Support Care Cancer (1995) 3:11–22

SPECIAL ARTICLE

David F. Cella

Methods and problems in measuring quality of life

Presented in part at the 6th International Symposium: Supportive Care in Cancer, New Orleans, La., 2–5 March 1994

D. F. Cella, Ph. D.
Division of Psychosocial Oncology,
Rush Medical Center,
1653 West Congress Parkway,
Chicago, IL 60612-3833, USA
Fax: (312) 563-2471

Abstract The US health-care transition demands increased accountability for medical care. This has contributed to increased interest in documenting valued medical outcomes, including improvements in health-related quality of life and treatment satisfaction. These data can only be obtained validly by asking patients directly about their current health state, perception of well-being, and satisfaction with care. A core set of well-validated instruments have been developed to measure health-related quality of life in patients with cancer. As these are employed with increasing frequency, rigorous quality assurance of data collection is critical. Because of the necessity of quality control, patient-reported data collection can be labor-intensive and prohibitively costly. However, time and cost-saving methods, such as centralized telephone survey methods or on-site direct data entry via interactive computer, can guarantee high-quality data while minimizing costs. Justification of the need for these methods and a brief description are provided.

Key words Data collection methods · Quality of life Health-related quality of life Quality assurance

Introduction

The term quality of life, or health-related quality of life (HQL), has emerged to organize and galvanize a collection of outcome-evaluation activities over the past two decades in cancer treatment research. Prior to this, length of survival, regardless of its quality, was considered to be the only primary outcome in oncology treatment research. It is now widely accepted that in most circumstances quality of survival is as important as quantity of survival. This implies that a severely toxic treatment must be evaluated for its detrimental impact as well as its survival benefit. It also raises a less obvious point: treatments can be considered efficacious if they improve the quality of life even in the absence of survival benefit. Thus, investigating the impact of cancer treatments on HQL is a two-tailed enterprise where treatment toxicity is traded not only with survival time but also with post-treatment function and well-being.

Health-related quality of life (HQL) evaluation entails a multidimensional quantification of patient functional status, usually as perceived by the patient [1, 7, 13, 14, 20, 22, 25, 37, 38, 47, 52, 58]. In the decades to come, treatment-intensification strategies that increase toxicity are likely to continue, given the advent of hematopoietic growth factors and improved antiemetic regimens. This further increases the importance of evaluating toxicity, patient function, and patient preferences for treatment. HQL evaluation differs from classical toxicity ratings in two important ways: (a) It incorporates more aspects of function (e.g., mood, affect, social well-being) than those which have typically been attributed to treatment; and (b) it focuses on the patient's perspective.

Evaluating methods of assessment

Along with the evolution of interest in HQL, many efforts to measure the construct have been created and promoted. A number of validated quality-of-life measures have become accepted for use in oncology in particular [2, 3, 15, 16, 26, 43, 45] and chronic illness in general [4, 8, 24, 30, 34, 36, 49, 60]. The diversity of available measures is potentially valuable in that it provides the user with choices based upon specific characteristics of a given disease site, clinical trial, or quality-of-life domain of interest. This paper provides the reader with some understanding of criteria to evaluate whether an HQL measure is likely to perform well in a clinical trial. Suggestions that can be helpful in the preparation of protocol documents have been published elsewhere [29].

There are many definitions of HQL [11, 13, 28, 44, 53]. Different measures of HQL are not necessarily equivalent and one must therefore be clear on the dimensions of HQL as measured by a particular instrument. Definitions of HQL may differ across study groups and still be measured reliably and validly within the parameters of a definition [19, 31, 59]. For example, most agree that important HQL domains include physical, mental and social dimensions. Whereas virtually all currently accepted HQL measures provide some ability to separate physical and psychological dimensions, social functioning is much less evenly represented. Some measures cover social well-being and function more than others. For example, deHaes et al. [19] do not measure social functioning as a component, and yet this scale can be evaluated for reliability and validity within its range of items.

Approaches to measuring quality of life

Over time, two approaches to measuring HQL have evolved: psychometric and utility. These approaches have evolved relatively independently of one another, largely because they were developed within different scientific disciplines. Psychometric approaches derive from psychology whereas utility approaches derive from economics. Only recently have investigators considered integrating these two approaches. This remains a critical challenge in HQL measurement.

Psychometric approaches

The psychometric approach includes generic health profile measurement (e.g., short forms from the Medical Outcomes Study [30, 60]) and specific instruments intended to measure the multidimensional impact of a specific disease, treatment or condition (e.g., the Functional Living Index – Cancer [45]). The psychometric approach places heavy emphasis upon an individual's response and response variability across individuals. An important contribution of the psychometric approach is that it provides measurement of subjective or perceived well-being. Psychometric measures may or may not include a summary or total score. When available, only rarely have these summary scores been connected to patients' value for their current health status. This poses a problem, because without a rating of patient preference, one cannot appropriately make a decision about the value of a given treatment to a given patient. Very often, one of two patients with identical disease and treatment options will decline therapy while the other will accept it enthusiastically. Because psychometric measures typically do not incorporate patient-specific weights for individual domains nor anchor states of health to a common standard, evaluating trade-offs between quality and length of life, or between one dimension of HQL and another, is difficult. This presents a challenge in a clinical trial where the primary purpose for integrating HQL measurement is to incorporate data on the impact of treatment on both length and quality of life into conclusions about treatment efficacy. The collection of patient preferences in clinical trials would allow the effect of treatment on quality-adjusted survival as well as on conventional outcome measures to be evaluated. Further, the addition of patient preference assessments to clinical trial outcome evaluation can make it possible to distinguish patients who favor one treatment over another when both may have an equivalent survival outcome. A strategy for doing this has been described by Till and colleagues [54].

Utility approaches

In contrast to the psychometric approach, the utility approach is explicitly concerned with decisions about treatment, usually at a policy level. In this approach, treatments are typically evaluated for their benefit compared in some way to their cost. The utility approach to health status measurement evolved from a tradition of cost/benefit analysis, into cost/effectiveness approaches and, most recently, cost/utility approaches [21]. The cost/utility approach extends the cost/effectiveness approach conceptually by evaluating the HQL benefit produced by the clinical effects of a treatment, thereby including the (presumed) patient's perspective. To be used this way, HQL must be measured as a utility since, by definition, utilities can be multiplied by time to yield a meaningful quantity. Two general cost/utility methods are the standard gamble approach and the time trade-off approach [55]. In the standard gam-

ble approach, people are asked to choose between their current state of health and a "gamble" in which they have various probabilities for death or perfect health (cure). The time trade-off method involves asking people how much time they would be willing to give up in order to live out their remaining life expectancy in perfect health. All utility approaches share in common the use of a 0–1 scale in which 0 = death and 1 = perfect health. In practice, most cost/utility analyses employ expert estimates of utility weights, or in some cases, weights provided by healthy members of the general public. It is often assumed that these weights are reasonable approximations of patient preferences. However, several studies have demonstrated that utilities obtained from patients are generally higher than those provided by physicians, which are, in turn, higher than utilities for the same health states obtained from healthy individuals [10]. There are practical impediments to collection of utilities directly from patients, including the complexity of the concepts involved and the requirement for an interviewer-administered questionnaire (often unfeasible in the cooperative group setting). In addition, utility assessments provide little information on important disease and treatment-specific problems and are probably less sensitive to changes in health status over time than psychometric data [12, 56]. Finally, the few studies that have been done involving simultaneous measurement of utilities and health status have found them at best to be moderately correlated, with measures of mood and depression correlating more highly than other measures with utilities [57].

A modified utility approach has been developed to evaluate the effectiveness of adjuvant chemotherapy for early-stage breast cancer [27]. This approach, the Quality-Adjusted Time Without Symptoms and Toxicity (Q-TWiST), discounts survival time spent with toxicity or symptoms relative to disease-free survival off therapy. Thresholds for decision-making were determined by modeling actual survival data, and judgments were made by the investigators regarding where patient preferences were likely to fall relative to these threshold values. There is no theoretical reason why actual patient preference data could not be used in the Q-TWiST analyses or other studies of quality-adjusted survival. If the relationship between psychometric data and utilities can be established, it will become possible to collect psychometric data and base utility estimates on the reports of patients rather than the best guesses of others.

In summary, the existing science of quality-of-life measurement is organized around a presumed (but theoretically unsubstantiated) dichotomy between psychometric and utility approaches. Neither approach alone is sufficient to understand clinical trial outcome data. The psychometric approach provides a detailed perspective of the patient, but it does not generally tell us how important a given problem or set of problems is to a group of patients. The utility approach informs us about the relative value of various health states; however, because of its emphasis on a single summary score, it fails to reflect the specific problems that might emerge. To date, it has also usually relied on surrogates rather than on patients to provide the utility weights. An individual provider cannot be expected to work intelligently with either one alone. The psychometric approach can uncover specific areas of difficulty or dysfunction, yet patients may not consider these areas to necessitate a change in treatment. On the other hand, the utility approach does not generally reveal the nature of specific problems or dysfunctions, which clearly hampers the provider's efforts in planning interventions or treatment changes. In fact, identification of health dimensions uniquely important to an individual and quantifying patient status within those dimensions has been proposed [35]. These approaches can and must be integrated for advances in the field to continue. Previous efforts to combine psychometric and utility approaches have been rare and, where present, poorly integrated [23]. An integrative approach could be applied in which a well-validated quality-of-life scale could be administered to a patient in a clinical trial (or in clinical practice, for that matter). This patient's total score could be converted to a standardized score that allows for both ease of communication and possible utility analysis.

The investigators task is to select the measure most likely to be effective for a given purpose. This is best accomplished by careful consideration of the purpose of the investigation, critical evaluation of the psychometric properties and known performance of available measures, and review of item content for relevance and appropriateness. Careful checking for relevance can prevent selection of an otherwise valid measure which will be insensitive for the application selected. For example, the short forms derived from the Medical Outcomes Study [30, 60] have a long history of development and demonstrate good psychometric properties, but may be inappropriate at the high end of HQL (e.g., adjuvant chemotherapy) because they emphasize mobility and physical function over social well being, sexuality and body image. The issue of disease severity cuts across virtually all self-report measures of HQL, in that it becomes difficult if not impossible to obtain self-report HQL data from very weak, cognitively impaired or emotionally upset patients. This is an unfortunate irony given that these patients are often the very ones where quality-of-life concerns take first priority in treatment decision-making. Efforts to use surrogate ratings have been largely disappointing, showing that health providers and, to a lesser extent, family members cannot be considered as reliable surrogate raters [48].

Related to validity is the issue of meaningfulness of the data obtained. A comparison of treatment arms might indeed result in differences in HQL, but how much of a difference is clinically meaningful, as opposed to statistically significant? For seven-point Likert scaling of symptoms, Jaeschke et al. [33] have suggested a difference of approximately 0.5 unit as a minimal clinically important difference. For other types of scaling (e.g., linear analogue), Jacobson and Truax [32] recommend a Reliable Change Index that estimates whether a change measured is real or a consequence of imprecise measurement.

Quality-of-life measures used in oncology

This section briefly summarizes some of the more commonly used and adequately validated measures of HQL that have been designated cancer-specific. The designation of cancer-specific is rather arbitrary in that some measures considered to be cancer-specific could be (and have been) applied in other diseases. Examination of item content of some of these measures reveals that indeed many of the concepts measured are generic rather than cancer-specific.

Psychometric measures

Spitzer Quality-of-Life Index (Spitzer QLI) [49]

Although not the first cancer-specific quality-of-life measure to appear in the literature (e.g., see [41]), the Spitzer QLI was certainly an early entry. Intended by its authors to be conceptually equivalent to a neonatal Apgar score [5], it was originally developed as a ten-point physician rating of five areas of functioning (activity, daily living, health, support, outlook). Since then many have used this observer rating scale as a patient-rated scale, with reasonable success [48]. The Spitzer QLI was carefully constructed using expert advisory panels comprised of patients and professionals, and has been subjected to study in at least 28 empirical investigations. In their review, Wood-Dauphinee and Williams [61] conclude that it is a well-validated global measure of HQL. Proxy ratings and reliability data for subscales of activity, daily living and health are more robust than those for support and outlook. The Spitzer QLI has demonstrated the ability to distinguish cancer patients with terminal disease from patients either with recent disease or ones who were engaged in active treatment [49, 61]. The Spitzer QLI has also been positively related to the Uniscale and Multiscale Measures of Quality of Life and self- and physician ratings of HQL in cancer patients [61], although the relationship with the Karnofsky Performance Status Scale (KPS) has been variable.

Ferrans and Powers Quality-of-Life Index (QLI) [24]

The QLI is a 68-item index of overall quality of life, which represents the aggregate of four health domains: health and physical functioning, social and economic, psychological/spiritual domain, and family domain. The instrument consists of two parts: the first measures satisfaction with 34 areas and the second measures their perceived importance. Scores are derived by weighting satisfaction scores with their importance [24]. The cancer version was an adaptation of an earlier general population version of the QLI, which was developed on the basis of an extensive review of the oncology literature and tested in breast cancer patients. Internal consistency reliability coefficients for the subscales ranged from 0.65 (family) to 0.93 (psychological/Spiritual), and the total index correlated highly with a measure of life satisfaction [24].

European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire – Core (EORTC-QLQ C30) [2, 3]

This is a 30-item instrument consisting of both dichotomous responses (yes/no) and responses that utilize a four-point rating scale ranging from "not at all" to "very much." The original 36-item QLQ [2] has been replaced with a 30-item version [3], which reduces the number of physical and emotional functioning items and replaces a single concentration and memory item with 2 separate items. The core instrument was developed from a conceptual model and measures physical functioning, role functioning, emotional functioning, and social functioning, along with disease symptoms, financial impact and global quality of life across different European and North American languages and culture. Aaronson et al. [2] report α values for individual scales ranging from a low of 0.59 for a 3-item subset of the physical functioning dimension to a high of 0.85 for the 2-item global quality-of-life dimension. Multitrait scaling techniques using 156 tests of item-discriminant validity yielded only one definite and three probable scaling errors and interscale correlations supported the notion of nonorthogonal dimensions ($P<0.001$) in quality of life. Finally, Aaronson et al. demonstrated that the seven scales significantly predicted differences in patient clinical status [2, 3].

Functional Living Index – Cancer (FLIC) [45]

This is a 22-item scale on which patients indicate the impact of cancer on "day-to-day living issues that represent the global construct of functional quality of life" [45], using a seven-point Likert-type rating. The scale

provides a total HQL score only. Initial psychometric evaluation indicated two factors (physical and emotional) accounting for a large proportion of the variance, and other smaller factors. Convergent validity studies on the FLIC suggest that the emotional factor is more highly correlated with other well-validated measures assessing depression and anxiety than with measures of physical functioning. Conversely, the physical factor of the FLIC is more highly correlated with measures of physical functioning than with measures of emotional distress. Despite these results suggesting at least two distinct factors, there remains only a single total score available for the instrument. The FLIC has been used extensively in oncology with predominantly positive results.

Functional Assessment of Cancer Therapy (FACT) Scales [15, 16]

This is a 34- to 50-item compilation of a generic core (28 items) and multiple specific subscales, which reflect issues or problems associated with different diseases (e.g., breast, bladder, colorectal, head and neck, lung, ovary and prostate cancer, and HIV infection), treatments (e.g., bone marrow transplantation), and symptom complexes (e.g., anorexia, incontinence) [16]. The scale was developed using a modular structure similar to that of the EORTC, but including patient "experts" in addition to multiple specialists to develop items. After developing the items with over 200 patients and 30 specialists, the general 33-item version (FACT-G) was validated on a second sample of 630 patients with a variety of cancers at different stages. The measure yields a total HQL score and subtest scores for physical well-being, social/family well-being, relationship with doctor, emotional well-being, functional well-being, and disease-specific concerns. Six additional experimental items request information regarding how much each dimension affects HQL, using a 0 (not at all) to 10 (very much so), rating scale.

The FACT-G is able to distinguish metastatic from non-metastatic disease. It also distinguishes between stage I, II, III and IV disease, between different levels of performance status, and between inpatients and outpatients from different centers.

A unique feature of the FACT scales is that they provide supplemental valuative ratings that allow patients to provide domain-specific utility weights. These scales were developed primarily out of the psychometric tradition; however there was an early eye toward movement into a utility approach as demonstrated by two unique features. First, the 47 items (38 general, 9 site-specific) that were selected for version 1 of the FACT were drawn from a larger pool of over 200 possible items according to patient ratings of item importance generated from the first-generation questionnaire. Second, each item on version 1 of the instrument required that the patient make two ratings: a rating of actual function or disability, and a rating of expectation that assesses whether a given symptom or rating was better or worse than expected.

Cancer Rehabilitation Evaluation System – Short Form (CARES-SF) [26, 43]

This is a 59-item self-administered rehabilitation and HQL instrument comprising a list of statements reflecting problems encountered by cancer patients. Patients complete a minimum of 38 to a maximum of 57 items, depending on their treatments as well as on other medical and demographic factors. Statements are rated in terms of how applicable it is to them using a five-point rating scale ranging from "not at all" to "very much." The measure yields a global score (summed ratings) reflecting overall HQL, five summary scores reflecting physical, psychosocial, medical interaction, marital and sexual dimensions, and 31 subscales. Adequate test/retest reliability (10 days, $r=0.92$ for global score, and ranges from 0.69–0.87 for subscales), internal consistency (α for five subscales ranges from 0.67–0.83), and concurrent validity with other HQL measures (r values range from -0.50 to 0.74, $P<0.0001$) are reported [43] and the shortened form is correlated with the longer, 139-item version at $r=0.98$. The global CARES score is sensitive to the extent of disease in colorectal, lung and prostate cancer patients, and to improvement in HQL in breast cancer patients over a 13-month period [26]. Summary scales, in part, have replicated global CARES scores, particularly in colorectal and lung disease [26].

Linear-Analogue Self-Assessment (LASA) scales

LASA scales use a 100-mm line with descriptors at each extreme. Respondents are required to mark their current state somewhere along that line, which is then measured as a score in centimeters or millimeters from the "0" point. There are three noteworthy LASA scales for cancer patients. The original LASA scale of Priestman and Baum [41] was a 10-item scale for studying HQL in advanced breast cancer. This was later extended to 25 items in a study comparing chemotherapy and hormone therapy for advanced breast cancer [6]. These items included 10 on symptoms and side-effects, 5 on physical functioning, 5 on mood, and 5 on social relationships.

The other two LASA scales of note are the 31-item measure of Selby et al. [46], which has been recently reduced to 29 items [9]; and the 14-item LASA of Padilla and colleagues [39, 40]. Much of the Selby measure [9, 46] was derived from the 12 sickness impact profile (SIP) categories [8], and supplemented with items

to measure pain, mouth sores, concern with appearance, and other breast-cancer-specific concerns. Test/retest reliability and internal consistency coefficients are above 0.70 [10, 50]. Concurrent validity coefficients with the appropriate SIP scales ranged from 0.28 to 0.98, most being above 0.60. Reliability coefficients on the Padilla et al. scale are acceptably high, with a factor analysis of 130 cancer patients revealing three factors (physical well-being, psychological well-being, symptom control) accounting for 73% of the total variance [40]. They have also developed a longer (23-item) measure for colostomy patients [39].

Linear-analogue scales are appealing because they are easy to administer and are usually presumed to have robust sensitivity due to interval scaling and a wide range of scores. They have also been criticized on the grounds that their sensitivity may be illusory and that it is difficult to know the minimal clinically significant difference. They also cannot be administered over the telephone, which can be limiting. However, they have performed rather well in metastatic breast cancer. For example, women receiving cytotoxic therapy were found to suffer more adverse physical reactions with a subsequent improvement in well-being on Priestman and Baum's scale, as long as there was an objective clinical response [41]. Later that decade, the much-quoted, counterintuitive results of Coates et al. [17] were reported in which women with metastatic cancer did better on continuous chemotherapy than those on intermittent chemotherapy. They used a very simple 5-item linear-analogue scale along with the Spitzer QLI. Finally, Tannock et al. [51] demonstrated trends toward better HQL in women receiving higher dosages of cytotoxic chemotherapy as opposed to lower doses, presumably because of the increased tumor response and survival advantage gained from the increased dosage. They used the Selby LASA. All of these studies point to the same general conclusion about management of metastatic breast cancer: that the advantages of continuous cytotoxic chemotherapy outweigh the costs, assuming sufficient dosing and assuming the presence of measurable response to therapy. Taken together, these findings can provide valuable guidance in patient counseling and management with respect to the costs and benefits of cytotoxic chemotherapy in advanced breast cancer. In fact, Tannock has put forth a set of guidelines for managing metastatic breast cancer based upon available treatment and HQL data [50].

Rand 36-item survey 1.0 (also known as SF-36) [30, 36, 60]

The Rand 36-item survey 1.0 (SF-36) is a self-administered 36-item measure of eight health concepts: physical functioning, limitations in role functioning due to physical health problems, social functioning, bodily pain, general mental health. Limitations in role functioning due to emotional problems, vitality, and general health perceptions [30, 36, 60]. It was developed to reproduce the previously well-validated, full-length scales, but in a shorter format. Responses vary as a function of the attribute measured, and range from dichotomous to a maximum of five possible choices. Its standardized scoring system yields a profile of eight health scores, which are summed scores of individual scale items (some of which have been reverse-scored), as well as summary indices. The SF-36 is reported to have satisfactory reliability (coefficients ranging from 0.73 to 0.94). Validity studies have demonstrated that it can distinguish patients with and without a chronic condition, discriminate levels of severity within a medical diagnosis, and reflect changes in health-related quality of life associated with changes in disease severity [30].

Utility measures

Quality-of-Well-Being Scale (QWB) [4, 34]

The QWB is actually a hybrid health-status/utility measure of HQL. Kaplan and Anderson [4] focus on the qualitative dimension of functioning rather than exclusively on the psychological and social attributes of health outcomes. The scale is a 25-item list of symptom/problem complexes (CPX) covering the domains of mobility, physical activity, and social activity, each representing related but distinct aspects of daily functioning. Community weights for each CPX control for its relative desirability, with higher weights reflecting more desirable states. The QWB is administered in a standardized interview and yields information about both specific states (CPX) and a total quality of well-being score (range = 0–1), expressed as the average of relative desirability scores. It is reported to demonstrate good test/retest reliability (r values ranging from 0.78 to 0.99, with most correlation coefficients being above 0.90) over a 1-day period across different populations and health problems [4], and adequate content, convergent, and discriminant validity [34]. Because it is not a "pure" utility measure, resulting QWB scores have on occasion been counterintuitive and therefore difficult to implement in health-policy decision making.

Quality-adjusted Time Without Symptoms and Toxicity (Q-TWiST) [27]

The only utility approach that was developed to be cancer-specific is the quality-adjusted time without symptoms and toxicity (Q-TWiST) approach, which attempts

to evaluate the effectiveness of adjuvant chemotherapy for early-stage breast cancer [27]. Similar to quality-adjusted life year (QALY) approach, it discounts survival time by reducing it according to a predetermined utility weight (0–1 range), which accounts for the impact of disease symptoms and treatment side-effects. Its developers have not yet generated the utility weights from patients themselves; rather they depend upon an assumed perspective. Given that it infers rather than measures patient preferences, the Q-TWiST approach may be regarded as related to but conceptually distinct from patient-rated HQL. It carries some advantages over other approaches in that it is inexpensive to derive and allows for adjustment of survival time with the (presumed) HQL of that time. It may be possible to integrate the Q-TWiST approach with a psychometric scale or a patient preference scaling approach that increases sensitivity of measurement from the perspective of the patient.

Quality assurance: our biggest problem

We have focused most of our effort in HQL evaluation on refining instrumentation, study design and statistical analysis. These areas are intellectually stimulating and personally rewarding to investigators. Ironically, all of these efforts can be thwarted by oversights in a less stimulating, less prestigious enterprise: quality assurance. Most clinical-trial organizations are not equipped to support the quality-assurance needs of an HQL evaluation in a clinical trial. The most vulnerable groups are those whose inexperience leads them to believe that quality assurance will not be a challenge. Quality-control procedures are likely to be most successful if they closely approximate existing quality-control mechanisms within the trial group. Nevertheless, the need for special added procedures often exists. Quality control in HQL studies is important at all phases of the study, from protocol development, to initiation of the study, and into follow-up of patients over time. Quality control needs differ at different points along the life of the study.

For an HQL effort to succeed, a centralized person or organization must be willing to take active and primary responsibility for the management of the project. Frequent contact, including the provision of ample opportunities for open communication, is an important and effective tool to maintain both the *quality* and the *quantity* of the collected data. An electronic mail user's group, with a specified mailbox name, can be very useful in allowing site investigators to check on a daily basis for new information and HQL trial updates. Also, frequent (e.g., semi-monthly) conference calls with the site interviewers and/or data collectors help to improve data quality by allowing less experienced personnel the opportunity to go over any questions or problems, and to obtain an update on their accrual, comparing it to their target accrual. Recently, the Canadian National Cancer Institute reported impressive quality control of HQL data on three of its trials, with overall compliance ranging from 95% to 99% [42]. These trials included English- and French- (and in one case, Italian-)speaking patients. They describe nine specific measures which contributed to their success:

1. Making quality of life a specific (i.e., mandatory) trial objective

2. Providing a clear rationale for studying HRQL in the protocol document

3. Including HQL administration instructions in the protocol document

4. Modifying data collection forms to remind data managers to gather data

5. Providing specific reporting schedules

6. Establishing successful completion of the HQL form as a *prerequisite* of eligibility with verification of questionnaire completion at the time of randomization

7. Providing computer-based reminders in advance of the due dates for questionnaire completion

8. Providing pretrial workshops for data managers on HQL rationale and administration procedures

9. Providing ongoing feedback to participants via letter and newsletter

All of these procedures can easily be applied to most multicenter trials with minimal effort, as long as they have the support of the leadership of the clinical trial organization and of the study chairs.

Protocol development

There are two issues related to protocol development that surface prior to any HQL study activation. First, the usual review process, in which study investigators and institutional principal investigators and biostatisticians examine the protocol, is inadequate for HQL studies, because data collection requires the learning of unfamiliar techniques by nurses and data managers. Therefore, protocol input from these disciplines, as well as from collaborating social scientists, is necessary in order to clarify any misunderstandings before they complicate the study procedures. It is important to establish that all disciplines are aware of each others' responsibilities within a particular HQL study, and this can be specified in the written protocol.

A second issue related to protocol development is the shortage of specialized expertise in statistical han-

dling of multidimensional, correlated data collected at multiple time points [18]. Statisticians in cooperative groups are typically confronted with unprecedented difficulty managing and analyzing data such as these. Many are not familiar with commonly used analytic options or statistical packages. It is important to clarify analytic plans *prior* to initiating a study.

Timing of measurement

Consideration of timing is deceptively complex. Detailed recommendations can be found elsewhere [58]. In general, it is advisable to keep the number of assessments to an interpretable minimum. When determining the specific assessment times, the investigator must balance treatment toxicities, the natural history of the disease, and time since initiating therapy along with a constant awareness of the study objectives. An additional level of complexity is added when comparing treatments of differing lengths to one another. The investigator is encouraged to consult with other colleagues who have experience with these treatments in order to catch any "blind spots" in planning these times that could render the comparison unfair. Finally, it is important to remind the investigator that patients should continue to be assessed for their HQL even if they discontinue therapy for some reason. A proposal for tracking down and studying these patients if they become lost to the institution should be specified.

Implementing HQL assessment

Although the details of implementation are of equal if not greater importance compared to the choice of instrumentation, the latter issue receives far more attention when planning the typical clinical trial. Unlike the task of instrument choice, which is completed before the trial begins, implementation demands continue throughout the trial and often beyond (e.g., when patients are followed until death). Unsuccessful implementation threatens the conclusional validity of the trial at many levels, including sampling bias (if all patients or a random subset do not participate), generalizability (if all institutions or cultures do not participate), and statistical conclusion validity (if there are missing data or inappropriate analyses planned).

Because of the unique nature of HQL data, staff and patients will require pre-study education and/or training about the nature of the HQL investigation, its purpose and its procedures. This can be a labor-intensive effort which requires central coordination and planning. As the study progresses, busy clinic schedules, normal staff turnover, and lack of accountability can all contribute to a systematic forgetting about the HQL component of the study over time. The result is patient attrition. Even if the protocol is carefully conceived, written and executed at study initiation, there remains a need for continued vigilance toward the risk of a declining rate of participation. Planned "booster" educational sessions and enforced accountability at each data collection site are mechanisms that can be considered to enhance quality control during follow-up.

Training and monitoring interviewers

The conclusions drawn from multicenter HQL studies will have significant implications for the interpretation of medical outcomes and patient preferences. Patients must be helped to feel as comfortable as possible, thereby maximizing the likelihood that they will provide veridical data. It is therefore important that interviewers be perceived as members of a similar culture as far as possible, in order to set patients at ease and facilitate removal of status barriers between examiners and respondents. HQL measures are all fairly easy to administer, provided that a minimum degree of preparatory training and monitoring occurs. Some standardization of administration must be established and monitored during the trial. For inexperienced data collectors, an initial pilot study could offer the opportunity for experience-based training which, when appropriately monitored, will improve a consistent administration technique. The procedure for administration of the HQL battery can also be standardized in a brief training manual or guide. Administration guidelines specific to the instrument to be given and the trial to be conducted should be provided whenever necessary. Standardized aspects of test administration must be consistently addressed at each site, and this is best monitored centrally after initial training.

Access to patients

Gaining access to patients may be a significant issue when assessing HQL. Although there may appear to be adequate numbers of patients or families in a particular setting, some studies languish because of accrual problems on the HQL component. This may be a sign of resistance. When low accrual is due to a poorly motivated staff, efforts to enhance their interest and commitment to participation, perhaps with built-in incentives, are important. For the patient, a HQL evaluation must be placed in a context so it is not perceived as gratuitous. Piloting can determine acceptability to patients and families, and written consent can prepare them for the nature of the inquiry.

Some solutions

Many practical problems emerge in the context of collecting HQL data in a clinical trial or, for that matter, in clinical practice. Many of these problems can be overcome or even circumvented by employing data collection strategies that draw upon recent advances in computer technology. Three such strategies will be described. The first is an augmentation of what is essentially today's standard (i.e., usual) approach. The second and third are more novel advances, including a decentralized direct on-site data-entry system, and a centralized off-site telephone data-collection system.

The augmented standard approach

The usual approach to gathering patient-reported data is to entrust the tracking and quality assurance monitoring to the hands of the existing clinical personnel. The assumption is that the outcomes are of sufficient intrinsic value, or that there are adequate extrinsic incentives, such as per capita reimbursement, to guarantee the collection of high-quality data. This assumption is almost never valid. Busy clinical staff may have the best of intentions but will produce data with multiple missing points, often to the point of uninterpretability, if they are entrusted to obtain high-quality data without help in the form of training, tracking, reminding and continuous quality assurance. All of these needs can be met, assuming adequate investment in the HQL portion of the clinical trial or effort.

Many clinical trials and health-care-delivery systems involve multiple clinical sites, each of which might place only a handful of patients per year on a given study or treatment. Low-volume clinical sites have a relative disadvantage over high-volume sites, because they usually do not have the resources to dedicate a full-time staff person to data management and quality assurance. However, with an organized effort at the local institutional level, high-quality data collection can occur. It is important to remain aware that HQL data differ from other trial data in two fundamental ways. First, they are obtained directly from the patient and therefore necessitate enlisting patient cooperation beyond that required for treatment adherence. Second, they cannot be retrieved from medical records if they are not measured at the specified time. This means that the person responsible on site is vital to the successful completion of all HQL protocols. This person must be motivated and able to stay abreast of upcoming patients due for evaluation. Quality-assurance procedures must be specified in the protocol and carried out on site.

Specific recommendations

It is recommended (a) that each participating institution designate a person who has responsibility for the HQL component of the study, (b) that each institution has a plan for keeping track of when HQL data are due on each individual patient, and (c) that the institution has a plan for promptly contacting patients who miss an assessment appointment. An acceptable window of time should also be specified after which data must be considered irretrievable. Procedures for this retrieval, including acceptable methods of data collection (proxy informant, mail, telephone, etc.), should be specified before beginning the trial.

When patients begin to complete the HQL form, they should be reminded of the time frame specified on the questionnaire (e.g., "past week"). If the patient requires assistance completing forms, this can be provided by a member of the treatment staff who has been trained to provide assistance without introducing bias, but not by family or friends of the patient. After the patient completes the HRQL form, it should be checked for completeness and accuracy. If items are left unanswered or if the responses are made incorrectly (e.g., circling a descriptive word when in fact a number was to be circled), they should be presented back to the patient with a request for clarification. If the patient does not want to answer, an explanation to this effect should be written in the margin and submitted to the data management center for the study.

Institutional tracking and quality assurance

A suggested local institutional method for tracking of patients is to assign two "cards" to each patient. These "cards" may take the form of two different sorts in a spreadsheet computer program, or they can be actual index cards in a filing box. Of course, if hardware and programming resources permit, it is more efficient and accurate to use "custom" spreadsheet or data base management programs to sort individual protocol data. One card (or record in a data base program) is sorted according to the date 2 weeks before the participant is due to complete the next HQL evaluation. This record should also contain the patient's name and phone number, the treating physician's name and phone number, and the study identification number. The other record is sorted alphabetically, and contains the location (i.e., date in file) where the other record can be found. This cross-referencing enables one to stay abreast of who should be receiving HQL evaluations in a given week and when any given patient is due for HQL evaluation. This ensures against loss of contact and greatly improves the likelihood that a patient will arrive within the window of time required by the protocol.

Before the patient's next visit, a parameter sheet describing which tests are required for the visit should be checked. If an HQL evaluation is among them, the patient can be called and prepared for this approximately 1 week prior to the appointment date.

Off-site telephone data collection (centralized)

Data collected off site (i.e., by telephone) can be put directly into the centralized data-management center. Patients can be called according to previously arranged appointments, in their homes at their convenience. Patients who do not have a telephone (or who wish not to be called at home) can be interviewed by telephone during their clinic visit. In the case of a clinical trial, informed consent will have been completed by patients at entry to the study. In the case of general tracking for clinical progress, patients can be advised that they will be called at home periodically to see how they are doing. Typically, a very positive relationship is established between the patient and the telephone interviewer, often borrowing from the positive relationship between the patient and the provider represented by the interviewer. The interviewer is advised to mail a copy of the forms ahead to the patients in advance of the scheduled interview, particularly if there are multiple forms to be completed. This gives the patient the opportunity to complete the forms in advance and simply read the answers over the telephone. Many patients will choose this option, as it reduces time spent on the telephone and adds to their convenience and sense of control over the parameters of their participation in the project.

Advantages of the off-site telephone approach include the fact that it centralizes the quality-control effort, and that it removes the data collection effort from the busy clinic setting (Table 1). It is an especially good choice when there are multiple clinical sites, or when there is a small number of patients per clinical site, because there are no on-site start-up costs for data collection. Such costs would be difficult to justify unless the patient-to-site ratio were relatively high (Table 1).

Table 1 Advantages and disadvantages of data collection methods

Method	Advantages	Disadvantages
Augmented standard	Minimal start-up costs	Highly vulnerable to missing evaluations *and* responses; staff turnover/training needs; effort; patient inconvenience
Off-site telephone (centralized)	High-quality control: less vulnerable to missing evaluations *and* responses; removes collection from busy clinic; best with multiple sites; best when patient:site ratio is low; patient convenience	Highest cost
Direct on-site entry (decentralized)	Moderate quality control: less vulnerable to missing responses; no forms required; patients enter data directly; best with fewer sites; best when patient:site ratio is high	Vulnerable to missing evaluations; patient must be on-site

Direct on-site data entry (decentralized)

A second new approach to obtaining patient-reported data in clinical trials and other medical treatment organizations is the direct on-site data entry approach. In this approach, the data are collected from the patients at the treatment site; however, not in the usual paper-and-pencil format. Instead, patients enter their responses to questions presented sequentially by an interactive computer program. This program can be custom-designed for virtually any combination of questions and time frames of assessment. This approach works best when the choices to be made by the patient are relatively simple (e.g., true/false; multiple choice). One limiting factor in this approach is that patients must be present at the clinic in order to provide their responses (Table 1). This limitation can be overcome with the use of touch tone telephone responses or individual interviews as a back-up method for those patients who cannot come in to the clinic.

To some extent, the advantages and disadvantages of the direct on-site method contrast with those of the off-site approach. Quality-control efforts are compromised by the fact that patients must come to clinic in order to be assessed, requiring a back-up strategy. Quality control of actual patient response to a set of questions is quite high, however, because the computer can be programmed to proceed only after a question is answered. A tremendous advantage of this approach is the removal of forms from the process of data collection, transmittal and entry. Patient responses are thereby transmitted error-free from the clinic to the data analysis center. This method is ideal for trials where there are relatively few sites, especially when the patient: site ratio is high. The cost of a computer at each

clinical site is offset by extensive cost savings entailed in form transcription and keypunch data entry.

Conclusion

This paper provides a brief update of some HQL study approaches and instruments commonly used in oncology. Some studies have already contributed to an understanding of the diverse costs and benefits of cancer therapies. Most of the progress has been in breast cancer and, to a lesser extent, lung, colorectal and prostate cancer. Further attention must be directed to less common (e.g., hematological) malignancies as well as intensive experimental therapies with severe toxicity and uncertain benefit (e.g., bone marrow transplantation with solid tumors). The "cancer-specific" issues in these areas may be sufficiently distinct to require new or appropriately adapted measurement. In order to make valid use of HQL data, there is a need for high-quality data collection. Most clinical trials groups who have never included HQL assessment underestimate the resources and commitment required for success. Fortunately, there are coming available some novel approaches to patient-reported data collection that circumvent or overcome many of the usual barriers to the collection of high-quality data. These include but are not limited to centralized telephone tracking and interviewing, and direct on-site data entry via interactive computer.

References

1. Aaronson NK, Beckman J (1987) The quality of life of cancer patients. (Monograph series of the European Organization for Research on Treatment of Cancer (EORTC), vol 17) Raven Press, New York
2. Aaronson NK, Ahmedzai S, Bullinger M, et al (1991) The EORTC Core Quality-of-Life Questionnaire: interim results of an international field study. In: Osoba D (ed) Effect of cancer on quality of life. CRC Press, Boca Raton, Fla, pp 185–203
3. Aaronson NK, Ahmedzai S, Bergman B, et al (1993) The European Organization for Research and Treatment of Cancer QLQ-C30: a quality of life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85:365–376
4. Anderson JP, Kaplan RM, Berry CC, et al (1989) Interrater reliability of function assessment for a health status measure: the quality of well-being scale. Med Care 27:1076–1084
5. Apgar V (1953) Proposal for new methods of evaluation of newborn infants. Anesth Analg 32:260
6. Baum M, Priestman T, West RR, Jones EM (1980) A comparison of subjective responses in a trial comparing endocrine with cytotoxic treatment in advanced carcinoma of the breast. Eur J Cancer 16 [Suppl 1]:223–226
7. Barofsky I, Sugarbaker PH (1990) Cancer. In: Spilker B (ed) Quality of life assessments in clinical trials. Raven Press, New York, pp 419–439
8. Bergner J, Bobbitt RA, Carter WB, Gilson BS (1981) The Sickness Impact Profile: development and final revision of a health status measure. Med Care 19:787–806
9. Boyd NF, Selby PJ, Sutherland HJ, Hogg S (1988) Measurement of the clinical status of patients with breast cancer: evidence for the validity of self assessment with linear analogue scales. J Clin Epidemiol 41:243–250
10. Boyd NF, Sutherland HJ, Heasman KZ, et al (1990) Whose utilities for decision analysis? Med Decis Making 10:58–67
11. Calman KC (1984) The quality of life in cancer patients – an hypothesis. J Med Ethics 10:124–125
12. Canadian Erythropoietin Study Group (1990) Association between recombinant human erythropoietin and quality of life and exercise capacity of patients receiving hemodialysis. Br Med J 300:573–578
13. Cella DF, Cherin EA (1988) Quality of life during and after cancer treatment. Comprehensive Therapy 14:69–75
14. Cella DF, Tulsky DS (1990) Measuring quality of life today: methodological aspects. Oncology 4:29–38
15. Cella DF, Lee-Riordan D, et al (1989) Quality of life in advanced cancer: three new disease-specific measures. Proc ASCO 8:315
16. Cella DF, Tulsky DS, Gray G, et al (1993) The Functional Assessment of Cancer Therapy (FACT) scale: development and validation of the general version. J Clin Oncol 11:570–579
17. Coates A, Gebski V, Stat M, for the Australian-New Zealand Breast Cancer Trials Group (1987) Improving the quality of life during chemotherapy for advanced breast cancer. N Engl J Med 317:1490
18. Cox DR, Fitzpatrick R, Fletcher AE, et al (1992) Quality of life assessments: can we keep it simple? J R Statist Soc 155:353–393
19. DeHaes JCJM, Knippenberg FCE van, Neijt JP (1990) Measuring psychological and physical distress in cancer patients: structure and application of the Rotterdam Symptom Checklist. Br J Cancer 62:1034–1038
20. Donovan K, Sanson-Fisher RW, Redman S (1989) Measuring quality of life in cancer patients. J Clin Oncol 7:959–968
21. Drummond MF, Stoddart GL, Torrance GW (1987) Methods for economic evaluation of health care programmes. Oxford University Press, Oxford
22. Fayers PM, Jones DR (1983) Measuring and analysing quality of life in cancer clinical trials: a review. Statist Med 2:429–446
23. Feeny D, LaBelle R, Torrance GW (1990) Integrating economic evaluations and assessments. In: Spilker B (ed) Quality of life assessments in clinical trials. Raven Press, New York, pp 71–83
24. Ferrans CE (1990) Development of a quality of life index for patients with cancer. Oncol Nurs Forum 17 [Suppl]:15–21
25. Fetting JH (1988) Evaluating quality and quantity of life in breast cancer adjuvant trials. J Clin Oncol 6:1795–1797
26. Ganz PA, Schag CAC, Lee JJ, Sim MS (1992) The CARES: a generic measure of health-related quality of life for patients with cancer. Quality Life Res 1:19–29
27. Gelber RD, Goldhirsch A, Cavalli F (1991) Quality-of-life adjusted evaluation of adjuvant therapies for operable breast cancer. Ann Intern Med 114:621–628

28. George L, Bearon L (1980) Quality of life in older persons: meaning and measurement. Human Sciences Press, New York
29. Gotay CC, Korn EL, McCabe MS, et al (1992) Quality-of-life assessment in cancer treatment protocols: research issues in protocol development. J Natl Cancer Inst 84:575–579
30. Hays RD, Sherbourne CD, Mazel RM (1993) The RAND 36-item health survey 1.0. Health Econ 2:217–227
31. Hays RD, Stewart AL (1990) The structure of self-reported health in chronic disease patients. Psychol Assess J Consult Clin Psychol 2:22–30
32. Jacobson NS, Truax P (1991) Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 59:12–19
33. Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status: ascertaining the minimal clinically important difference. Control Clin Trials 10:407–415
34. Kaplan RM, Bush JW, Berry CC (1976) Health status: types of validity and the Index of Well-Being. Health Serv Res 11:478–507
35. Llewellyn-Thomas HA, Sutherland HJ, Tritchler DL, et al (1991) Benign and malignant breast disease: the relationship between women's health status and health values. Med Decis Making 11:180–188
36. McHorney CA, Ware JE, Rogers W, et al (1992) The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the medical outcomes study. Med Care 30:MS253–MS265
37. Moinpour CM, Feigl P, Metch B, et al (1989) Quality of life end points in cancer clinical trials: review and recommendations. J Natl Cancer Inst 81:485–495
38. Osoba D (ed) (1991) Effect of cancer on quality of life. CRC Press, Boca Raton, Fla
39. Padilla GV, Grant MM (1985) Quality of life as a cancer nursing outcome variable. Am Nurs Sci 8:45–60
40. Padilla GV, Presant C, Grant MM, et al (1983) Quality of life index for patients with cancer. Res Nurs Health 6:117–126
41. Priestman TJ, Baum M (1976) Evaluation of quality of life in patients receiving treatment for advanced breast cancer. Lancet I:899–901
42. Sadura A, Pater J, Osoba D, et al (1992) Quality of life assessment: patient compliance with questionnaire completion. J Natl Cancer Inst 84:1023–1026
43. Schag CAC, Ganz P, Heinrich RL (1991) Cancer rehabilitation evaluation system – short form: a cancer specific rehabilitation and quality of life instrument. Cancer 68:1406–1413
44. Schipper H, Levitt M (1985) Measuring quality of life: risks and benefits. Cancer Treat Rep 69:1115–1123
45. Schipper H, Clinch J, McMurray A, Levitt M (1984) Measuring the quality of life of cancer patients: the Functional Living Index – Cancer: development and validation. J Clin Oncol 2:472–483
46. Selby PJ, Chapman JAW, Etazadi-Amoli J, et al (1984) The development of a method for assessing the quality of life of cancer patients. Br J Cancer 50:13–22
47. Skeel RT (1989) Quality of life assessment in cancer clinical trials – it's time to catch up. J Natl Cancer Inst 81:472–473
48. Slevin ML, Plant H, Lynch D, et al (1988) Who should measure quality of life, the doctor or the patient? Br J Cancer 57:109
49. Spitzer WO, Dobson AJ, Hall J, et al (1981) Measuring the quality of life of cancer patients: a concise QL-Index for use by physicians. J Chronic Dis 34:585–597
50. Tannock IF (1991) Management of breast and prostate cancer: how does quality of life enter the equation? In: Tchekmedyian NS, Cella DF (eds) Quality of life in oncology practice and research. Dominus, Williston Park, NY, pp 81–88
51. Tannock IF, Boyd NF, DeBoer G, et al (1988) A randomized trial of two dose levels of cyclophosphamide, methotrexate and fluorouracil chemotherapy for patients with metastatic breast cancer. J Clin Oncol 6:1377
52. Tchekmedyian NS, Cella DF (eds) (1991) Quality of life in current oncology practice and research. Dominus, Williston Park, NY
53. Till JE, McNeil BJ, Bush RS (1984) Measurement of multiple components of quality of life. Cancer Treat Symp 1:177–181
54. Till JE, Sutherland HJ, Meslin EM (1992) Is there a role for preference assessments in research on quality of life in oncology. Qual Life Res 1:31–40
55. Torrance GW (1986) Measurement of health state utilities for economic appraisal: a review article. J Health Econ 5:1–30
56. Tsevat J, Goldman L, Soukup JR, Lee TH (1990) Stability of utilities in survivors of myocardial infarction. Med Decis Making 10:323
57. Tsevat J, Cook EF, Soukop JR, et al (1991) Utilities of the seriously ill (abstract). Clin Res 39:589A
58. US Department of Health and Human Services (DHHS), Public Health Service, National Institutes of Health (1990) Quality of life assessment in cancer clinical trials: report of the Workshop on Quality of Life Research in Cancer Clinical Trials. Prevention and Control, National Cancer Institute, Division of Cancer Prevention and Control, Bethesda, MD
59. Ware J (1991) Measuring functioning, well-being and other generic health concepts. In: Osoba D (ed) Effect of cancer on quality of life. CRC, Boca Raton, Fla, pp 7–23
60. Ware JE, Sherbourne CD (1992) The MOS 36-Item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30:473–483
61. Wood-Dauphinee S, Williams JI (1991) The Spitzer Quality-of-Life Index: its performance as a measure. In: Osoba D (ed) Effect of cancer on quality of life. CRC Press, Boca Raton, Fla, pp 169–184