cure” for missing data is prevention through careful study design and execution.43
- Most PRO applications in cancer have focused on the 4 highest incidence tumor sites (breast, colorectal, lung, and prostate), and greater attention is needed to other cancers.
- Comparatively few studies provide head-to-head comparisons of the performance of seemingly similar, high-quality PRO measures. Such comparative studies could provide a clearer picture of the similarities and differences of measures when applied within a given population and help inform the choice of measure for research studies and other applications.
- New approaches to questionnaire development and analysis grounded in modern measurement techniques such as item response theory (IRT) modeling hold considerable promise for improving the scientific soundness of PRO assessment. We return to this theme in the final section.

The circumstances under which HRQOL measures bring significant information value
to outcomes assessment over and above that provided by traditional biomedical endpoints need to be identified, particularly when the study’s primary endpoint is survival or disease-free survival. For purposes of COMWG deliberations, HRQOL measures were defined as providing added value when these measures were instrumental in interpreting a study’s findings and would be expected to influence clinical recommendations. The COMWG authors’ judgments were primarily based on their study-by-study assessments of HRQOL findings compared with those for biomedical outcomes. The specific observations are as follows:

• Ganz and Goodwin concluded that the value added of HRQOL measures in breast cancer treatment assessment varies importantly with the type of study. HRQOL measurement can play an important, even pivotal, role in studies where biomedical outcomes are essentially equivalent (such as breast-conserving surgery versus mastectomy in the primary management of breast cancer); that evaluate psychosocial interventions; or that examine the late (downstream) effects of therapy. On the other hand, HRQOL data may be of secondary importance in comparing interventions that are curative in intent and differ substantially in biomedical outcomes, especially survival, as may be found in the adjuvant setting. Ganz and Goodwin also noted that biomedical endpoints may be sufficient for evaluating treatments for metastatic breast cancer when there are substantial differences in toxicity and other clinical outcomes.

• Litwin and Talcott emphasized that HRQOL may be particularly important for decision making in prostate cancer since there is still not decisive evidence about the survival advantages of the leading therapeutic options. They reported that across a range of treatment studies, prostate cancer-specific measures tend to have greater sensitivity for detecting HRQOL changes compared with generic measures.

• Earle and Weeks found that the information value of HRQOL in lung cancer treatment studies depends substantially on the type of investigation. In curative treatment, HRQOL effects are frequently transient and may be unlikely to affect clinical practice. In nonrandomized studies, particularly Phase I and II trials without comparator arms, HRQOL endpoints can provide important insights into how the patient regards treatment toxicities. In randomized treatment studies, HRQOL measures can capture the trade-off between symptom improvement and toxicity or the impact of delayed disease progression more effectively than any single biomedical endpoint.

• Moinpour and Provenzale noted the potential value of HRQOL data to inform colorectal cancer treatment decisions, especially when there are multiple effective strategies, only small differences in survival across strategies, or the emphasis is on psychosocial outcomes. Yet, most studies they reviewed found little change in HRQOL over time and only small HRQOL differences across treatment arms. They believe it is premature to draw firm conclusions because most studies were significantly flawed by inadequate sample sizes or nonignorable missing data.

• Mandelblatt and Selby pointed out that for evaluating the HRQOL impacts of cancer screening and chemoprevention interventions, generic measures such as the SF-36 may not be sensitive enough to detect certain important outcomes. They advocated further investigation of preference-based measures (such as the HUI) to capture the net impact on HRQOL of a panoply of subjective effects, including the anxiety, relief, reassurance, and/or discomfort.

• Zebrack and Cella found that a number of well-validated generic and general cancer instruments (including the SF-36, EORTC QLQ-C30, and FACT G) performed adequately in studies of cancer survivors. But they noted concerns about content validity since few existing instruments capture such survivor-specific issues as fear of disease recurrence, chronic physical compromise, and post-traumatic growth.

• Ferrell concluded that HRQOL assessment can play a vital role in end-of-life care, providing information beyond symptom measurement that can contribute to better decision making for the patient.
Snyder reported that a wide variety of HRQOL measures have been used to assess the subjective impacts of cancer caregiving on informal caregivers, with 2 particular instruments (the Caregiver Reaction Assessment and the Caregiver Quality of Life Index) performing well by the MOT attributes.

Finally, in a separate study that served to complement the COMWG’s work, NCI’s Community Clinical Oncology Program (CCOP) undertook a comprehensive review of all NCI-supported symptom-management trials initiated since 1987. In sum, just over half of these trials assessed “global quality of life” using a total of 22 distinct instruments (though the FACT G and the Uniscale were adopted in over half the trials). The conceptual framework for most applications consisted of a posited simple, one-way relationship between symptom relief and global quality of life; the possibly complex interplay between the different quality-of-life dimensions and types of symptoms was rarely recognized. Across these symptom-management trials, there was “no consistent relationship” found between global quality-of-life measures and either the symptoms being targeted or the interventions studied.

Regardless of the focus of the cancer trial—treatment, symptom management, screening, or prevention—several consistent themes emerge. To maximize the information value of PRO assessment in cancer trials, one needs a clear rationale for including PROs; an explicit conceptual model to guide the measurement task; appropriate instrumentation; and a sound plan for data collection, analysis, and reporting.

These are precisely the matters at issue when decision makers have to judge the merits of a claim that a particular drug provides “clinical benefit” based on improvement in PROs, as discussed below.

**FDA Guidance on PRO Measurement**

More than a decade in the making and much anticipated by the pharmaceutical industry, clinical trialists, and outcomes measurement researchers, the FDA’s draft “Guidance for Industry-Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims” was released for comment in February 2006. According to the FDA, the Guidance is intended to increase the efficiency of the FDA’s communication with industry about PRO endpoints in trials, streamline its review of the adequacy of PRO endpoints used to support product-labeling claims, and “provide optimal information about the patient’s perspective of treatment benefit at the time of product approval.” By way of definition, the FDA states the following:

A PRO measurement is any aspect of a patient’s health status that comes directly from the patient (ie, without the interpretation of the patient’s responses by a physician or anyone else). In clinical trials, a PRO instrument can be used to measure the impact of an intervention on one or more aspects of patients’ health status, ranging from the purely symptomatic (response to a headache), to more complex concepts (eg, ability to carry out activities of daily living), to extremely complex concepts such as quality of life, which is widely understood to be a multidimensional concept with physical, psychological, and social components.

Although the FDA’s final guidance on PROs has yet to be issued, it may be reasonable to expect it to include a number of themes enunciated in the draft document. With greater specificity than in the past, the FDA discusses the rationales for using PRO measures in medical product development: (1) some treatment effects (eg, pain intensity and pain relief) are known to define and communicate to industry sponsors of clinical trials the appropriate role of PRO measurement in decisions about drug and device approval and labeling. Then we briefly discuss 2 recent initiatives in which the FDA has participated to improve the application of PRO measures in clinical trials.
only to the patient; (2) patients provide a unique perspective on treatment effectiveness (since “improvements in clinical measures of a condition may not necessarily correspond to improvements in how the patient feels or functions”); and (3) formal assessment by patients may be more reliable than informal interviews with providers or other sources of information about the patient’s condition.

On balance, this proposed guidance suggests that PRO measures used in studies to support drug-labeling claims need to meet the same general standards of scientific rigor and clinical usefulness expected for more traditional clinician-reported outcome measures such as treatment toxicity scores. A forthcoming paper by FDA-associated authors identifies 5 “sources of bias” that, taken together, explain why HRQOL-based efficacy claims have not to date been accepted by the FDA for inclusion in anticancer product labels.”53 The sources of bias include lack of randomization (especially as occurring in single-arm trials); lack of blinding (since masking treatment assignments from patients or providers can prove difficult); missing data; multiplicity of endpoints (arising when one fails to adjust statistically for the resulting increased likelihood of obtaining a “significant” finding by chance); and intrinsic meaning (whether the HRQOL findings consider all relevant information, are internally consistent, and exhibit clinical relevance to the population studied) (see Note 2).

Along this line, it may be pertinent to note what FDA oncology officials reported in a 2003 article in the Journal of Clinical Oncology.3 Of 57 regular approvals for cancer drugs over 1990 to 2002, tumor response was the approval basis for 26, supported by relief from tumor-specific symptoms in 9 of the 26. Symptom relief “provided critical support” for approval in 13 of the 57 cases. And although many of the 53 marketing applications for approval based on nonsurvival endpoints “used surrogate endpoints for a better life, no approvals were based on instruments measuring health-related quality of life.”

But note that whatever the final form of the FDA’s PRO guidance, it applies expressly to the industry-submitted claims for product approval in the United States. Regulatory authorities elsewhere in the world have, in fact, taken a somewhat different perspective in recent years, particularly regarding the value of HRQOL measurement. A recent review of drug approvals in the European Union nations over 1995 to 2003 found that 34% of the dossiers submitted by product sponsors for evaluation reported HRQOL and other PRO measure findings, “with cancer-related treatments most frequently including PRO data.”55 Within the United States, the extent to which such an FDA guidance will influence the use of PROs in nonindustry trials is unclear. In particular, no more than 10% of NCI-supported cancer treatment trials are conducted to support FDA submissions (E. L. Trimble, personal communication, March 2006). Rather, most NCI Phase III treatment trials compare new interventions with standard therapeutic approaches to inform treatment decisions in the cancer community. A 2003 examination of all NCI-sponsored cancer treatment trials found that 31% (59 of 189) of Phase III trials and 4% (34 of 810) of Phase I, II, and I/II trials had 1 or more HRQOL endpoints.56

NCI’s Patient-Reported Outcomes Assessment in Cancer Trials (PROACT) Conference

Although NCI has never issued a formal guidance document on the use of PROs, NCI scientists have published extensively on the appropriate inclusion of quality-of-life measures in trial protocols (see Gotay et al57,58). Regarding decisions on the use of HRQOL in trials in recent years, NCI staff have essentially worked with investigators on a trial-by-trial basis. The guiding principle has been to encourage inclusion of HRQOL assessment in a trial when there is “an HRQOL hypothesis that will add to the existing body of knowledge, will generate hypotheses for future studies, or stimulate a change in clinical practice.”56

The NCI’s Clinical Trials Working Group, created to guide the NCI’s efforts to restructure its clinical trial enterprise, recommended in its 2005 final report that a “funding mechanism and prioritization process [be established] to ensure that the most important…quality of life studies” are carried out appropriately alongside clinical trials.59
With these developments as a backdrop, the NCI conducted an international conference in September 2006, focusing on “Patient-Reported Outcomes Assessment in Cancer Trials: Evaluating and Enhancing the Payoff to Decision Making.”6 Cosponsored by the American Cancer Society (ACS), the conference examined the circumstances under which the use of PROs, including HRQOL, in cancer trials could yield the greatest payoff to decision making; best practices for the application of PROs in a range of trials (Phase I/II, Phase III, and symptom management); and high-priority topics for future research. Conference findings, slated for publication in the Journal of Clinical Oncology in late 2007, are intended to inform the early deliberations of NCI’s new Symptom Management and Health-related Quality of Life Steering Committee (S. B. Clauser, personal communication, March 2007) (see Note 3).

Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT)

In parallel, there is at least one more initiative that could well have an impact on PRO measurement standards and practice, and it is called IMMPACT. Organized in 2002 as a voluntary association of participants from academia, government agencies (including the FDA and the NIH), patient advocacy organizations, and the pharmaceutical industry, IMMPACT’s mission has been to develop consensus recommendations for improving the design, execution, and interpretation of clinical trials of treatment for pain.13,60,61

PROS IN THE ASSESSMENT OF CANCER CARE IN THE COMMUNITY

In recent years, there have been a number of nonrandomized, “observational” studies—sometimes carried out at one or a few institutions and typically not population-based—that use PRO measures such as HRQOL to examine the burden of cancer on individuals or the effectiveness of interventions.62 In this section we highlight 3 large-scale, population-based research projects that demonstrate the application of PRO measures to investigate the outcomes and quality of cancer care from the patient’s perspective.

Prostate Cancer Outcomes Study (PCOS)

The first population-based evaluation of HRQOL for prostate cancer patients on a multiregional scale, PCOS was initiated by the NCI in 1994 in collaboration with 6 geographically dispersed cancer registries in the NCI’s Surveillance, Epidemiology, and End Results (SEER) program.63 On a final sample of over 3,500 men, PCOS investigators combined cancer registry and detailed medical records data to obtain information on specific diagnostic procedures, prostate-specific antigen (PSA) values, clinical stage and grade of tumor, details of treatment (prostatectomy, hormonal therapy, evidence of “watchful waiting”), and acute complications of treatment. Of particular import, study participants were surveyed by mail at 6, 12, 24, and 60 months after the initial diagnosis to elicit their own reports on such issues as urinary/bladder control, bowel habits, sexual function, satisfaction with care, and overall impact of their condition(s) on health status. Several complementary HRQOL measures were employed, including the SF-36 instrument, the E-VG-G-F-P summary rating scale, and some survey items developed expressly for PCOS.64 Through 2006, PCOS had generated about 25 publications, with findings such as these:

- Men treated for prostate cancer had, on average, the same general HRQOL 2 years after diagnosis, regardless of initial choice of therapy. Those reporting bother related to urination or impotence were more likely to have poorer HRQOL scores.65
- Five years post diagnosis for localized prostate cancer, men who chose radical prostatectomy continued to report worse urinary incontinence than those receiving external beam radiation therapy. However, overall sexual function did not differ between these 2 groups, largely because erectile function continued to decline for external beam patients.66
- For men who chose active treatment for clinically localized prostate cancer, the majority were satisfied with their treatment decision. However, following a decision to undergo either radical prostatectomy or androgen-deprivation therapy, Hispanic men
reported less satisfaction than non-Hispanic white men.67

**Cancer Care Outcomes Research and Surveillance Consortium (CanCORS)**

With PCOS showing the way, the NCI in 2001 launched CanCORS, a comprehensive study of cancer care treatment and outcomes in large, population-representative cohorts of individuals newly diagnosed with either lung or colorectal cancer.11 The consortium, composed of 6 research groups and a statistical coordinating center funded by the NCI and 1 group funded by the US Department of Veterans Affairs, has 2 principal aims: first, to determine how the characteristics and beliefs of cancer patients and their providers, as well as the characteristics of health care organizations, affect treatment decisions, clinical outcomes, and PROs; and second, to examine the association between specific treatments and patient outcomes.

To accomplish this, CanCORS investigators have identified roughly 5,000 lung cancer and 5,000 colorectal cancer patients, drawn from 5 geographically defined regions of the United States (which together include 5 integrated health care delivery systems), 15 VA medical centers, and a range of community-based practice sites. For each included patient, a rich longitudinal picture of cancer diagnosis, treatment, and outcomes is being constructed by linking registry data with information from medical records, administrative files, and surveys conducted of the patient and his or her cancer care providers (including, for a sample of patients, their informal caregivers).

Through telephone interview surveys, CanCORS investigators have collected PRO data using a variety of measures, including symptom/problem checklists, SF-12 items, lung cancer-specific items and colorectal cancer-specific items from the EORTC suite of HRQOL instruments, the EQ-5D preference-based measure, the EVG-G-F-P global health status scale, a visual analog scale (0 to 100), and questions on the patient’s perception of and satisfaction with cancer treatment planning and decision making.68

The scientific papers flowing from CanCORS over the next several years will likely provide unprecedented insights into the value, as well as possible limitations, of PRO measures for evaluating the effectiveness of cancer care in real-world community settings.

**National Initiative for Cancer Care Quality (NICCQ)**

Organized by the American Society of Clinical Oncology in 2000 with support from several professional societies, patient advocates, and private foundations, the NICCQ was an in-depth study of cancer care quality in a (final) sample of nearly 1,800 patients diagnosed with either breast or colorectal cancer across 5 geographically dispersed regions of the United States.12,69 Of particular interest for present purposes is that 9 of the 36 quality measures for breast cancer and 6 of the 25 quality measures for colorectal cancer focused expressly on “respect for patient preferences and inclusion in decision making”—measures necessarily requiring the patient’s own report. For these measures, overall adherence was 86% for breast cancer and 89% for colorectal cancer.69

**PROS IN CLINICAL ONCOLOGY PRACTICE**

There is growing recognition that routine measurement of PROs such as HRQOL in oncology practice has the potential to improve cancer care planning, monitoring, and management for patients and survivors.70–75 The rationales put forward include promoting better communication and shared decision making by patients and providers; assessing the health status of patients entering therapy and identifying treatable problems; determining the degree and sources of the patient’s decreased ability to function; distinguishing among types of problems, including physical, emotional, and social; detecting adverse effects of therapy; monitoring the effects of disease progression and response to therapy; informing decisions about changing treatment plans; and predicting the course of disease and the outcomes of care.

Regarding the final point, a comprehensive analysis by Gotay et al76 examined 66 published investigations from 1986 to 2005 of the relationship between the cancer patient’s reported HRQOL and length of survival. In 60 of these
66 studies, HRQOL was significantly related to survival time after controlling for biological variables in multivariable analyses. Indeed, HRQOL demonstrated a stronger relationship to survival prognosis than standard clinical predictor variables. A compelling recent example was provided by Efficace et al., who found that baseline social functioning scale scores of the EORTC QLQ-C30 were significant predictors of survival time for metastatic colorectal cancer patients.

Interestingly, much of the literature on implementing or evaluating PROs in oncology practice comes from Europe and Canada, not the United States where the use of PRO measures to inform patient-level decision making remains “rare.” Given the potential benefits, it is reasonable to ask why there has not been more active investigation, if not adoption, of PRO measurement in oncology care generally. Donaldson identifies patient, clinician, and health system factors that may be at work. For patients, potential hurdles include response burden, confidentiality concerns, and the possibility that PRO assessment may touch on sensitive topics. For clinicians, there may be doubts about whether PRO instruments are truly useful vehicles to inform cancer care decision making and about the time and practice resources required for PRO assessment (especially in the general absence of third-party payment for such activities). For health care delivery systems, potential barriers include the anticipated resource costs of collecting and managing PRO data, as well as concerns that such data could eventually be used by payers, purchasers, and others to monitor the quality of care.

To enhance the use of PRO measures in oncology practice, Donaldson urges the adoption of new information infrastructures and technologies, combined with redesign of cancer care delivery itself. Such changes could lower data collection costs, ensure confidentiality, and facilitate the day-to-day use of PRO information in patient–provider decision making.

System changes notwithstanding, provider reimbursement for PRO data collection and use will likely remain an important issue. In that regard, we note a striking, if time-limited, example of how third-party payers could go about reimbursing providers to collect HRQOL data in day-to-day oncology practice. It was provided by the “Demonstration of Improved Quality of Care for Patients Undergoing Chemotherapy” put in effect by the US Centers for Medicare & Medicaid Services (CMS) for (calendar year) 2005. Under terms of the Demonstration, medical oncologists delivering chemotherapy to Medicare enrollees could bill $130 per patient encounter if they collected and submitted data on the patient’s current status regarding nausea/vomiting, pain, and fatigue, as measured by items taken from the Rotterdam Symptom Checklist, a prominent HRQOL instrument. A preliminary analysis of the data by the CMS indicated that only a minority of patients suffered significant symptoms (defined as checking the Rotterdam boxes “quite a bit” or “very much”): 2% for nausea/vomiting, 8% for pain, and 26% for fatigue. To be sure, this 2005 CMS quality-of-life demonstration was not without its controversy (see Note 4). Nonetheless, it provides both a precedent and some practical information about how the “pay for performance” mechanism might be employed to incentivize PRO assessment in routine cancer care.

To accelerate this transition, it is important that patients, providers, and health systems become increasingly familiar with the PRO assessment process and ongoing efforts to make it feasible and attractive. In that regard, 2 recent international conferences sought to provide comprehensive assessments of the state of the science in PRO assessment in oncology practice. In June 2007, the International Society for Quality of Life Research (ISOQOL) sponsored a special conference on “Patient-Reported Outcomes in Clinical Practice.” The conference, held in Budapest, addressed a wide spectrum of issues: PRO use in clinical practice from the perspectives of various stakeholders (eg, patients and providers); theoretical underpinning for using PROs in clinical practice; applications of PROs in clinical practice; and topics pertaining to data collection, analysis, and interpretation. In October 2003, the Mayo Clinic sponsored the conference “Quality of Life III: Translating the Science of QOL into Clinical Practice.” The conference, held in Scottsdale, Arizona, investigated a range of issues bearing on the benefits and costs of incorporating HRQOL measures into clinical practice.
PROS IN POPULATION SURVEILLANCE OF CANCER PATIENTS AND SURVIVORS

At present in the United States, there is no data collection mechanism or process in place to monitor the national burden of cancer or to chart progress against that burden from the perspective of the cancer patient or survivor. By contrast, capitalizing on a strong cancer registry system, there has been the long-standing capability to track cancer incidence and mortality on an annual basis by cancer disease type and demographic subgroup. This is the case for all nations that keep cancer statistics.

The question thus arises: at the population level, how might we go about measuring and monitoring progress against the suffering due to cancer and, at the same time, progress toward improving the quality of life of all those touched by cancer?

These issues of PRO data availability take on added import with the ACS’s recent revision of its “2015 Nationwide Objectives,” which are designed to facilitate fulfillment of its 2015 Challenge Goal in the area of quality of life. This Challenge Goal calls for “Measurable improvement in the quality of life (physical, psychological, social, and spiritual) from the time of diagnosis and for the balance of life of all cancer survivors by the Year 2015.” In support of the Goal, ACS’s National Board of Directors has adopted 4 new Nationwide Objectives intended to foster, over time, both improvement in the quality of life of individuals affected by cancer and the capacity to measure changes on a population-wide basis. Specifically, ACS’s Quality of Life Objectives call for (1) improving access to care (by ensuring health care coverage for all by 2015); (2) reducing the impact of out-of-pocket costs on receipt of care by individuals diagnosed with cancer (so that by 2015, no more than 2% of cancer patients report cost-related access-to-care difficulties); (3) improving pain control by greatly strengthening pain management policies within the states by 2015; and (4) having in place national surveillance systems by 2015 to support population-based measurement of quality of life for individuals affected by cancer. That these objectives do not also include specific national targets for quality-of-life improvement simply reflects the fact, underscored in objective 4, that the data for such population-based monitoring are presently unavailable in the United States.

The question is how to accelerate the evolution to cancer data systems that incorporate not only traditional clinical and epidemiological parameters, but PROs. Building on pertinent commentary and recommendations from several recent analyses, we see at least 3 strategies, and they are not mutually exclusive:

- Exploit more fully the currently available data sources and mechanisms that collect PRO information on cancer patients or survivors. Prime examples include the National Health Interview Survey (NHIS), Medical Expenditure Panel Survey (MEPS), Medicare Health Outcomes Survey (HOS), and Behavioral Risk Factor Surveillance System (BRFSS) (see Note 5).
- Importantly, while these population-based surveys identify respondents who have been diagnosed with cancer, none was designed to provide statistically robust, population-based information on health status by disease category (eg, for cancer or types of cancer). In some cases (eg, MEPS), the total number of cancer patients sampled may be comparatively small. In other cases (eg, the Medicare HOS), the sampling frame is not representative of all cancer patients or survivors. Consequently, a second broad strategy is to enhance existing survey mechanisms by increasing the number of cancer patients interviewed, enlarging the sample frame to encompass all relevant segments of the population, or both. Even then, a problem common today may persist: there is rarely a way to link the self-identified cancer respondent being surveyed to other information about their cancer diagnosis and treatment. Consequently, we would still have limited capability to analyze the clinical or other determinants of the PRO assessments provided by these respondents.
- Thus, a third strategy is to accelerate development of a “national cancer data system” that not only links high-quality registry data with medical records, insurance claims data, and other administrative information, but also samples cancer patients and survivors to collect population-based PRO data. Such a data
system would support ongoing statistical analyses of patient, provider, and health system factors accounting for variation in PRO outcomes.

The technology for carrying out such an ambitious effort is fully available now, even as the nation progresses apace toward an interoperable electronic health care information system. Research initiatives like PCOS, CanCORS, and the NICCQ demonstrate that we already know “how to do it.” The question is whether public agencies and private cancer organizations will devote the resources and then work in concert to create an enduring cancer data infrastructure.

**ENHANCING THE SCIENCE OF PRO MEASUREMENT AND APPLICATION**

Our understanding of how to define, measure, and use PROs in cancer has grown substantially in recent times, with progress accelerating over the past decade or so. No doubt, approaches to PRO measurement and application to decision making will continue to evolve and improve. We conclude with brief commentary on opportunities and challenges in 3 particular areas.

**Improving the State of the Science in PRO Measurement**

While PRO measures can undoubtedly benefit from “continuous quality improvement” in multiple ways, there are several specific issues that merit and are receiving close attention.

**Strengthening the Conceptual Foundations of PRO Measurement**

As underscored in a number of COMWG analyses and in NCI’s review of HRQOL in symptom–management studies, much work remains in strengthening the conceptual underpinnings of PRO measurement models. Ferrans has proposed additional work in a number of areas, including stronger attention to the cause–effect relationships among variables; distinguishing more clearly between “objectively” measured biomedical outcomes and “subjectively” measured PROs; and identifying not only the PRO domains that are conceptually important to measurement, but their interrelationships. Hays et al. has advocated “structural equation modeling” as an analytic platform for exploring conceptual model–measurement model relationships, as well as rigorous investigation of other MOT attribute issues, such as construct validity.

**Enhancing the Interpretability of PRO Data**

Two matters have drawn much attention in recent years: defining what constitutes a MID in a PRO measure score and understanding whether the patient’s adaptation to illness over time influences not only his or her PRO measure scores, but the very meaning of those scores to the patient (“response shift”).

Perhaps the most discussed if not also most challenging interpretative issue concerns defining the MID. The issue arose at multiple points in the COMWG’s deliberations. Working on this issue in parallel with but independently of the COMWG was a Clinical Significance Consensus Meeting Group (ClinSig) organized by Mayo Clinic statistician Sloan and comprising 30 experts from academia, industry, and government. ClinSig produced a 6-paper monograph examining such topics as individual versus group differences in the meaning of clinical significance; assessing HRQOL changes over time; and strategies for communicating HRQOL findings to patients, providers, and other decision makers. The monograph highlighted the strengths and limitations of the 2 principal approaches for demonstrating clinical significance in a PRO measure. The “distribution-based approach” expresses the PRO-measured treatment effect in terms of some underlying statistical distribution of possible effect results; for example, it has been argued that an effect may be regarded as clinically significant if it is greater in absolute value than one half the standard deviation of effect sizes. The “anchor-based” approach compares the PRO-measured treatment effect to concurrent changes in some independent standard (or anchor) that itself is regarded as clearly interpretable and meaningful (eg, ability to perform usual activities).

Writing on the topic of clinical significance for the COMWG, Osoba (who was also a member of ClinSig) concluded there is preliminary evidence suggesting that a meaningful change in an HRQOL score appears to be about 7% of the full breadth of the measurement scale,
perhaps bracketed by 5% and 10%. For example, this implies that a score change from, say, 50 to 57 on a 100-point scale constitutes a perceptible and meaningful HRQOL improvement. Such a difference (7%, with a range of 5% to 10%) is consistent with what Cohen and others have termed a “small” to “medium” effect size, which is often regarded as approximating a MID in effect sizes. Hence, there is preliminary evidence that the MID for an HRQOL measure is roughly the same under current articulations of distribution-based and anchor-based approaches to defining the MID.

On the other hand, there is no direct evidence from either its draft guidance to industry or recent publications by its staff scientists that the FDA openly embraces or rejects any of these approaches to determining what constitutes a clinically meaningful difference in an HRQOL measure (or a PRO measure, generally).

The second intriguing interpretation matter is response shift, the hypothesized phenomenon that the very meaning of an individual’s evaluation of a subjective construct like HRQOL changes over time (see Schwartz and Sprangers). Such response shift within an individual is said to reflect changes in either his or her (1) internal standards of measurement; (2) valuation of the individual HRQOL domains; or (3) definitions or perceptions of what the domains mean. For example, the following stylized scenario is consistent with the occurrence of a response shift. Suppose that a (cancer-free) individual rates his own health as Very Good (on the E-VG-G-F-P global scale) and the health of someone described with localized prostate cancer as Fair. Suppose this same individual is subsequently diagnosed with localized prostate cancer and now—living his life with this cancer—rates his overall health as Good. The pre/post change in his rating of localized prostate cancer from Fair to Good could reflect internal adaptations arising from either 1, 2, or 3 above.

Ferrans noted that a better understanding of response shift could lead to conceptual models of HRQOL that provide a better portrayal of dynamic relationships among “objective” biomedical outcomes and (the comparatively more malleable measures of) HRQOL.

Applying Modern Psychometric Approaches to Enhance PRO Assessment

As noted earlier, modern measurement approaches such as IRT modeling may offer significant opportunities to enhance the rigor and efficiency of PRO data collection and analysis compared with what is possible using conventional psychometric approaches (based on “Classical Test Theory” [CTT]).

IRT modeling not only allows PRO item responses to inform the scale score assigned to an individual as with CTT, but also uses (large samples of) individual item responses to estimate a PRO “scale score” for each item. Such information allows the investigator to pose items to a respondent strategically so as to rapidly “hone in” on his or her most likely position and, thus, statistically optimal score on the PRO measurement scale.

In consequence, it can be demonstrated that IRT modeling can generate PRO measurement scales with stronger reliability across the full breadth of the scale and also possibly greater validity in terms of being able to distinguish among true individual-level differences in outcome. IRT (unlike CTT) enables one to “cross-walk”—and, thus, compare directly—the scores between instruments purporting to measure the same health-related construct. IRT also facilitates a statistically rigorous investigation of what has been called “differential item functioning,” which is whether a given PRO instrument performs the same or differently when applied to populations that differ by cultural background, geography, or other considerations.

Perhaps most significantly, IRT modeling provides the theoretical underpinnings for the construction of “item banks” to facilitate PRO measurement via computer-adaptive testing (CAT). Under this approach, IRT modeling is used to choose the PRO survey items, taken from existing fixed-item instruments or else newly created, that constitute the membership of the bank. To assess where a given individual lies on a given PRO scale (for example, to assign the individual a score on a physical functioning scale), the CAT procedure selects survey items for the individual in a strategic fashion, with the answer to the previous questions driving the
selection of the next question. The upshot is that the individual's scale score can be estimated with acceptable precision using significantly fewer questions than with traditional fixed-item instruments. (Precisely the same methodology is used today in computer-assisted administration of the Scholastic Aptitude Tests and other standardized exams.)

In reality, virtually all PRO measure applications in the published literature to date are based on fixed-item instruments. So the promise of IRT modeling, demonstrated in a number of recent experimental studies, awaits the broad “market test.” And that test is about to come. In 2004, as part of its new Roadmap Initiative, the NIH launched a 5-year, $25 million project to develop the Patient Reported Outcomes Measurement Information System (PROMIS). PROMIS is building web-based, public-domain item banks to support CAT for selected health symptoms and HRQOL domains affected by a variety of chronic diseases, including cancer. A network of 6 research and data collection sites and a statistical coordinating center are pursuing the following aims: electronic administration of individually tailored PRO questionnaires via a number of secure platforms (eg, computers, Internet, telephone); collection of PRO data in research studies, including clinical trials; and, eventually, the capability of providing instant-turnaround PRO reports to patients, providers, and researchers.

Targeted Therapies and Longer Life Expectancies: PRO Measurement in the 21st Century

While today’s cancer therapies are frequently effective at arresting the progression of disease, they often have significant toxic effects since they generally affect healthy tissue and organs in addition to cancerous cells; surgery, of course, carries its own particular risks. Targeted therapies are at the center of a paradigm shift in cancer treatment (see Gotay). They work directly to interfere with the cancer cell growth process, thus greatly reducing damage to adjacent healthy cells. On the other hand, some targeted therapies may require ongoing administration so that patients are to maintain contact with their oncologist over time—indeed, they may never be able to “complete” treatment in a decisive way. Rather, the individual may live for years with his or her cancer under control, but at risk to other chronic disease problems.

Consequently, do targeted therapies require a distinctly new approach to PRO assessment? If so, in what ways? For example, to focus solely on cancer symptoms or single HRQOL domains—in the spirit of published observations by FDA senior oncology staff—could prove problematic for new types of cancer therapy where the nature of the health status impacts is simply not known early on. Rather, it may be necessary to ascertain the patient’s own perspective on the short-term effects of these new treatments through qualitative interviews.

The possible long-term and late effects of cancer treatments are receiving heightened attention now, initially for pediatric cancers and more recently for adult cancers. Adverse downstream consequences of therapy may include excess fatigue, sexual dysfunction, other physical effects (such as arm edema), and cognitive problems. Consequently, it is important that PRO assessment extend across the survivorship period, perhaps through greater emphasis on postmarketing surveillance of therapies.

The challenges outlined here point to the ongoing importance of monitoring and enhancing the content validity of PRO instruments (see Table 2). For example, it is reasonable to ask whether certain HRQOL instruments developed and validated in younger cancer patients are the most appropriate measures for older cancer patients. Likewise, as Zebrack and Cella have pointed out, HRQOL instruments frequently used among older adult cancer survivors may not adequately address the concerns of young adult and adolescent survivors.

Meeting the Needs of Decision Makers

Matching the PRO Measure to the Task at Hand

For each potential PRO application, an enduring challenge is selecting the most appropriate PRO measure(s) from the large menu of available options. From the COMWG analyses, we conclude the following:
When the aim is to compare interventions in clinical trials or observational studies, very specific and highly targeted PRO measures may be most useful, especially for detecting “small,” though clinically important, differences.

When the aim is to use PRO data at the bedside, the instrument(s) need to provide clear and interpretable data to the clinician and the patient.

Research on patient-provider decision making may be well addressed through preference-based HRQOL measures, often in conjunction with nonpreference-based measures.

Cost-effectiveness analyses of cancer interventions that attempt to account for the impact on both quality of life, as well as life expectancy (through use of the Quality-Adjusted Life Year metric), require preference-based measures of HRQOL.

When the aim is to compare the HRQOL of a cancer population to that of the population at large (as might arise in studies of cancer prevention or cancer survivorship), it is appropriate to use a generic HRQOL measure for which population norms have been computed.

When the aim is to use PRO data for national, state, or even local policy guidance (for example, to monitor the cancer burden in relation to other diseases or to evaluate interventions that address the cancer burden while recognizing competing disease risks), it is appropriate to use generic measures that facilitate cross-disease comparisons.

### Toward a Research Agenda on PROs in Cancer Decision Making

As a review of the thousands of citations in the COMWG’s 32-chapter *Outcomes Assessment in Cancer* will show, there is a large, rapidly growing, and increasingly sophisticated literature on the psychometric development and research application of PROs across the cancer continuum. A closer look at these citations will also suggest there is only a small, still nascent effort by investigators to understand and improve the way that patients, providers, payers, regulators, and other policy makers use these measures in decision making. (A noteworthy exception is the FDA’s current active engagement with PRO applications.)

Consequently, the time is ripe for a well-designed program of research to identify where PRO measures have made a difference, where they have not, the reasons for success or failure, and how to enhance their usefulness. The emphasis would be on case studies and other in-depth qualitative analyses to gain a deep understanding of the structure, processes, and outcomes of decision making in specific cancer contexts, with an eye to the nature of the information that proved decisive for the choices being made. The cancer contexts of interest are precisely those examined in this paper: clinical trials and observational studies examining the outcomes of cancer interventions; efforts to collect and use PRO data to inform patient-provider decision making in oncology practice; and population surveillance of progress against the cancer burden from the perspective of the cancer patient and survivor.

Presently, our knowledge of the role that PRO measures play or could play is still based more on anecdote than analysis.

We need a new synergism wherein the PRO measurement literature informs the decision maker, who in turn influences the direction of the literature. It is no stretch to call it a PRO-active approach.

### NOTES

1. The organizations were the ISOQOL, the International Society for Pharmacoeconomics and Outcomes Research, the Health Outcomes Committee of the Pharmaceutical Research and Manufacturers of America, and the European Regulatory Issues on Quality of Life Assessment.

2. At a February 2006 conference jointly sponsored by the Mayo Clinic and the FDA to invite public discussion and recommendations on the draft guidance, several HRQOL measurement scientists (among the roughly 300 attendees) frankly asked whether the guidance might serve to provide an “unlevel” playing field for PRO assessment by holding PROs to a higher explicit standard than clinical outcome measures. FDA officials at the
conference responded that the intent over the long term is to hold all trial endpoints to appropriately high standards. Most recently, a number of prominent PRO researchers have formed a new organization called “Assessing the Symptoms of Cancer using Patient-Reported Outcomes” (ASCPRO) in response to the publication of the FDA PRO draft guidance. ASCPRO will focus on methodological issues in assessing cancer-related symptoms (see http://www.ascpro.org).

3. The breadth of organizational leadership and participation in the PROACT conference is perhaps an indication of the current depth of interest in the application of PROs. The conference was cosupported by 3 NCI extramural divisions (Cancer Control and Population Sciences, Cancer Treatment and Diagnosis, and Cancer Prevention), cosponsored by the ACS, and guided by a scientific program committee with representatives from the NCI Cooperative Groups and CCOP research bases, the FDA, the ACS, the pharmaceutical industry, and the academic community of cancer outcomes researchers.

4. Launched roughly in parallel with CMS’s decision to significantly reduce Medicare reimbursement payment rates to oncologists providing office-based chemotherapy, this demonstration was projected to provide an extra $300 million in total payments to medical oncologists in 2005. A Department of Health and Human Services Inspector General report requested by the then-chair of the Senate Finance Committee (Charles Grassley, R-IA) reviewed activity over the first 6 months of the demonstration and noted several concerns.79 Beneficiaries were required to make the standard Medicare Part B 20% copay (here, $26) for the quality-of-life assessment. There was evidence of variation in how the data were collected and submitted; there was no direct incentive to focus on interventions to relieve symptoms; and there was concern that oncologists were being paid extra to collect information that should already have been a part of routine care assessment.79 In 2006, CMS significantly altered the focus of the oncology demonstration, providing supplementary payments to medical oncologists based on self-reported adherence to certain patient evaluation and management guidelines.80

5. The NHIS conducted by the US National Center for Health Statistics annually surveys 40,000 households covering 100,000 individuals, with questions posed on functional status. The MEPS, supported by the US Agency for Health Care Research and Quality, surveys about 25% of the NHIS-sampled households in greater depth, collecting such PRO data on health status via the SF-12 instrument (a compressed version of the SF-36) and also instrumental activities of daily living. The HOS, supported by the US Centers for Medicare and Medicaid Services, annually surveys roughly 200,000 Medicare beneficiaries enrolled in Medicare HMO plans and includes the SF-36. The BRFSS has a state-based sampling frame that targets approximately 150,000 individuals annually across the nation and collects self-report data on health status, HRQOL, and a range of behavioral and demographic variables.85

REFERENCES


7. Doroshow JH. Restructuring the National Cancer Clinical Trials Enterprise: Clinical Trials Working Group Implementation Update. Presented at: Patient-Reported Outcomes Assessment in Cancer Trials: Evaluating and Enhancing the Payoff to Decision Making; September 20–21, 2006; Bethesda, MD.


Erratum

In the May/June 2007 issue, in the article “Minimal Access Cancer Management” (CA Cancer J Clin 2007;57:130–146), an error appeared in the text on page 138. The sentence that read, “Laparoscopic resection of the colon was first described in 199272” was incorrect. In fact, the first report was in 1991 (Jacobs M, Verdeja J-C, Goldstein HS. Minimally Invasive Colon Resection [Laparoscopic Colectomy]. Surg Laparosc Endosc 1991;1:144–150). The authors regret the error and thank Dr. Juan-Carlos Verdeja for bringing it to their attention.