QUALITY OF LIFE ASSESSMENT

An Abstract from the Masters in Science thesis titled: DEVELOPMENT OF THE PROSTATE CANCER RADIATION TOXICITY QUESTIONNAIRE : George B. Rodrigues, B.Sc, MD, FRCPC

Development of HRQoL Instruments: General Methodology

1 Introduction

The definition of quality of life is complex and often depends on the investigator or organization (Dijkers 1999). The World Health Organization defines health as “a state of complete physical/mental, social well-being and not merely the absence of disease or infirmity” (World Health Organization 1947). Conceptually, quality of life can be divided into health-related (with physical, psychological, social, and spiritual subdomains) and non-health related (with personal, social interaction, societal, and environmental subdomains). Health-related quality of life (HRQoL) is increasingly utilized for the measurement of treatment effects in oncology clinical trials, to measure the burden/impact of disease in individuals and groups, and for use by individual patients, investigators, clinicians, and policy-makers to assist in decision making (Guyatt and Veldhuyzen 1989).

HRQoL measures are often considered to be subjective because of their perceptual nature and the fact that the patient reports HRQoL either directly or indirectly (Aaronson 1989, Aaronson 1991). This is in contrast to objective measures that are externally observable and measurable along dimensions that

have either a physical or mathematical basis. Although HRQoL is subjective, there are techniques that make it possible to develop measures of HRQoL that make them valuable additions to “more objective“ measures of therapeutic effects.

One of the challenges in the design of commonly accepted HRQoL endpoints is the definition of the ”clinically important difference”. In general it is defined as the change in HRQoL (or one of its subdomains) that clinicians (and/or patients) consider to be sufficiently relevant to consider a change in therapy or to consider one therapy to be superior to others. Controversy exists in regards to the methodology of determination of clinically important differences (Redelmeier 1996).

Instruments are created to measure various levels of HRQoL ranging from an overall assessment of well-being, to broad domains of HRQoL (e.g. physical), to components of the domain (e.g. radiation toxicity). Types of HRQoL instruments include those that measure: 1. general HRQoL applicable to all individuals; 2. domain(s) of HRQoL applicable to all individuals; 3. domain(s) of HRQoL applicable to a subset of individuals with a certain characteristic, disease, or impairment (disease- or symptom-specific HRQoL instrument); 4. clinical measures (pain scales, depression scales); and 5. issues tangential to HRQoL (personality tests). Properties of methodologically sound HRQoL instruments include multidimensionality, and acceptable psychometric properties such as reliability, validity, and responsiveness (Bergner 1987, Guyatt and Feeny 1993, Hays 1993).

HRQoL instruments can serve one or more of three potential functions. These are instruments that are discriminative, predictive, and evaluative in nature (Kirshner 1985). A discriminative index attempts to distinguish individuals or groups by some underlying HRQoL dimension where no gold standard is available. This type of index could be utilized to assess associations between

HRQoL and other symptom/toxicity scales. A predictive index attempts to distinguish individuals/groups into categories when a gold standard is available. This is most commonly applied in HRQoL studies looking at a short versus long (the gold standard) instrument. Good predictive measures should correlate highly with the appropriate gold standard and are useful when the measure is less risky or costly than the gold standard or can screen for cases of disease/impairment earlier in the disease process. Evaluative instruments measure change in individuals/groups over time. Evaluative instruments are generally used in clinical trials in order to assess differences between groups of individuals. All instruments require assessment of their psychometric properties such as internal consistency, reliability, and validity. In addition, instrument interpretability in terms of external calibration can increase the utility of HRQoL instruments.

Cancer treatments often involve a trade off between response and survival versus acute and late toxicity. Inherent in all of these variables are HRQoL changes. HRQoL can be affected by both the success/failure of therapy and its side effects. This realization of the importance of both sides of the therapeutic ratio has led to increases in the use of HRQoL instruments in oncology. Well-developed core general cancer questionnaires such as the Quality of Life Questionnaire – C30 (QLQ-C30©) and the FACT-G© exist with a myriad of specific treatment and disease modules (Sprangers 1993, Sprangers 1998). In addition, many stand-alone HRQoL oncology instruments also exist spanning the variety of disease sites.

2 The Main Steps in Developing a HRQoL Instrument

2.1 Instrument Construction

2.1.1 Determination of HRQoL Instrument Purpose

Health-related quality of life (HRQoL) indices can be used to measure general domains such as social, emotional, physical, and role functioning as well as more disease- or symptom- specific questionnaires (e.g. prostate cancer and fatigue, respectively). These HRQoL measures can be constructed in order to serve discriminative, predictive, and/or evaluative purposes (see section 3.1). Many HRQoL instruments can either serve one or several of these construction goals. The general methodological process of creating a HRQoL measure involves the development of a questionnaire (specifying measurement goals, item generation/formatting and item reduction) and then subsequent questionnaire testing to establish psychometric properties (pretesting, reliability, validity, responsiveness, and interpretability).

In the initial development of a HRQoL measure, the ultimate purpose(s) of the instrument in terms of discriminative, predictive, and evaluative properties need to be considered. In addition, the domain(s) of interest need to be prospectively and clearly defined. A clear definition of the domain and disease/impairment among the population of interest assists in designing appropriate item construction and validation protocols. Prospective definition of the number and types (general or disease/symptom specific) of domains is essential. A balance between breath and depth of the questionnaire needs to be struck where the usual tradeoff is between loss of detail at a gain of broader scope or vice versa. In addition, disease processes may have several distinct HRQoL effects. A decision of whether all, some or one symptom/impairment is to be included in a HRQoL instrument needs to be made. The patient population of interest needs to be clearly defined in terms of inclusion and exclusion criteria, age, literacy level, language ability, and the presence of other diseases/impairments that may affect HRQOL. The development of HRQoL instruments in conjunction with a narrowly defined study population may limit subsequent use in other populations; therefore, the initial definition of the population of interest is important.

2.1.2 HRQoL Instrument Logistics

Logistical issues also need to be addressed in terms of the length (maximum number of items) and administration (self-administered vs. assisted vs. interviewer and document vs. telephone vs. diary) of the questionnaire. Self-administered questionnaires have the advantage of obtaining the direct input of from the patients' perspectives. For diseases that impair the patient's ability to communicate, family feedback can sometimes serve as a proxy. Interviewer (or health care provider) ratings have the advantages of being practical and rapid. However, these ratings can suffer from low inter-physician and inter-physician/patient reliability. Most cancer toxicity scales are interviewer-based scales.

Document-based (i.e. where questionnaire items and responses are delivered and collected in a paper format) questionnaires have the advantage of a practical and efficient questionnaire delivery that can be easily repeated. However, document questionnaires are less flexible than interview-based questionnaire where an interviewer asks questions from a semi-structured prearranged script. Diaries can also provide useful frequent information from patients; however, compliance and retrospective filling out of questionnaire items may degrade its utility.

2.1.3 Item Selection

The first step in the construction of the questionnaire is item generation through the creation of a large pool of potentially relevant items. Items can be generated from the following sources:

1. Investigators' personal judgement.
2. Discussion with clinical colleagues and expert opinion.
3. Review of the relevant medical literature.

4. Review of analogous HRQoL instruments.
5. Unstructured interview with patients.
6. Patient focus group discussions.

The type of items included in the questionnaire will depend on the purpose of the questionnaire (discriminative, predictive, and evaluative). Item selection for discriminative questionnaires should be focused on clinically important items that universally apply to patients within the study population. Responses to items chosen for this type of questionnaire should also be stable over short periods of time. Predictive questionnaires should select all clinically important items that have a potentially strong statistical association with the gold standard measure. In evaluative instruments, items that are likely to reflect domain changes over time should be selected. A broader selection of items is useful in the development of an evaluative instrument compared to discriminative and predictive instruments because the goal is to measure all clinically important treatment effects. Another issue in questionnaire construction is the time over which the item applies (i.e. a reference to the time period in which you wish the patient to answer a specific question) and can range from immediate to as long as a month. The time specified for each questionnaire depends on the purpose and content of the instrument.

2.1.4 Item Scaling

Item scaling refers to the response options for each potential question in a HRQoL instrument. The simplest scale is a binary scale with two options (e.g. yes/no). This type of question scale is good for discriminative and some predictive instruments. For predictive instruments, the optimal scale will be the one that maximizes the correlation of the instrument scores with the gold standard. The optimal scale varies between individual items and between different instruments. For evaluative instruments, the ideal response scale is one that allows the individual to provide information on clinically relevant changes. In

general, binary scales are insufficient for this purpose because they are insensitive to gradations in response. Five to 9-item Likert scales and visual analogue scales are usually employed for this purpose.

A Likert scale is defined as a categorical, qualitative, closed-end scale ordered in a hierarchical sequence (Figure 1). A visual analogue scale uses a line anchored at each end by the dimension extremes in which individuals are asked to place a mark on the line (usually 10 cm) to indicate their level of symptom/impairment (Figure 2). The score for each item is based on the distance from the left-sided anchor line. Investigations into the relative advantages and disadvantages of Likert versus visual analogue scales have been performed in multiple settings. However, heterogeneity of criteria for judging superiority of one scale over another exists. These include the magnitude of between-subject variability, test-retest reliability, strength of correlation with other validated measures, performance on factor-analysis, and measurement of responsiveness. Thus, no clear consensus has emerged on the value of one scale methodology over any other (Guyatt and Townsend 1987).

Once items have been identified and scale methodology selected, the initial questionnaire pool can be generated. This is followed by item reduction in order to select only items that are relevant to achieve the instrument's purpose.

Figure 1: Prototypical Likert scale

In the past 4 weeks, how often have you had blood in your bowel movements?

Never
Sometimes
Frequently
Most of the time

All of the time

Figure 2: Prototypical Visual Analogue Scale

In the past 4 weeks, how often have you had blood in your bowel movements?

3.2.2 Item Reduction

The process of taking all items generated in an initial item pool and condensing them into a core set of final items is called item reduction. The process of item reduction can depend on the purpose of the instrument as well as prospectively determining criteria for question deletion.

For discriminative questionnaires, questions that are completely identical in terms of their discriminative properties are of limited or no use. That is, if all patients respond identically to a given question, that question should be deleted. Questions need to measure similar concepts in order to be useful. Internal consistency refers to the ability of each item of a question set to measure a given construct. Internal consistency is directly proportional to the number of items in the questionnaire or domain of interest and the correlation between items in measuring the domain of interest. Internal consistency can be measured by the Cronbach alpha statistic. The alpha statistic ranges from –1 to +1 with a higher value seen as a superior value. Deleting items that correlate poorly with other items can increase the alpha statistic (internal consistency). Cronbach alpha analysis with deleted variables can assess whether removing items from the questionnaire domain can improve internal consistency. Alpha values of 0.60 or

higher are usually considered to be acceptable internal consistency; however, acceptable values for alpha can vary between disciplines. Levels above 0.70 are usually considered being good with levels above 0.80 considered to be excellent (Bergner 1987, Hays 1993). Domains with poor internal consistency may not contain all items relevant to the domain construct.

For predictive questionnaires, the criteria for keeping and rejecting a question is to simply keep questions that predict the gold standard and delete questions that do not. Evaluative questionnaires are primarily interested in changes over time. Thus, questions that are unable to detect real changes over time (non-responsive) as a result of a treatment or the natural history of the disease should be deleted. Adding non-responsive items only increases "response burden" for the patient and may reduce the response rate.

Several other item reduction methods exist in the literature. One approach uses patient assessment of frequency and importance of the potential items. The questionnaire item impact factor is determined by the product of frequency and importance as rated by patients. Questions with low impact factor are deleted from further analysis. Another approach uses the deletion of non-discriminative questions with limited heterogeneity. If all individuals answer questions identically (or near identically), the discriminative utility of such questions are limited and should be considered for removal. In addition, questions that highly correlate with each other can be reduced because they are redundant. This is usually advisable in discriminative questionnaires but note that correlation between items can change over time in evaluative instruments. Factor analysis uses mathematical modeling to determine which items should be included. Items that are correlated with each other are grouped together as a "factor". Items not correlated with any other items are eliminated. Common sense domain labels based on clinical experience are then assigned to the various groups and scales are subsequently constructed.

3.2.3 Reliability Testing

The reliability of an instrument relates to the stability of domain scores when no change in the underlying attribute occurs over time. Highly reliable instruments should give identical domain values over time under these circumstances. Reliability is an important feature of discriminative instruments as consistency of response between individuals is an important feature of such a questionnaire. Predictive instruments require even greater reliability as both systematic and random errors must be taken into account when reliability coefficients are calculated. Evaluative instruments require stable intra-subject variation and stable domain scores over periods of insignificant clinical change. Conversely, if significant change occurs over a period of time, that change should be reflected in the questionnaire domain scores (see below).

Often, reliability is measured by a statistical comparison of an administration of a questionnaire (test questionnaire) followed by a second questionnaire administration (retest questionnaire) ideally under identical circumstances. Reliability of a questionnaire, as measured by the test-retest method, can depend on the conditions of questionnaire administration, daily HRQoL changes, length of time between administrations, changes in attribute over that period of time, internal questionnaire consistency and validity. If substantial change occurs between administrations of questionnaires, corresponding reliability coefficients may diminish in magnitude.

Statistically, reliability of a questionnaire can be estimated by assessing the correlation of the test-retest relationship. Individuals are asked to complete a questionnaire on two separate occasions over a period of time that is felt to reflect a stable HRQoL state. Domain scores are calculated and the appropriate correlation coefficient is calculated. A Pearson correlation coefficient could be calculated for this purpose; however, this statistic does not adjust for systemic differences in mean score between administrations. The intraclass correlation

coefficient method uses a one-way Analysis of Variance (ANOVA) method to adjust the underlying correlation for systematic changes in scores (Weir 2005). The one-way intraclass correlation coefficient assumes that raters for each domain are selected at random. The two-way intraclass correlation coefficient (fixed model and random model) allows for further refinement of the reliability estimate assuming different respondent assumptions. The two-way random effects model assumes that the same raters assess the domains of interest and that the raters were randomly chosen from a greater population of potential raters. The two-way fixed effect model assumes that the same raters assess the domains of interest and that the raters consist of the entire list of individuals that can rate the domains. In addition, intraclass correlation coefficients can be calculated for both single and mean scores for each rater. Since, the assessment of reliability can vary on the model chosen for assessment, the selection of a model for the calculation of intraclass correlation coefficients should be based on the population chosen and the type of scores generated. In general, for a given assessment model a reliability coefficient of 0.70 or greater using these techniques considered to represent a questionnaire with acceptable reliability for HRQoL assessment.

3.2.4 Validity Testing

Validation of a HRQoL instrument and its subscales assesses the extent that the instrument measures what it intends and purports to measure (Bergner 1987). This is vital, especially because a gold standard for HRQoL is usually not available. Validation is usually not an “all or none” procedure. It involves collecting increasing levels of evidence that a HRQoL instrument is performing adequately for the purpose it was designed for. Various types of validity exist and are described below with examples (Table 3).

1. Content validity:

Content validity is defined as the extent that items are representative of the HRQoL domain being measured. Content validity can consist of face validity (subjective assessment of validity by expert opinion) and sampling validity (the degree by which the instrument is felt to be comprehensive).

2. Criterion validity:

Criterion validity is defined as relationship between a measure and its corresponding gold standard (if available). Concurrent validity is a sub-type of criterion validity assessing criterion correlation at a point in time. This is usually used for comparisons between short and long questionnaires or between a new HRQoL instrument and an existing physical parameter. Low concurrent validity can occur because the new or old instrument is inferior or because the instruments measure different aspects of the underlying domain. Predictive validity is an assessment of correlation between a measure and its gold standard where the event is at some time in the future.

3. Construct validity:

Construct validity is an assessment of what the instrument is actually measuring. This is accomplished by developing *a priori* hypotheses about how an instrument is to behave in various situations and study populations. Testing of these hypotheses will either lead to improvements in the instruments or increasing confirmation of the instruments' validity.

4. Responsiveness:

The ability for a measure to reflect underlying change in an underlying HRQoL domain demonstrates the instruments' validity. This generally has been the least studied aspect of HRQoL instrument validity.

5. External validity:

Assessment of new findings from validity studies need to be interpreted in the light of previous validation studies in different individuals, settings, times, investigators, and study populations.

Statistically, most validation procedures involve calculation of a Pearson correlation coefficient (also known as a product-moment correlation coefficient). Its purpose is to estimate the level of association of one measure with another (Zou 2003). The Pearson statistic measures the extent by which two instruments numerically rank subjects in the same order and depict the same relative magnitude of difference between subjects. It is directly proportional in the range of scores observed and inversely proportional to the extent those subjects occupy different numerical ranks in the two instruments. It is a parametric statistic (assumption of normality) and correlation coefficients can range from –1 (perfect inverse relationship) through 0 (no relationship) to +1 (perfect direct relationship). Mathematically the estimate of the true correlation coefficient equals the estimates of the covariance of (x,y) divided by the square root of the product of the variance of x and the variance of y. Other statistical procedures used in validation studies can include Student's t-test and effect size calculations.

Table 3: Examples of Validity Testing

Form of Validity	Example
Face	A subjective assessment by experts (i.e. oncologists) and patients (cancer patients) of a putative HRQoL instrument in which the following question is answered: "Does the PCRT instrument appear to contain items that are relevant to the questionnaires' purpose of assessing HRQoL effects of late prostate cancer radiation toxicity" No statistical methodology is usually employed; however, a range of expertise is usually employed to generate face validity.
	An assessment of the comprehensive nature of the questionnaire. Does the questionnaire assess both the

Sampling	frequency/intensity of the symptom (i.e. dysuria) and the potential impact (i.e. bother from dysuria) of that symptom. In addition, do the questions span the appropriate physical, social, emotional, and physical domains of interest?
Criterion	Development of a HRQoL instrument (e.g. PCRT late toxicity HRQoL instrument) that can predict for a gold standard binary (i.e. yes/no) event such as grade 3-5 toxicity (as measured by a validated toxicity scale). Other important binary events can include death and tumor control/progression.
Concurrent	Similar to criterion validity but the assessment is at a point of time. An example of concurrent validity is the development of a cancer-related short HRQoL instrument to predict for an analogous longer version HRQoL instrument.
Construct	Multiple hypotheses to be tested in comparison to other HRQoL instruments (e.g. PCRT vs. PCQoL, SF-36®, FACT-G©) or in different populations (e.g. differences in PCRT scores with brachytherapy vs. external-beam RT).
Responsiveness	Assessments of changes over time of a HRQoL instrument either after therapy (progression of PCRT scores over time) or in response to therapy (assessments of differences in PCRT score pre/post RT). The determination of clinical important changes is vital to the interpretation to changes in HRQoL.
External	The assessment of a HRQoL instrument needs to be interpreted in the light of other questionnaires (PCRT vs. PCQoL, Swedish and Chicago RT questionnaires). In addition, external validity can be assessed by introduction of the questionnaire to new populations (e.g. translations of the EORTC QLQ-C30© general cancer instrument into many languages.