Quality of Life Research (2007) 16: 143–164
DOI 10.1007/s11136-006-9116-x

Cancer outcomes measurement

Through the lens of the Medical Outcomes Trust framework

Joseph Lipscomb[1], Claire F. Snyder[2] & Carolyn C. Gotay[3]
[1]*Department of Health Policy and Management, Rollins School of Public Health, Emory University, Rm 642, 1518 Clifton Road, NE, Atlanta, GA, 30322, USA (E-mail: jlipsco@sph.emory.edu);* [2]*Division of General Internal Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA;* [3]*Cancer Research Center of Hawai'i, University of Hawai'i, Honolulu, HI, USA*

Accepted in revised form 16 August 2006

Abstract

Background: In 2001, the U.S. National Cancer Institute established the Cancer Outcomes Measurement Working Group (COMWG) to evaluate and advance the state of the science in patient-reported outcome (PRO) measurement, with a focus on health-related quality of life (HRQOL). To guide its work, the COMWG adopted the revised Medical Outcomes Trust (MOT) attributes and review criteria for evaluating health status and quality-of-life instruments. *Objective*: With the MOT attributes providing the organizing principle, this paper summarizes and draws inferences from key COMWG findings about the methodological soundness of HRQOL assessment in cancer and steps required to move the field forward. *Results and Conclusions*: Across a range of cancer research applications, especially clinical trials, a variety of generic, general cancer, and cancer site-specific measures of HRQOL have demonstrated adequate reliability, validity, responsiveness, feasibility, and cultural and language adaptation. Methodological challenges remain in the interpretability of HRQOL measures, though substantial progress has been made in defining a ''minimum important difference'' in scale scores. Much work remains in forging a stronger link between the conceptual model and measurement model in HRQOL instrumentation. Progress along all MOT attributes will likely accelerate with the growing application of modern psychometrics, particularly item response theory modeling, which provides the underpinnings for item banking and computer-adaptive assessment of HRQOL. Future research should emphasize prospectively designed studies to evaluate PRO measures within the MOT framework and in-depth investigations of the role of PRO measures in cancer decision making at all levels.

Key words: Health-related quality of life, Oncology, Outcomes research, Patient-reported outcomes

Introduction

In 2001, the U.S. National Cancer Institute (NCI) created the Cancer Outcomes Measurement Working Group (COMWG) to evaluate the state of the science in cancer outcomes measurement and recommend approaches for moving the field forward [1]. To guide its assessments, the COMWG adopted the revised Medical Outcomes Trust (MOT) attributes and criteria for evaluating the psychometric performance of health status and quality-of-life instruments [2]. The resulting analyses, taken together, represent the most comprehensive effort to date to apply the MOT

framework in a consistent fashion to evaluate outcomes measurement within an entire disease area [3].

With the MOT attributes providing the organizing principle, this paper draws broadly from COMWG findings and recommendations to identify current strengths, limitations, and opportunities for improvement in cancer outcomes measurement.

Background and methods

COMWG structure and development

For cancer outcomes research to achieve its potential to inform decision making, cancer outcome measures must be scientifically sound and regarded as meaningful and useful by patients, survivors, family members, providers, payers, and regulators, among others [4]. In response, the NCI established the COMWG, comprising 35 experts drawn from academia, government, industry, and the cancer patient and survivorship communities (see Appendix). The majority of COMWG members were cancer researchers, selected from medicine (with 9 of 12 clinicians being oncologists), nursing, psychology, and social work. Members included experts in economics, biostatistics, psychometrics, and health services research generally. The perspectives of the cancer patient and survivor were given particular emphasis through two members recommended by NCI's Director's Consumer Liaison Group.

The COMWG was not a federal advisory committee or decisional body, but rather an NCI-constituted working group whose individual members addressed specific research questions posed by the NCI. (The authors were members of the COMWG, with JL and CCG serving as co-chairs and CFS, as the NCI-designated "initiator" of the working group.)

COMWG focus

Although traditional biomedical endpoints, especially survival and disease-free survival, remain of central importance in cancer decision making, outcome measures that reflect the perspective of the individual touched by cancer are of increasing interest to researchers and policy makers. Such patient-reported outcome (PRO) measures include health-related quality of life (HRQOL), perceptions of and satisfaction with care, and the economic impact of cancer. Because these measures not only hold great promise but also pose significant methodological challenges, they were the COMWG's primary focus.

The COMWG accorded particular attention to HRQOL, that PRO category generating the largest literature, both theoretical and empirical, and the most debate about measures, methods, and applications. HRQOL was defined broadly, to encompass *patient-reported* symptoms, functional status, and global well-being; and to be either single-dimensional or multi-dimensional, depending on the nature of the application. Instruments for measuring HRQOL in cancer can be *generic* (not specific to cancer but broadly applicable to ill and well individuals), *general cancer* (for use with cancer patients regardless of disease site), or *cancer site-specific*.

COMWG members assayed HRQOL measurement across the cancer continuum, defined to include prevention, screening, treatment, survivorship, and end of life. The reviews and analyses focused on the four most prevalent cancer disease sites in the U.S. – breast, colorectal, lung, and prostate. In 2003, about 55% of all new cancer cases and just over 50% of cancer deaths were attributable to these four sites [3].

COMWG operations

Over roughly a 3-year period, individual COMWG members reviewed and evaluated specific aspects of cancer outcomes measurement, met as a group three times to discuss findings, exchanged data and ideas with each other electronically, and prepared written reports on specifically assigned topics for submission to the NCI. The primary source of data for the majority of these reports was the peer-review literature through 2002. Several reports also included *de novo* psychometric analyses of secondary data sets to illustrate important methodological points. In addition, 13 focus groups (involving in total over 100 cancer patients,

survivors, and outcomes researchers) were conducted in 2001 to provide participating COMWG members with information and perspectives not available from the literature.

The COMWG member reports now comprise, along with four supplementary papers, the 32 chapters of the edited volume, *Outcomes Assessment in Cancer: Measures, Methods, and Applications* [3]. For an overview of COMWG findings and implications directed especially to cancer researchers and policy makers, see [5].

MOT framework

In 2002, the Scientific Advisory Committee (SAC) of the non-profit Medical Outcomes Trust published its revised set of attributes and criteria for evaluating the psychometric performance of health status and quality-of-life instruments [2]. The SAC updated the original MOT framework, released in the mid-1990s, "to take account of expanding theories and technologies" for instrument development [2, p. 193]. Within the current MOT framework, instruments are evaluated along eight defined attributes (see Table 1): conceptual and measurement model, reliability, validity, responsiveness, interpretability, respondent and administrative burden, modes of administration, and cultural and language adaptations. *All COMWG reports that assessed outcome measurement instruments adopted this framework*, strengthening the coherence of the working group's final product.

As the revised MOT framework recognizes and as some of the COMWG findings discussed below underscore, modern measurement approaches such as item response theory (IRT) modeling may offer significant opportunities to enhance the scientific rigor (and also cost efficiency) of PRO assessment. Specifically, IRT modeling not only allows survey item responses to inform the scale score assigned to a person (as with Classical Test Theory [CTT] approaches to measurement), but also allows person responses to inform item parameter estimation. Thus, IRT modeling brings fundamentally more information to bear in psychometric analyses than CTT. Whether and how this additional information leads to improved outcomes assessment was a recurring topic in COMWG deliberations.

Organization of the paper

The next seven sections discuss and draw implications from key findings of the COMWG, as filtered through the lens of the MOT framework. Each section corresponds to a specific MOT attribute, with the exception that respondent and administrative burden and modes of administration are consolidated. Tables 2 and 3 summarize key COMWG findings and conclusions relevant to these seven sections. The final section emphasizes that close adherence to the MOT framework may be viewed as a means to an important end: decision relevance. To that end, the paper concludes by identifying research directions to enhance the scientific strength and usefulness of cancer outcome measures.

Conceptual and measurement model

The generic [6], general cancer [6], and cancer site-specific [7–10] HRQOL instruments reviewed by COMWG members generally fared well along six of the seven MOT-specified review criteria under this attribute. As indicated in Table 2, these six criteria entail specifying the concept(s) to be measured, describing the target population and its role in developing instrument content, reporting information on the dimensionality of the instrument's scale(s), providing evidence of adequate variability across each scale, indicating the intended level of measurement (e.g., ordinal vs. cardinal), and describing the procedures and rationale for deriving scale scores from raw data [2, pp. 196, 198]. In addition, these six requirements appear to be met by many instruments commonly used to assess HRQOL in cancer survivors [11], cancer patients at end of life [12], and cancer caregivers [13].

The remaining MOT review criterion here – "conceptual and empirical bases for item content and combinations" [2, p. 196] – asks whether the specification of the HRQOL measurement model (that is, the instrument) was guided by a well-articulated conceptual model. On this point, the COMWG chapters just cited offer little affirmative evidence. Ferrans, who examined conceptual model – measurement model issues in some detail, acknowledged the availability of many well-established HRQOL instruments in cancer,

Table 1. Attributes of health status and quality-of-life instruments as identified by the MOT[a]

Conceptual and measurement model	The conceptual model provides the rationale for and articulation of specific concepts, or instrument domains, of importance and also their interrelationships in measuring the outcome of interest (e.g., health-related quality of life) in a population of interest (e.g., breast cancer patients). The measurement model (ideally) is the operational counterpart to the conceptual model, with the specified domains taking concrete form as constructs to be measured *via* the items included in the instrument.
Reliability	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* or *inter-rater* reliability.
Validity	The degree to which an instrument measures what it claims to measure. In the MOT framework, three distinct concepts are identified. *Content validity*: the degree to which available evidence supports the claim that each domain of an instrument, as defined through its item content, is appropriate for its intended use.*Criterion validity*: the degree to which the scores from an instrument relate to some designated gold-standard measure of the concept.*Construct validity*: the degree to which evidence supports a proposed association of scores based on theoretical implications associated with the constructs being measured. In principle, construct validity requires specification of a conceptual model and corresponding measurement model, and then an analysis of whether the instrument successfully measures the implied constructs. In practice, construct validity is typically demonstrated by statistical confirmation of the hypothesized relationships among instrument scores and other selected variables logically relating to these scores (e.g., a positive association between physical functioning scores and labor market participation among non-elderly patients). Also, for multidimensional instruments, there may be analysis of whether the estimated scales are measuring the distinct constructs hypothesized.
Responsiveness	Connotes the ability of an instrument to detect outcome changes over time. The allied concept of *sensitivity* refers to the ability to detect point-in-time differences in a cross-section of respondents. Responsiveness often viewed as important aspect of longitudinal construct validation.
Interpretability	The degree to which readily understood meaning can be attached to the quantitative scores from an instrument.
Burden and alternative modes of administration[b]	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, interviewer administered, trained observer rating, performance-based measures, and computer-assisted approaches (including computer-adaptive testing (CAT) using item banks).
Cultural and language adaptations	The degree to which an instrument that is being translated into another language or cultural setting is conceptually and linguistically equivalent to the original; such equivalence is generally evaluated by comparing the instruments' various measurement properties.

[a]Adapted directly from [2].
[b]The distinct MOT attributes of burden and alternative modes are discussed together here, because they are frequently inter-related.

but concluded that the field "has proceeded generally in an atheoretical manner..." [14, p. 27]. Instrument developers and users have generally focused more on identifying the individual domains of HRQOL than on investigating the potential inter-relationships among domains. In addition, researchers have not paid sufficient attention to whether HRQOL may be influenced by mediating factors both internal and external to the individual (e.g., personality traits and education level, respectively).

To strengthen the conceptual foundations of HRQOL assessment, Ferrans recommended additional work on several fronts:

– clarifying causal relationships in the conceptual model, including the appropriate distinction between "causal" and "indicator" variables

Table 2. HRQOL instrument[a] performance in cancer treatment

	Breast	Colorectal	Lung	Prostate
Conceptual and measurement model	*Overall findings across diseases and the cancer continuum*: Most of the generic, general cancer, and cancer-site specific instruments selected for intensive review by COMWG members appeared to meet six of the seven MOT-specified review criteria for assessing instrument performance on this attribute [6–13]. These six criteria are, in summary: concept to be measured, target population involvement in content derivation, information on dimensionality and distinctiveness of scales, evidence of scale variability, intended level of measurement, and rationale for deriving scale scores [2, p. 196]. However, COMWG members found little evidence that instruments met the one remaining review criterion, dealing with the conceptual and empirical basis for item content – that is, the conceptual model – measurement model relationship. In line with this, Ferrans [14] concluded that HRQOL instrument development has tended to be atheoretic, with a strong emphasis on appropriate selection of individual domains but less attention to potential domain inter-relationships and the possible mediating influence of internal and external factors on assessed HRQOL.			
Reliability	Cronbach's α for internal consistency reliability was generally within acceptable ranges for studies reviewed. Coefficient α was 0.92 and 0.85 for the FACT G and FACT B, respectively; ranged from 0.60 to 0.86 for the FACT G subscales; and was 0.88 for the FACT-based TOI-PFB scale [7]. For the EORTC BR23, α ranged from 0.70 to 0.91 in a U.S. sample and from 0.57 to 0.89 in a Dutch sample [7].For a generic HRQOL measure (HADS), α ranged from 0.72 to 0.94; generally similar results were reported for two general cancer measures, the RSCL and the FLIC [6].Strong test–retest reliability was reported for FACT B ($r = 0.85$) and FACT G ($r = 0.92$).	FACT G and FACT C showed good internal consistency reliability in samples of advanced CRC patients ($\alpha = 0.88$ and 0.91, respectively); an English language sample of CRC patients with a mix of disease stages (0.84 and 0.87), and a Spanish sample of CRC patients with mixed stages (0.89 and 0.88) [10]. The subscales of the EORTC QLQ-CR38 had α scores ranging from 0.38 to 0.88, with most greater than 0.70 [10].This same investigation found subscale-specific test–retest scores ranging from 0.53 to 0.92, with most falling in the 0.75–0.85 range.	For subscales of EORTC QLQ-LC13, α scores ranged from 0.53 to 0.83, with strong performance of multi-item dyspnea subscale, but less so for pain subscale ($\alpha = 0.53$–0.54). For FACT L, α reported at 0.68, and for FACT-based TOI index it was 0.89. For LCSS, $\alpha = 0.82$; this lung-cancer-specific instrument also demonstrated strong test–retest reliability ($r > 0.75$ for all items) and inter-rater agreement ranging from 95% to 100% at 8 centers [9]. In a multi-country study of 305 lung cancer patients, EORTC QLQ C30 scales had α scores ranging from 0.52–0.54 (role) to 0.80–0.85 (fatigue). Reliability tended to improve with repeated measurement. For the POMS, a generic HRQOL measure, $\alpha = 0.92$ in a study of small-cell lung cancer [6].	For the nine prostate cancer-specific HRQOL instruments assessed, internal consistency reliability for scales and subscales ranged from about 0.60 to 0.93, with most scores well above 0.70 [8]. For example, $\alpha = 0.87$ for the FACT G, and ranged from 0.64 to 0.84 for its subscales, with the prostate cancer subscale of the FACT P having $\alpha = 0.69$. Another study found α ranging from 0.75 to 0.83 when the FACT G was used with metastatic prostate cancer patients [6]. For the UCLA PCI, $\alpha > 0.90$ for urinary and sexual function, but 0.65 for bowel function [8]. For three of the nine instruments, test–retest reliability was reported; correlations ranged from about 0.64 to 0.93, with most greater than 0.80 [8].

Table 2. Continued

	Breast	Colorectal	Lung	Prostate
Validity	Content validity pursued through interviews with patients and oncology professionals (FACT G, FACT), sometimes supplemented by literature reviews (EORTC QLQ-BR23) [7].Construct validity of EORTC QLQ-BR23 demonstrated through ability to detect HRQOL differences among metastatic and other patients (known-groups method). For FACT B and FACT G, concurrent construct validity supported through correlations with FLIC and POMS; divergent validity indicated through low correlation with Marlowe-Crown Social Desirability Scale [7].	Modular construction of EORTC QLQ-CR38 and FACT C reflects careful efforts by developers to capture colorectal cancer impact. But these and other instruments may not be sufficiently sensitive to detect some CRC effects, such as post-operative disruption of diet, social, and sexual function.Evidence supporting construct validity of EORTC QLQ-CR38 included multitrait scaling results showing strong item-scale correlations; the ability to distinguish HRQOL between two known groups (metastatic vs. earlier-stage CRC); and descriptive statistics showing that for most subscales, scores tended to be distributed symmetrically across the full range [10]. Convergent and divergent construct validity for FACT C supported by analysis showing interscale and overall scale correlations with POMS in the expected directions [10]. FACT C also discriminated between patients based on performance status and extent of disease (known-groups).	Content validity of FACT L claimed on basis of substantial patient and provider input; for LCSS, developers reported 96% agreement on item content among patient, physician, and nurse groups. Item content validation for EORTC QLQ-L13 conducted by multi-country project group of professionals [9]. Construct validity of EORTC QLQ-L13 and LCSS supported by ability to discriminate among known groups defined by performance status or disease stage; for FACT L, evidence for construct validity includes strong positive correlations (about 0.60) with FLIC and with FACT G. LCSS also claims criterion validity on basis of positive correlations (ranging from 0.47 to 0.67) with a variety of HRQOL instruments [9]. For the general cancer instrument EORTC QLQ C30, construct validity was well supported by known-groups analysis among lung cancer patients, though less well supported in terms of being able to discriminate among patients by disease stage [6].	Content derivation and validation for the nine prostate cancer-specific instruments varied considerably. EPIC developers were informed by patient focus groups, a range of oncology professionals, survey researchers, and also a literature review. For other instruments, item content was derived from, and tested on, some but not all of these sources [8].For the nine instruments, construct validation was examined generally through multi-trait analyses that sought to validate the instrument's posited scale structure or through correlational analysis to demonstrate the relationship between a scale and some external (validating) indicator of HRQOL [8]. For example, multi-trait results were consistent with the EPIC's posited three domains (urinary, sexual, bowel). The overall PROSQOLI scale score was significantly correlated with most but not all subscale scores for a HRQOL measure comprising the EORTC QLQ C30 and the QLM-14, a trial-specific quality-of-life module [8]. Construct validity of FACT G to assess metastatic prostate cancer was suggested by positive correlation with KPS and Spitzer QLI [6].

Table 2. Continued

	Breast	Colorectal	Lung	Prostate
Responsiveness	FACT B and FACT G (and several subscales) sensitive to 2-month changes in PSR scores [6, 7]. With KPS as benchmark, EORTC QLQ-BR23 responsive over time only to side effects (a Dutch sample), body image (a Spanish sample); not responsive in a U.S. sample [7]. The CARES (a general cancer instrument) showed significant HRQOL improvement from cancer diagnosis to 1 year post [6].	In analysis of patients undergoing either radiation or chemo, with performance status change as a basis for judging patient change over time, EORTC QLQ-CR38 subscales generally detected treatment-specific patient-level changes as expected [10]. Similar analyses of responsiveness of FACT G, FACT C, and the FACT-based TOI-PFC yielded broadly comparable findings: subscale scores, and also total scores, generally changed over time as predicted by changes in patients' separately measured performance status scores.	EORTC QLQ-L13 scores changed during treatment concurrently with symptom decrease and toxicity increases during treatment. FACT L sensitive to changes in performance status over 2 months post-treatment, while LCSS able to detect symptom improvement following treatment [9]. EORTC QLQ-C30 did not detect pre- post-treatment difference overall, but did find significant changes along certain scales and globally once patients grouped by performance status [6].	Only two of the nine prostate-specific instruments (FACT P and the Clark-Talcott measure) had reported evidence supporting responsiveness, and one other (by Dale and colleagues) yielded data showing (point-in-time) sensitivity [8]. In particular, patients whose performance status did not decline over time had higher scores on FACT P and most subscale scores [8].For SF-36 (a generic instrument), significant improvement in the Role-Emotional scale, but not in other scales, for patients undergoing conformal radiation.
Interpretability	Kemmler et al. [15] directly compared FACT G and EORTC QLQ-C30, and concluded they measure significantly different aspects of HRQOL and their scores should not be directly compared. Effect sizes for EORTC QLQ-BR23 found to be moderate to large [7]. Scoring norms for FACT B are available from a validation study [7].	Both EORTC QLQ-CR38 and FACT C appear promising, but value as outcome measures will become clearer with experience. Effect sizes have been calculated and discussed for FACT C applied to advanced stage CRC patients. General issue is the ability to adequately capture the full range of CRC disease and treatment side-effects, including for chemotherapy [10].	Normative data for interpreting significance of a scale score or change in score available for EORTC QLQ-LC13, FACT L, and LCSS. For FACT L, a 2-point change is claimed to be a clinically meaningful difference; not explicitly defined for other two instruments [9].	For none of the nine prostate-specific instruments was there a reported effort to determine what constitutes a clinically meaningful difference in scores [8].
Burden and alternative modes of administration	Most instruments are designed to be self-administered, though in practice may also be completed via face-to-face interview. Mean administration time ranges from 5 to 18 min, with most 10 min or less [6, 7].	When self-administered, FACT C requires 5–10 min to complete. Completion time for self-administered EORTC QLQ-CR38 was just over 10 min in one study, and reported in another study to range from about 10 to 20 min [10].	EORTC QLQ-LC13, FACT L, and LCSS all designed for self-administration, though LCSS requires face-to-face interview to discuss visual analog scale. Average completion times, respectively, are 11–12, 10, and 5–8 min [9].	All nine prostate instruments designed to be self-administered. Average completion time 10–15 min (for the three instruments for which data reported) [8].

Table 2. Continued

	Breast	Colorectal	Lung	Prostate
Cultural and language adaptation	At least five HRQOL instruments that have been translated into multiple languages meet minimal criteria for psychometric performance: FLIC, CARES and CARES SF, EORTC QLQ C30, and FACT G [36]. Choice among these depends on the requirements for the study at hand. Need remains for improved standardization of translation and validation process, and good guidelines for instrument use. So far, these objectives most closely met by EORTC and FACT [36]. All seven generic and all five general cancer HRQOL instruments assessed by Erickson [6] are available in multiple languages (with at least 33 versions of the FACT G, 28 of the EORTC QLQ C30, and 50 of the SF-36).			
	EORTC QLQ-BR23 is available in at least 16 languages; FACT B, in at least 13 [7].	EORTC QLQ-CR38 is available in at least ten languages with at least six more in process; FACT C in at least 22 languages [10].	EORTC QLQ-L13, FACT L, and LCSS are available in at least 22, 20, and 12 languages, respectively [9].	Two of the nine prostate-specific instruments available in multiple languages (25 for FACT P, and 5 for UCLA-PCI) [8].

[a]Instruments cited in table:
CARES = Cancer Rehabilitation Evaluation System; CARES SF = CARES Short Form; EORTC QLQ-C30 = European Organization for Research and Treatment of Cancer Quality of Life Questionnaire; EORTC QLQ-BR23 = EORTC QLQ C30 + 23-item breast cancer module; EORTC QLQ-CR38 = EORTC QLQ C30 + 38-item colorectal cancer module; EORTC QLQ-LC13 = EORTC QLQ C30 + 13-item lung cancer module; EPIC = Expanded Prostate Index Composite; FACT G = Functional Assessment of Cancer Therapy – General; FACT B = FACT G + (10-item) breast cancer module; FACT C = FACT G + (10-item) colorectal cancer module; FACT L = FACT G + (10-item) lung cancer module; FACT P = FACT G + (12-item) prostate cancer module; FLIC = Functional Living Index – Cancer; HADS = Hospital Anxiety and Depression Scale; KPS = Karnofsky Performance Scale; LCSS = Lung Cancer Symptom Scale; POMS = Profile of Mood States; PSR = Eastern Cooperative Oncology Group (ECOG) Performance Status Rating; PROSQOLI = Prostate Cancer Specific Quality of Life Instrument; QLM-P14 = Quality-of-Life Module used in conjunction with EORTC QLQ-C30 for prostate HRQOL assessment; RSCL = Rotterdam Symptom Check List; SF-36 = Medical Outcomes Study Short-Form 36; Spitzer QLI = Spitzer Quality of Life Index; TOI-PFB = Trial Outcome Index – Physical Well-Being, Functional Well-Being, and Breast Cancer Subscale (based on FACT B); TOI-PFC = Trial Outcome Index – Physical Well-Being, Functional Well-Being, and Colorectal Cancer Subscale (based on FACT C); ULCA PCI = UCLA Prostate Cancer Index.

Table 3. HRQOL instrument[a] performance along the cancer continuum

	Prevention & screening	Survivorship	End of life	Caregiver impact
Conceptual and measurement model	Main points from Table 2 apply to HRQOL assessment across the cancer continuum. Substantial observed differences across the cancer continuum in the domain specification and item content of current instruments suggest that the appropriate conceptual model for HRQOL assessment may likewise vary across the continuum.			
Reliability	Not reported	For 14 selected generic and general cancer HRQOL instruments identified as commonly used in survivorship studies, internal consistency reliability was generally in the acceptable range, but varied substantially across instruments and also for subscales within a given instrument [11]. For example, for subscale analyses of the EORTC QLQ-C30 and FACT G, Cronbach's α ranged from 0.65 to 0.92 and from 0.56 to 0.89, respectively, in survivorship studies. Test–retest reliability ranged from $r = 0.92$ for the FACT G to $r = 0.51$–0.67 for the CES-D, with most reported correlations above 0.70 [11].	For six instruments often adopted to assess HRQOL at EOL, adequate internal consistency reliability has been reported. Cronbach's α ranges from 0.93 for the COH-QOLS to 0.77 for the MVQOLI, with the BHI, MQOL Revised, ESAS, and MSAS having α values (or ranges of values) intermediate to these [12]. Test–retest reliability correlations ranged from 0.89 (COH-QOLS) to the 0.58–0.63 range (for BHI), with r varying from 0.86 to 0.45 for the ESAS depending on time interval between assessments [12].	Among the more than 50 instruments identified in studies to measure the impact of caregiving on the caregiver, the CRA and CQOLQ were selected for detailed evaluation [13]. For each there is evidence that reliability is adequate. Cronbach's α ranged from 0.80 to 0.90 in an initial validation study of the CRA and was at similar levels in most other published studies, though was about 0.70 when a Dutch translation of the instrument was used. The CQOLC produced α's in the 0.87–0.91 range, and also a test–retest reliability correlation $r = 0.95$ (over a 14-day period) [13].
Validity	In studies of the impact of chemoprevention, genetic testing, or screening on the individual's short-term psychological well-being (e.g., anxiety, depression), symptom status, or overall health status, most findings have been derived from clinical trials or other non-population-based studies. Generalizability of results to the overall population can be challenged [18].	Regarding content validity, for cancer survivors HRQOL is a multi-dimensional construct, and several generic and general cancer instruments (e.g., SF-36, FACT G, EORTC QLQ-C30) show promise by tapping into a broad spectrum of effects; however, they do not cover such survivor-specific issues as fear of recurrence, chronic physical compromise, or post-traumatic growth [11]. Regarding construct validity, for 12 of the 14 instruments, there was evidence supporting either concurrent or divergent validity. For 9 of the 14, known-groups analyses were reported to support a claim of either construct or criterion validity [11].	Item content for all six instruments generally derived in consultation with clinical experts and HRQOL researchers, and often further informed by analyses of published literature [12]. Efforts to demonstrate construct validity have included factor analyses to confirm domain structure (BHI); correlational analyses to show either convergent or divergent validity (MVQOLI, MQOL Revised, and ESAS); known-groups comparisons (MSAS); and multiple regression analysis to identify a set of variables exhibiting convergent/divergent behavior vis á vis the HRQOL measure (COH-QOLS) [12].	Content validity of CRA and CQOLC supported by the approaches used to develop them. For CRA, items generated based on indepth interviews with caregivers, following literature review. For CQOLC, items produced through iterative process involving caregivers, patients, and health care professionals [13]. For each instrument, evidence of construct validity based on correlations in the expected directions with similar measures (convergent validation). In addition, domain structure of CRA supported by exploratory/confirmatory factor analysis [13].

Table 3. Continued

	Prevention & screening	Survivorship	End of life	Caregiver impact
Responsiveness	Generic HRQOL measures (e.g., SF-36) may not be sufficiently responsive to detect short-term effects. Instead, need to investigate utility-based measures to capture and value net impact of multiple effects: anxiety, relief, reassurance, discomfort, time costs of intervention [18].	For 5 of the 14 instruments analyzed, there was evidence to support the claim that changes in survivor functioning could be detected over time [11].	Supportive evidence was reported for each of the six instruments, with variation in the type of evidence adduced. BHI sensitive to multiple symptom changes in the terminally ill; MVQOLI scores align with behaviors as patients approach death; MQOL Revised able to predict patient-rated good-average-bad days over time; graphical analyses of ESAS scores suggest responsiveness to patient change; COH-QOLS responsive to symptom changes; and MSAS shown to correlate with treatment-induced tumor response [12]. In addition, ESAS claims criterion validation through positive correlations with KPS and validated subscales of FACT G and MSAS [12].	In one cross-sectional sample, CQOLC was significantly negatively correlated with both ECOG PSR and with number of treatment modalities patient received; however, in another sample, correlation between CQOLC and performance status was low. For CRA, instrument developers say factor analyses showing domain structure is longitudinally stable and supports CRA's suitability for measurement of change over time [13].
Interpretability	Little research on what constitutes clinically meaningful or important difference in short-term HRQOL score. However, in the context of a cost-utility analysis, clinical importance of capturing short-term utility impact of intervention may be inferred from whether doing so significantly influences overall cost-effectiveness of intervention.	Important to consider using generic and general cancer instruments for which population norms data are available, to capture impact of multiple influences, including comorbidities, on HRQOL and to facilitate comparisons of cancer survivor HRQOL with general population; such instruments include SF-36, FACT G, EORTC QLQ-C30.	Modest-to-small samples of normative data available for several instruments (COH-QOLS, BHI, ESAS), but efforts still ongoing to compare and interpret HRQOL scores across different patient and population groups. Consideration of clinically meaningful score changes reported for MQOL Revised, ESAS, BHI, MVQOLI) [12].	No published discussions of such issues as need for normative data, or what constitutes clinically important difference in scores.

Table 3. Continued

	Prevention & screening	Survivorship	End of life	Caregiver impact
Burden and alternative modes of administration	Most major generic HRQOL instruments are self-administered and require 10 min or less to complete [6]. Direct utility assessment of individual short-term impacts very time consuming [18], but application of preference measurement "systems" (e.g., HUI, QWB, EQ-5D) requires about 2–10 min (respondent self-administered) [19].	Most of the 14 generic and general cancer instruments noted above are self-administered and require about 10–15 min to complete [6].	All six instruments are designed to be self-administered or completed with assistance of interviewer. Time to complete ranges from 20 to 30 min (COH-QOLS, MSAS), to 10 min or less (MVQOLI) [12].	Both CRA and CQOLC are self-administered, requiring about 10 min to complete. (Dutch variant of CRA was administered through face-to-face interview.) [13]
Cultural and language adaptation	Main points from Table 2 apply regarding adaptation of HRQOL instruments across languages and cultures.			
	Most major generic HRQOL instruments [6] and preference measurement systems [19] are available in multiple languages.	Several of the 14 instruments are available in multiple languages, including the SF-36, EORTC QLQ C30, FACT G, and FLIC [6].	All six instruments available in English and two also in one other language (COH-QOLS, Spanish; and MQOL Revised, French) [12].	Both CRA and CQOLQ developed in English, with CRA also available in Dutch [13].

[a]Instruments cited in table:
BHI = Brief Hospice Inventory; CES-D = Center for Epidemiologic Studies – Depression; COH-QOLS = City of Hope Quality-of-Life Scale; CQOLC = Caregiver Quality of Life Index – Cancer; CRA = Caregiver Reaction Assessment; EORTC QLQ-C30 = European Organization for Research and Treatment of Cancer Quality of Life Questionnaire; ESAS = Edmonton Symptom Assessment Scale; FACT G = Functional Assessment of Cancer Therapy – General; FLIC = Functional Living Index – Cancer; MQOL, Revised = McGill Quality of Life Questionnaire; MSAS = Memorial Symptom Assessment Scale; MSQOLI = Missoula-VITAS Quality of Life Index; SF-36 = Medical Outcomes Study Short-Form 36.

(e.g., between having diarrhea and the resulting activity limitation);
- distinguishing more clearly between objectively measured health status and subjectively measured quality of life (positively correlated, but not synonymous);
- examining further the phenomenon that HRQOL scores differ (all else equal) depending on whether the instrument defines quality of life in terms of perceived status, the evaluation of that status, or both;
- investigating patient adaptation over time (response shift) and the resulting influence on HRQOL scores and their interpretation;
- conducting more head-to-head comparisons of HRQOL instruments purporting to measure similar constructs; and (in general)
- demonstrating in real-world applications how a conceptual model of HRQOL can provide the starting point for HRQOL instrument development.

Reliability

The HRQOL literature in cancer still reflects almost entirely the Classical Test Theory perspective, with Cronbach's α as the measure of internal consistency and some variant of the intraclass correlation coefficient (r) to index the strength of agreement or reproducibility; see Table 1. COMWG analyses [6–13] generally found Cronbach's α and reproducibility coefficients for HRQOL questionnaires to be within the commonly accepted ranges for adequate performance; see Tables 2 and 3. From a MOT perspective, this means α estimates greater than 0.70 for group comparisons and 0.90 for individual-level comparisons [2].

As the MOT framework recognizes [2] and as Reise [20] emphasized, each of these traditional CTT reliability measures is computed as a summary statistic for the measurement scale as a whole. The strong possibility that reliability – either internal consistency or reproducibility – might vary along the scale, or even by item, cannot be addressed through CTT approaches.

IRT modeling, on the other hand, acknowledges the possibility that the reliability of a survey item may vary depending on the particular level of the HRQOL construct being measured by the scale. For example, for an individual with severe mobility limitations, survey items asking about the ability to move about the bedroom or house will generally provide more useful information for identifying the individual's position along the physical functioning scale than items asking about the ability to walk a golf course or to run a mile. As Hambleton [21] discussed, IRT modeling tailors the selection of survey items to the particular individual being scaled, so that fewer items are needed to achieve any given level of scoring precision (or, alternatively, greater precision can be achieved for any given number of items asked).

In IRT, the reliability of each item is represented by an "information function," whose inverse reflects the item's standard error of measurement [20, 21]. These functions can be combined across items to derive a summary measure of the reliability of the entire scale. At the same time, the functions can be analyzed individually to determine if there are regions of the scale where measurement precision could be improved through additional item development.

Validity

Defined as the extent to which an instrument measures what it claims to measure, validity assumes a tripartite classification in the MOT framework: content, criterion, and construct; see Table 1.

Content validity

As indicated in Tables 2 and 3, a number of the HRQOL instruments commonly used in studies of cancer treatment [6–10], survivorship [11], end of life [12], and caregiver impact [13] exhibit strong content validity, according to their developers and users. To guide the creation and validation of survey items, developers have conducted (1) focus groups involving patients, providers, researchers, or the lay public; (2) critical reviews of the item content of previously developed instruments; and (3) surveys of the pertinent scientific literature [6].

Nonetheless, several COMWG analyses cited important opportunities for improvement. Moinpour and Provenzale [10] concluded that for patients undergoing colorectal surgery, post-operative effects involving diet, social, and sexual functioning may not be adequately covered in

existing instruments. Zebrack and Cella [11] found that current multidimensional HRQOL constructs may not capture certain elements important to cancer survivors, such as fear of recurrence or chronic physical compromise. Williams, a prostate cancer survivor, observed that current HRQOL instruments fail to measure the depth of suffering faced by patients and their families [22].

Citing one approach to enhancing content validity, Erickson [6] noted that at least two prominent instrument systems combine general cancer measures with modules specific to cancer disease site, symptom complex, and toxicity effects. The systems cited were the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ) set, and the Functional Assessment of Chronic Illness Therapy (FACIT). She urged expansion of such a modular strategy to create a broad HRQOL "measurement system" containing content-appropriate instruments for the entire cancer continuum. Barry and Dancey [23] advanced a similar strategy for developing instruments to capture the impact of specific cancer therapies, including their side effects, on HRQOL.

Criterion validity

In fact, criterion validity has played only a minor practical role in the assessment of quality-of-life instruments, "since they measure postulated constructs that are experimental and subjective [24, p. 298]." Similarly, the MOT notes that criterion validation is rarely pursued for self-reported health status measures, "because of the absence of widely accepted criterion measures..." [2, p. 200].

In COMWG analyses, criterion validation was infrequently reported; and when it was, the criterion selected was generally open for discussion. For example, developers of the Lung Cancer Symptom Scale (LCSS) claimed criterion validity on the basis of significant positive correlations with *other* HRQOL instruments [9].

Construct validity

In the absence of a consensus criterion, construct validation becomes the primary means for determining whether a measure provides an unbiased assessment of the construct of interest.

COMWG members found substantial evidence of construct validity for virtually all of the HRQOL instruments selected for intensive analysis; see Tables 2 and 3. Frequently, such evidence took the form of convergent validation, e.g., the strong positive correlation between the EORTC QLQ-Lung cancer (LC13) scores and performance status for lung cancer patients [9]. Also reported were successful efforts at divergent validation, e.g., showing that the Functional Assessment of Cancer Therapy-General (FACT G) was poorly correlated with the Marlowe-Crown Social Desirability Scale in breast cancer patients [7]. Sometimes convergent and divergent validations were pursued concurrently, e.g., by examining the pattern of inter-scale correlations for the FACT Colorectal cancer (FACT C) and the Profile of Mood States in colorectal cancer patients [10]. Researchers have often employed factor analytic approaches, e.g., to the Rotterdam Symptom Checklist (RSCL) [6] and Brief Hospice Inventory [12], to confirm the presence of hypothesized constructs. At the developmental stage, a comparison of scale scores with clinical data from medical records may yield valuable information for assessing construct validity [6].

A number of challenges remain in identifying valid approaches to HRQOL assessment in one particular part of the cancer continuum: prevention and screening for asymptomatic individuals [18]. Assessing the short-term HRQOL impacts of chemoprevention, genetic testing, and screening for disease has been a focus of research in this part of the continuum. One important issue is whether HRQOL findings reported so far, which come largely from clinical trials or other non-population-based studies, are generalizable to the broad community of individuals at risk to cancer [18].

The findings summarized in Tables 2 and 3 derive from a cancer outcomes research literature dominated by CTT-based HRQOL measures, and it is natural to ask whether IRT-based measures would be expected to demonstrate stronger construct validity. As Reise noted, research on this topic remains in its early stages [20]. Still, some additional observations are possible, building on Reise's COMWG chapter and other recently published work. First, in a given respondent sample, IRT-based and CTT-based HRQOL scores may be very highly correlated. An informal

survey of applications by Reise and Henson [25] indicated that correlations above 0.98 are ''routinely'' encountered. Similarly, McHorney et al. [26] reported the correlation between Rasch-based (IRT) and Likert-based (CTT) scores on the Physical Functioning Scale in the Medical Outcomes Study to be 0.98 (see their Figure 1, p. 455). Consequently, to the extent that one attempts to demonstrate construct validity by correlating HRQOL scores with some external variable or measure (e.g., labor force participation rates), it may make little difference in practice whether an IRT or CTT approach is used. (Two highly correlated measures will likely be similarly correlated with any selected third measure.)

That said, Reise and Henson [25] emphasized that, ''...the optimal scaling of individual differences with IRT can make a difference in practice and can dramatically change substantive conclusions'' (p. 100). They cite several studies in which IRT modeling essentially clarified or corrected CTT-generated problems in the analysis of individual differences in trait level. Likewise, McHorney et al. [26] identified several potential advantages of the Rasch variant of IRT modeling for interpreting HRQOL scores at the individual (not group) level.

Responsiveness

All COMWG chapters analyzing HRQOL measure performance reported on responsiveness or sensitivity, and the findings were generally (but not always) positive for the instruments reviewed.

For example, Litwin and Talcott [8] cited evidence that the FACT Prostate cancer (FACT P) is responsive based on concurrent changes in recorded performance status for prostate cancer patients. Moinpour and Provenzale [10] reported that EORTC QLQ-Colorectal cancer (CR38) scores varied as hypothesized with changes over time in the performance status of colorectal cancer patients. Ganz and Goodwin [7] found that the FACT Breast cancer (FACT B) and the FACT G (and several of its subscales) were sensitive to 2-month changes in Eastern Cooperative Oncology Group (ECOG) Performance Status Ratings. Earle and Weeks [9] noted that the EORTC QLQ-LC13 and LCSS were both responsive to treatment-related changes in symptom or toxicity status, and that the FACT Lung cancer (FACT L) was sensitive to changes in ECOG performance status. Ferrell [12] summarized encouraging results for several instruments measuring HRQOL at end of life, including the concurrent movement of subscales of the City of Hope Quality of Life Scale with symptom changes.

Snyder [13] found studies reporting somewhat conflicting evidence on whether the Caregiver Quality of Life Index – Cancer responds as might be expected with observed changes in patient performance status. Mandelblatt and Selby [18] concluded that generic HRQOL measures may not be sufficiently responsive to detect the short-term impacts of prevention or screening interventions. They recommend, instead, that utility-based measures may be required to capture the net impact of such diverse effects as anxiety, relief, reassurance, discomfort, and the time costs of participation; for a discussion of the application of such utility measures to cancer, see Feeny [19].

In the MOT framework, responsiveness is largely about the performance characteristics of an instrument – its ability to detect differences over time, even ''small'' ones [2, p. 201]. But as COMWG members frequently asked in their deliberations, how small is still large enough to matter? As the MOT criteria acknowledge, responsiveness and interpretability are related concepts. Assessing the responsiveness (or sensitivity) of an instrument involves two tasks: determining whether there is a statistically significant difference in treatment effects over time (or cross-sectionally), and, if so, determining whether the observed difference is large enough to be clinically important. The latter inquiry requires a clear understanding of how to interpret the meaning and significance of instrument change scores, as discussed just below.

Although these points apply in principle whether the measurement model is IRT- or CTT-based, at least one recent analysis of migraine headache trials found impressive evidence that IRT-based HRQOL measures are more responsive than CTT-based measures [27].

Interpretability

The MOT attribute of interpretability proved to be very challenging for the COMWG, as can be

inferred from Tables 2 and 3. Several analyses, including most notably Erickson's assessment of generic and general cancer measures [6], identified population norms for some instruments that could be used to interpret the scores assigned to cancer patients (for example, for the Medical Outcomes Study Short-Form 36 (SF-36)). Others COMWG analyses identified studies in which scores on a particular HRQOL measure were positively (or negatively) associated with scores on some other outcome measure, thus providing interpretative information, e.g., the Ganz and Goodwin [7] summary of a study [15] concluding that the EORTC QLQ-Core instrument (C30) and FACT G measure significantly different aspects of HRQOL. In a few cases, instrument developers or users focused on effect size as a means to interpret instrument performance [10].

In general, however, the HRQOL literature reviewed by COMWG members was oriented to uncovering clinically interesting findings, which frequently would be reported – and also interpreted – by the original study authors as "significant" if they met conventional thresholds of *statistical* significance. Still, two particularly important issues were identified in the COMWG's deliberations:

Defining meaningful and important differences in an outcome measure

The most prominent interpretative challenge is determining whether a given observed change in a latent variable construct like HRQOL is "meaningful" or "important." [In fact, a Clinical Significance Consensus Meeting Group (ClinSig), working in parallel but independently of the COMWG, has produced a six-paper monograph exploring aspects of this issue [28].]

Writing for the COMWG, Osoba [29] noted the two major approaches for demonstrating clinical importance in a HRQOL measure: distribution-based and anchor-based. In the former, one judges the meaningfulness of a HRQOL change score according to some summary statistic internal to the measurement process itself, e.g., a change greater than one-half the standard deviation in the distribution of change scores. In anchor-based approaches, one judges the meaningfulness of a change score by how well it accords with parallel changes in other, readily interpretable measures hypothesized to relate to HRQOL. Osoba concluded that:

- With anchor-based approaches, a "small" perceptible change to patients in physical, social, or emotional functioning, or in global HRQOL, appears to be about 7% on average (and ranging from 5% to 10%) of the scale breadth.
- This 7% (range from 5% to 10%) perceptible difference level is consistent in magnitude with what Cohen [30] and others have termed a "small" to "moderate" effect size, which is often regarded as approximating the "minimum important difference" (MID).
- Consequently, there is preliminary evidence that the MID is approximately the same whether derived from anchor-based or distribution-based approaches.

Understanding response shift

As analyzed by Schwartz and Sprangers [31, 32], response shift occurs when the very meaning of an individual's evaluation of a construct like HRQOL changes over time. This is said to reflect changes over time in the individual's (1) internal standards of measurement of the construct, (2) valuation of the domains comprising the construct, or (3) definition or perception of the construct itself. Ferrans [14] concluded that a deeper understanding of response shift could open the way to conceptual models that better account for the complex dynamic between changes in "objective" biomedical outcomes and comparatively more malleable measures of HRQOL. To the extent a response shift is at work, it will almost certainly influence both the responsiveness of a HRQOL measure and the interpretation of evidence supporting construct (or criterion) validity. Further investigation of these challenging issues is clearly warranted [33].

Burden and alternative modes of administration

These two MOT attributes will be discussed jointly (see Table 1), since one of the main messages from the COMWG's psychometric analyses is that item response theory modeling opens the way for a mode of administration – built around item banking and computer-adaptive testing (CAT) – that may reduce respondent burden significantly.

COMWG psychometricians noted that IRT modeling provides the only sound theoretical basis for the construction of item banks, which are prerequisite for CAT [20, 21]. Computer-adaptive approaches permit the researcher to obtain any given level of HRQOL measurement precision with fewer questions asked than with traditional fixed-item survey forms (the only form of instrumentation provided for under Classical Test Theory). Similarly, for any given number of scale items posed to the respondent, a more statistically precise score can be computed under CAT than with fixed-item instruments. As indicated in Tables 2 and 3, across the array of HRQOL fixed-item measures reviewed by COMWG members, average administration time generally varied from about 5 and 15 min [6–11, 13]. Some instruments commonly used in end-of-life studies require 20–30 min to complete [12], possibly owing to the disability in the population. Virtually all of the HRQOL instruments cited in these tables are designed for self-administration, though some are readily adaptable for interviewer administration.

The degree to which item banking and CAT can reduce respondent burden, increase measurement precision, or both will be tested in the months ahead, given the 2004 launch by the U.S. National Institutes of Health of a 5-year, $25 million initiative to develop the Patient Reported Outcomes Measurement Information System (PROMIS) [34]. PROMIS will develop public domain item banks and CATs for selected health symptom and HRQOL domains affected by a variety of chronic diseases, including cancer. The arenas of application will include clinical trials, other research studies, and eventually patient care assessment.

The construction of items banks for CAT is a complex undertaking, involving IRT evaluation and calibration of possibly hundreds of candidate items and the development of new items to fill gaps over certain ranges of the HRQOL scale continuum [21]. Very likely there are significant economies-of-scale in the development, ongoing maintenance, and periodic updating of item banks. Yet, difficult issues may arise regarding ownership and intellectual property rights [35]. Moreover, it is important to foster an intellectually open, vibrant research environment that encourages work on new survey items, analytical approaches, and strategies to improve item banking over time [4]. Such developments could enhance the attractiveness of routinely incorporating patient-reported outcomes assessment as one component of population-based cancer outcomes research [36].

Even as novel data collection approaches are explored, practical issues related to traditional pencil-and-paper approaches to HRQOL data collection still require close attention, Fairclough emphasized [17]. For example, under what circumstances do various approaches to data collection (self-administered questionnaires, in-person interviews, over-the-phone interviews) yield convergent or divergent results? How do these results vary by the clinical and socio-demographic characteristics of respondents? How might multiple modes of administration be used in concert to minimize the likelihood of missing data? Fairclough proposed ''piggy-backing'' such inquiries onto clinical trials or other studies already collecting HRQOL data.

Cultural and language adaptations

For cross-cultural assessment of HRQOL in clinical oncology, Aaronson [16] concluded there are at least five instruments meeting minimum criteria for psychometric performance when used in group comparisons: the Functional Living Index – Cancer (FLIC), Cancer Rehabilitation Evaluation System (CARES) and CARES Short Form (CARES SF), RSCL, FACT G, and the EORTC QLQ-C30. Although the instruments are very broadly similar, they have substantially different item content, as well as strengths and limitations that vary with the application at hand. He urged greater standardization and monitoring of the instrument translation process, and the development of guidelines for ensuring higher quality products. At present, this has been done to a greater extent for the EORTC QLQ-C30 and FACT.

A potentially salient threat to validity in this context is differential item functioning (DIF), wherein the psychometric performance of a given survey item differs systematically across cultural or geographic settings [16, 35]. In instances where DIF is detected (whether through CTT or IRT approaches), researchers can consider how to correct for possible racial/ethnic/cultural biases that can skew study findings.

Concluding observations

As the National Cancer Institute has emphasized, a major aim of cancer outcomes research is to identify, and develop as needed, measures of patient-reported outcomes that are *methodologically sound* and *provide substantial value to a range of decision makers* [1, 4]. The Medical Outcomes Trust framework provides compelling guidance for evaluating and improving the scientific quality of PRO measures, most specifically for HRQOL. For that reason, NCI's Cancer Outcomes Measurement Working Group adopted the MOT attributes and review criteria to guide its assessment of the state of the science in cancer outcomes measurement. The previous seven sections have summarized major findings and recommendations from the COMWG, organized according to the MOT attributes and presented with the MOT review criteria in mind.

The COMWG's work builds from, and capitalizes on, a literature in cancer outcomes measurement that is growing in quantity and quality. The central tasks for working group members entailed reviewing, synthesizing, and evaluating this literature, as a springboard for making recommendations to the NCI. Perhaps never before has this much credible information been brought together about the methodological soundness, and shortcomings, of outcomes measurement in a major disease area.

At the same time, a close reading of *Outcomes Assessment in Cancer* [3] – and a careful review of the published literature generally – will yield very little hard information about the perceived value of PRO data to cancer decision makers: patients, survivors, families, providers, payers, regulators, or those establishing quality-of-care standards. Likewise, little is known formally about the role of PRO data in the actual planning, execution, or appraisal of cancer-related decisions – whether by the patient and her provider, by the drug formulary manager, or by organizations establishing treatment guidelines. [A significant exception arises with the U.S. Food and Drug Administration, which has assessed the role of PROs in cancer drug approval [37]. The FDA recently issued a draft guidance on the appropriate development and use of PROs in industry-sponsored studies to support product approval [38].]

Consequently, additional research is needed on assessing and enhancing the perceived scientific credibility and usefulness of PRO measures to the full range of cancer decision makers. Work should proceed on two broad fronts: (1) *prospectively designed studies* to examine the strengths and limitations of PRO measures, using the MOT framework, in a variety of research settings (beyond randomized clinical trials); and (2) *in-depth investigations*, including case studies, of the roles that PRO measures do play, or could play, in real-world decision making.

Such studies would draw upon a range of disciplines and perspectives: psychology (not only advanced psychometrics [39], but cognitive and behavioral approaches); economics [40, 41]; statistics [42]; and the decision sciences [43]. In planning and carrying out these PRO studies, investigators should consider:

- Giving strong emphasis to IRT modeling, for reasons discussed at multiple points in this paper.
- Applying structural equation modeling to facilitate a rigorous analysis of the conceptual model – measurement model relationship. Such models facilitate investigation of multiple cause–effect relationships and interaction effects, and can provide a coherent framework for conducting construct validation [44].
- Collecting preference-based and non-preference-based measures of HRQOL on the same set of respondents. Doing so would set the stage for a variety of cross-validation analyses, provide a possible means for aggregating multidimensional latent variable models of utility (see the discussion by Wilson [39]), and facilitate the conduct of cost-utility analyses [40].
- Pursuing quantitative and qualitative analyses (e.g., IRT modeling and cognitive interviewing) in tandem to improve understanding of the interplay between instrument content, measures of psychometric performance, and the perceived usefulness in real applications [45]. The spotlight would be not only on HRQOL, but other patient-reported information, including perceptions of and satisfaction with care [46] and assessment of patient needs [47].

The importance of both rigor and relevance in outcomes measurement – and of not assuming that

achieving one guarantees the other – has been emphasized by Dowie [48]. He distinguishes ''knowledge validity'' – does the outcome measure in fact measure what it intends to? – from ''decision validity'' – does the measure provide the necessary information for the decision at hand? Complementary discussions are found in the COMWG analysis by Revicki [49], and the invited contributions from Spilker [50] and the Health Outcomes Committee of the Pharmaceutical Research and Manufacturers of America (PhRMA) [51].

Indeed, few would dispute the joint importance of rigor and relevance in cancer outcomes measurement. The challenge ahead lies in creating an inter-disciplinary agenda – drawing from the measurement, behavioral, and social sciences – that advances the field on both fronts. Although the focus throughout this paper has been exclusively on cancer, we strongly suspect that outcomes assessment within and across *all* disease domains would benefit significantly from a research agenda that creates new synergisms between HRQOL researchers and decision makers.

Appendix: NCI Cancer Outcomes Measurement Working Group, 2001–2004

Co-Chairs

Joseph Lipscomb, PhD
Chief, Outcomes Research Branch, Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute

Carolyn C. Gotay, PhD
Professor, Cancer Research Center of Hawai'i
University of Hawai'i

Working Group Initiator

Claire F. Snyder, PhD
Expert, Outcomes Research Branch, Applied Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute

Working Group Participants

Neil K. Aaronson, PhD
Head, Division of Psychosocial Research & Epidemiology
The Netherlands Cancer Institute, Professor, Faculty of Medicine
Vrije Universiteit

Michael J. Barry, MD
Chief, General Medicine Unit
Massachusetts General Hospital

David Cella, PhD
Professor of Psychiatry and Behavioral Science
Northwestern University Feinberg School of Medicine
Director, Center on Outcomes Research and Education
Evanston Northwestern Healthcare

Janet E. Dancey, MD
Senior Clinical Investigator, Investigational Drug Branch, Cancer Therapy Evaluation Program
Division of Cancer Treatment and Diagnosis
National Cancer Institute

Charles Darby
Social Science Administrator
Agency for Healthcare Research and Quality

Craig C. Earle, MD, MSc
Assistant Professor of Medicine, Harvard Medical School
Dana-Farber Cancer Institute

Pennifer Erickson, PhD
Associate Professor, Departments of Biobehavioral Health and Health Evaluation Sciences
Pennsylvania State University

Diane L. Fairclough, DrPH
Professor, Colorado Health Outcomes Center and
Department of Preventive Medicine and Biometry
University of Colorado Health Sciences Center

David H. Feeny, PhD
Professor of Pharmacy and Pharmaceutical Sciences
Departments of Economics and Public Health Sciences
University of Alberta

Carol Estwing Ferrans, PhD, RN, FAAN
Professor, College of Nursing,
University of Illinois at Chicago

Mark Wilson, PhD
Professor, Graduate School of Education
University of California, Berkeley

Brad Zebrack, PhD, MSW
Cancer Survivor/Advocate
NCI Director's Consumer Liaison Group
Assistant Professor, School of Social Work
University of Southern California

References

1. U.S. National Cancer Institute. Cancer Outcomes Measurement Working Group. Retrieved February 7, 2006, from http://www.outcomes.cancer.gov/methods/measures/comwg/.
2. Scientific Advisory Committee of the Medical Outcomes Trust (Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, Stein REK). Assessing health status and quality-of-life instruments: Attributes and review criteria. Qual Life Res 2002; 11: 193–205.
3. Lipscomb J, Gotay CC, Snyder C. Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005.
4. Lipscomb J, Donaldson MS, Arora NK, et al. Cancer outcomes research. J Natl Cancer Inst Monogr 2004; 33: 178–197.
5. Gotay CC, Lipscomb J, Snyder CF. Reflections on findings of the Cancer Outcomes Measurement Working Group: Moving to the next phase. J Natl Cancer Inst 2005; 97: 1568–1574.
6. Erickson P. Assessing health status and quality of life of cancer patients: The use of general instruments. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 31–68.
7. Ganz PA, Goodwin PJ. Quality of life in breast cancer – what have we learned and where do we go from here? In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 93–125.
8. Litwin MS, Talcott JA. Measuring quality of life in prostate cancer: Progress and challenges. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 126–159.
9. Earle CC, Weeks JC. The science of quality-of-life measurement in lung cancer. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 160–177.
10. Moinpour CM, Provenzale D. Treatment for colorectal cancer: Impact on health-related quality of life. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 178–200.
11. Zebrack B, Cella D. Evaluating quality of life in cancer survivors. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 241–263.
12. Ferrell BR. Assessing health-related quality of life at end of life. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 264–285.
13. Snyder C. Assessing the subjective impact of caregiving on informal caregivers of cancer patients. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 329–345.
14. Ferrans CE. Definitions and conceptual models of quality of life. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 14–30.
15. Kemmler G, Holzner B, Kopp M, et al. Comparison of two quality-of-life instruments for cancer patients: The Functional Assessment of Cancer Therapy-General and the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-C30. J Clin Oncol 1999; 17: 2932–2940.
16. Aaronson NK. Cross-cultural use of health-related quality of life assessments in clinical oncology. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 406–424.
17. Fairclough DL. Practical considerations in outcomes assessment for clinical trials. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 346–361.
18. Mandelblatt JS, Selby JV. Short-term outcomes of chemoprevention, genetic susceptibility testing, and screening interventions: What are they? How are they measured? When should they be measured? In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 216–240.
19. Feeny DH. The roles for preference-based measures in support of cancer research and policy. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 69–92.
20. Reise SP. Item response theory and its applications for cancer outcomes measurement. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 425–444.
21. Hambleton RK. Applications of item response theory to improve outcomes measurement: Developing item banks, linking instruments, and computer-adaptive testing. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 445–464.

22. Williams JE. Patient advocate perspective on health-related quality of life issues with prostate cancer survivors. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 286–289.
23. Barry MJ, Dancey JE. Instruments to measure the specific health impact of surgery, radiation, and chemotherapy on cancer patients. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 201–215.
24. Fayers PM, Machin D. Quality of Life: Assessment, Analysis, Interpretation. Chichester, England: John Wiley & Sons, 2002.
25. Reise SP, Henson JM. A discussion of modern versus traditional psychometrics as applied to personality assessment scales. J Personal Assess 2003; 81(2): 93–103.
26. McHorney CA, Haley SM, Ware JE Jr.. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods.. J Clin Epidemiol 1997; 50(4): 451–467.
27. Kosinski M, Bjorner JB, Ware JE Jr, et al. The responsiveness of headache impact scales scored using 'classical' and 'modern' psychometric methods: A re-analysis of three clinical trials. Qual Life Res 2003; 12: 903–912.
28. Sloan JA, Cella D, Frost MH, et al. Assessing clinical significance in measuring oncology patient quality of life: Introduction to the symposium, content overview, and definition of terms. Mayo Clin Proc 2002; 77: 367–370.
29. Osoba D. The clinical value and meaning of health-related quality-of-life outcomes in oncology. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 386–405.
30. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
31. Schwartz CE, Sprangers MAG. Adaptation to Changing Health: Response Shift in Quality-of-Life Research. Washington, DC: American Psychological Association, 2000.
32. Sprangers MAG, Schwartz CE. Integrating response shift into quality-of-life research: A theoretical model. Soc Sci Med 1999; 48: 1507–1515.
33. Gotay CC, Lipscomb J, Snyder C. Reflections on COMWG findings and moving to the next phase. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 568–583.
34. National Institutes of Health. Patient-Reported Outcomes Measurement Information System: Dynamic Tools to Measure Health Outcomes from the Patient Perspective. Retrieved February 10, 2006, from http://www.nihPROMIS.org/.
35. McHorney CA, Cook K. The ten D's of health outcomes measurement for the 21st century. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 590–609.
36. Gotay CC, Lipscomb J. Data for cancer outcomes research: Identifying and strengthening the empirical base. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 522–549.
37. Johnson JR, Williams G, Pazdur R. End points and United States Food and Drug Administration approval of oncology drugs. Journal of Clinical Oncology 2003; 21: 1404–1411.
38. U.S. Food and Drug Administration. Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (DRAFT). Available at http://www.fda.gov/cber/gdlns/prolbl.pdf. Accessed March 20, 2006.
39. Wilson M. Subscales and summary scales: Issues in health-related outcomes. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 465–479.
40. O'Brien BJ. Cost-effectiveness analysis in cancer: Toward an iterative framework for integration of evidence from trials and models. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 503–521.
41. Hornbrook MC. On the definition and measurement of the economic burden of cancer. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 480–502.
42. Sloan JA. Statistical issues in the application of cancer outcomes measures. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 362–385.
43. Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs., 2nd ed., Cambridge: Cambridge University Press, 1993.
44. Hays RD, Revicki D, Coyne KS. Application of structural equation modeling to health outcomes research. Eval Health Prof 2005; 28(3): 295–309.
45. Willis GB, Reeve BB, Barofsky I. The use of cognitive interviewing techniques in quality of life and patient-reported outcomes assessment. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 610–622.
46. Darby C. Measuring the patient's perspective on the interpersonal aspects of cancer care. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 290–304.
47. Gustafson DH. Needs assessment in cancer. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 305–328.
48. Dowie J. Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions. Health Econ 2002; 11: 1–8.
49. Revicki DA. Use of health-related quality of life measures by industry and regulatory agencies in evaluating oncology therapies. In: Lipscomb J, Gotay CC, Snyder C (eds),

Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 550–567.
50. Spilker B. The world of outcomes research: Yesterday, today, and tomorrow. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 584–589.
51. Copley-Merriman K, Jackson J, Boyer JG, et al. Industry perspective regarding outcomes research in oncology. In: Lipscomb J, Gotay CC, Snyder C (eds), Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge: Cambridge University Press, 2005: 623–638.

Address for correspondence: Joseph Lipscomb, Department of Health Policy and Management, Rollins School of Public Health, Emory University, Rm 642, 1518 Clifton Road, NE, Atlanta, GA, 30322, USA
Phone.: +1-404-7274513; Fax: +1-404-7279198
E-mail: jlipsco@sph.emory.edu