Objective Measurement of Subjective Well-being

Elizabeth A. Hahn

In everyday situations and during unforeseen circumstances, each of us evaluates the impact of a particular decision in terms of its effect on our quality of life. Although the construct is subjective and is best assessed by self-report, researchers have created acceptable definitions and useful ways to measure it. The following definition is widely accepted for health-related quality of life (HRQOL): "...patients' appraisal of and satisfaction with their current level of functioning as compared to what they perceive to be possible or ideal." (Cella & Cherin, 1988). There are many instruments available to assess HRQOL dimensions such as physical or emotional well-being, as well as disease- or treatment-specific dimensions (Berzon et al., 1995).

Quality of Life in Cancer Treatment

HRQOL is an important consideration in cancer treatment, and healthcare providers seek to improve both the quantity and the quality of their patients' lives. Some cancer types, such as metastatic breast cancer, cannot be cured with currently available therapeutic agents, so the objectives of treatment are directed toward other goals (symptom relief, functional status, prolongation of life). In these patients, the quality of their survival may be as important as the length of their survival. In other types of cancer, the optimal treatment is unknown, and decision-making can best be made by taking into account patient preferences and HRQOL. For example, information about the impact of a disease and its treatment on HRQOL is invaluable for the prostate cancer patient who must decide between 'watchful waiting' vs. surgery, radiation therapy or hormonal therapy, each of which has its own risks and benefits. When treatment costs and health outcomes vary, healthcare providers can use information about preferences and HRQOL to optimize outcomes management.

The focus on HRQOL as an important clinical endpoint in cancer treatment is international in scope. With the availability of multiple language versions of HRQOL instruments, researchers and clinicians are beginning to evaluate the effects of cultural differences on HRQOL measurement. Cross-cultural evaluation of HRQOL and pooling of international research data require unbiased measures of the defined constructs that can detect clinically important differences between patients. Detected differences must not be caused by items that may function differently depending upon patient characteristics.

Elizabeth Hahn

Elizabeth Hahn is a Research Associate with the Institute for Health Services Research and Policy Studies at Northwestern University, and Director of Biostatistics and Data Management Systems at the Center on Outcomes, Research and Education (CORE) at Evanston Northwestern Healthcare. She is a medical sociologist and biostatistician with extensive experience in the design, implementation, coordination and statistical analysis of clinical trials and survey research studies. She also serves as a statistical consultant to international collaborative groups regarding research design and analysis. Her current research includes a focus on methodological and cross-cultural issues in the measurement of health-related quality of life and treatment satisfaction for patients with cancer and other chronic illnesses.

In 1999, she was awarded a two-year grant by the Agency for Healthcare Research and Quality to develop and evaluate a computer-based measurement program for quality of life assessment in low literate cancer patients. She is also the principal investigator on a project to develop a treatment satisfaction scale for cancer, HIV and other chronic illnesses, and a project to evaluate literacy assessment methods and patient preferences and attitudes towards literacy screening.

Cross-Cultural Equivalence

Several types of cross-cultural equivalence have been discussed in the literature, with varying degrees of agreement on definitions and hierarchy (Flaherty et al., 1988; Hui & Triandis, 1985). The universalist approach to cross-cultural research acknowledges that HRQOL concepts may differ across cultures and that this must be evaluated prior to performing comparative analyses. This paper illustrates the use of objective measurement to evaluate item equivalence (commonly defined as items that are relevant and acceptable in both cultures, and that measure the latent trait similarly) and metric/scalar equivalence (the construct is measured on the same metric and locates similar individuals at the same point on the scale).

METHODS

Quality of Life Instruments

The Functional Assessment of Cancer Therapy-Breast (FACT-B; Brady et al., 1997) developed in English, is available in 18 other languages, including German. It includes a general assessment of physical, functional, social/family and emotional well-being as well as a nine-item subscale to assess breast-cancer specific concerns. There are five response categories for the items: "not at all" ("berhaupt nicht" in German), "a little bit" ("ein wenig"), "somewhat" ("m((ig"), "quite a bit" ("ziemlich") and "very much" ("sehr"). The English version of the nine items in the breast cancer subscale are:

I have been short of breath
I worry about the risk of cancer in other family members
I am self-conscious about the way I dress
I worry about the effect of stress on my illness
One or both of my arms are swollen or tender
I am bothered by a change in weight
I feel sexually attractive
I am able to feel like a woman
I am bothered by hair loss

The FACT-B is part of the Functional Assessment of Chronic Illness Therapy (FACIT) quality of life measurement system (Cella, 1997). The initial cultural adaptation of FACIT instruments is based on a sequential approach for the development of internationally applicable quality of life measures, i.e., the instruments are translated from English into other languages (Bullinger et al., 1993). The adaptation methodology involves an iterative forward-backward translation, extensive review and evaluation by bilingual health professionals, and pretesting with patients (Bonomi et al., 1996; Lent et al., 1999).

Patients

The U.S. sample was a subset of 1,616 cancer patients enrolled in a validation study of the FACT-B during 1994-1997. White, English-speaking breast cancer patients (n=195) were selected as a comparison group for the Austrian patients (n=118) who completed the questionnaire in German while receiving treatment at two outpatient clinics during 1995.

Rasch Measurement Model

Rasch (1960) developed the logistic measurement model for the probability of a "correct" response with dichotomous data. This project used an extension of the model for rating scale data i.e., items with ordered response categories such as those used in the FACT-B (Wright & Masters, 1982). The model has three components: 1) an estimate of each patient's "ability" to achieve a high score (high HRQOL), 2) an estimate of each item's "difficulty" (the degree to which an item would be unlikely to be answered in a manner reflecting a high HRQOL) and 3) response "thresholds" for each "step" in the rating scale (there are m-1 steps in an m-category scale). The decisive property of Rasch models is that the person abilities and item difficulties can be estimated independently by means of conditional maximum likelihood estimation, resulting in sample-free question calibration and test-free patient measurement. In the rating scale model, the thresholds can be estimated once for a set of questions.

Item and Metric/Scalar Equivalence

The extent to which items in a questionnaire perform similarly across different reference groups is of critical interest when determining whether a given questionnaire can be used as an unbiased basis for comparing groups. The Rasch model allows us to identify items displaying differential item functioning (DIF). The most important indicator of DIF is not whether items systematically differentiate relevant subgroups, but whether they do so in an unmodeled (i.e., unpredicted) way. Unmodeled differences reflect differential interaction between some items and some persons, which in turn confuses interpretation of results. Items that differentiate groups can be identified and investigated as to their content to determine the likely source of DIF. DIF detecting procedures were applied in four steps: 1) After evaluating and anchoring the step threshold estimates on the entire sample, separate item calibrations were obtained for the two samples. 2) The calibrated item difficulties were plotted against each other. 3) An identity line and statistical control lines (95% confidence limits) were drawn on the plots to guide interpretation and assessment of possible bias (Wright & Masters, 1982). 4) Items identified as possibly biased (displaying DIF) were reviewed to obtain direction on interpreting the plots and determining the appropriate disposition of the item, given the content and the context of the misfitting item. The end product of these analyses and plots is an unbiased subset of items to be used for obtaining patient HRQOL measures on a common, linear metric. The patient measures, rather than raw scores, can then be used for analysis.

The nine breast cancer-specific items in the FACT-

B were evaluated to determine the extent to which they define a unidimensional construct of disease-specific HRQOL. All of the negatively worded items e.g., "I have been short of breath", were reversed in the analyses and item calibrations were reported as logits (log-odd units), with a higher value representing greater item difficulty. The WINSTEPS computer program (Linacre & Wright, 1998) was used to conduct the Rasch model analyses, and SAS software was used to make item difficulty plots.

RESULTS

Patient Characteristics

The majority (57%-60%) of patients (all women) in both groups had no current evidence of disease and few limitations in performance status (81% were classified at the highest level of functioning). The groups were also similar in terms of prior treatment history and current living arrangement. The U.S. group was slightly older and had a higher proportion of patients currently undergoing chemotherapy or receiving hormonal therapy.

Rasch model analyses

Using response thresholds from the combined analysis, separate item calibrations were obtained for the two patient groups and plotted against each other. Only one item ("I am self-conscious about the way I dress") functioned differently across groups. It was more difficult for the Austrian patients. A translation error was discovered in the German language version of this item, which may account for its apparent misfit. The other eight items in the module functioned similarly across groups, suggesting that they can be used to create unbiased measures of HRQOL in Austrian and U.S. breast cancer patients.

DISCUSSION

There is a growing body of literature on cross-cultural evaluation of HRQOL, yet few researchers have appreciated the advantages offered by objective measurement models to control bias and to construct reproducible linear measures. Estimating sample-free item calibrations and test-free person measures provides assurance that the analysis of HRQOL will not be impeded by measurement difficulties.

The limitations of traditional analysis methods to detect bias across different groups of subjects are discussed by Wright, Mead & Draba (1976). Common methods include regression using an external criterion of bias, comparison of factor structures, item-by-group interaction terms in analysis of variance and comparison of the proportion of subjects answering each item correctly. While these methods provide important information about how items function in different groups, they cannot adjust for unequal distributions of person abilities (sample dependency), heterogeneity of item difficulty variance and nonlinearity of raw scores. Rasch measurement model specifies that each item has an inherent property (difficulty level) that does not depend upon any particular sample, and that each person has a characteristic ability (in this case, level of HRQOL) that does not depend upon the particular items used in a test/instrument.

The study reported here demonstrates the usefulness of the Rasch model in evaluating the cross-cultural equivalence of HRQOL instruments. Statistical as well as conceptual criteria were used to determine which items were functioning differently in Austrian and U.S. breast cancer patients. The identification of biased items does not invalidate the questionnaire, but rather enables a better estimate of each cultural group's HRQOL.

References

Berzon R.A., Donnelly M.A., Simpson R.L. Jr., Simeon G.P., & Tilson H.H. Quality of life bibliography and indexes: 1994 update. Qual Life Res, 1995, 4: 547-569.

Bonomi A.E., Cella D.F., Hahn E.A., Bjordal K., Sperner-Unterweger B., Gangeri L., Bergman B., Willems-Groot J., Hanquet P., & Zittoun R. Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Qual Life Res, 1996, 5, 309-320.

Brady M.J., Cella D.F., Mo F., et al. Reliability and validity of the Functional Assessment of Cancer Therapy-Breast (FACT-B) quality of life instrument. J. Clin Oncol, 1997, 15: 974-986.

Bullinger M., Anderson R., Cella D.,& Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. Qual Life Res, 1993, 2, 451-459.

Cella D.F. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT Scales) - Version 4. Evanston, IL: Center on Outcomes Research and Education (CORE), Evanston Northwestern Healthcare & Northwestern University, November, 1997.

Cella D.F., & Cherin E.A. Quality of life during and after cancer treatment. Compr Ther, 1988, 4, 69-75.

Flaherty J.A., Gaviria F.M., Pathak D., Mitchell T., Wintrob R., Richman J.A., & Birz S. Developing instruments for cross-cultural psychiatric research. J. Nerv Ment Dis, 1988, 176, 257-263.

Herdman M., Fox-Rusby J., & Badia X. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res, 1997, 6, 237-247.

Hui C.H., & Triandis H.C. Measurement in cross-cultural psychology: a review and comparison of strategies. J. Cross-Cultural Psychol, 1985, 16, 131-152.

Lent L., Hahn E., Eremenco S., Webster K., & Cella D. Using cross-cultural input to adapt the Functional Assessment of Chronic Illness Therapy (FACIT) scales, Acta Oncologica, 1999, 38: 695-702.

Linacre J.M., & Wright B.D. A User's Guide to BIGSTEPS/WINSTEPS/MINISTEP: Rasch-Model Computer Programs. Chicago: MESA Press, 1998.

Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut, 1960 (Chicago: University of Chicago Press, 1980).

Wright B.D., Masters G.N. Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press, 1982.

Wright B.D., Mead R., & Draba R. Detecting and correcting test item bias with a logistic response model. MESA Research Memorandum Number 22, 1976.