VALUE IN HEALTH 16 (2013) 848–854

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

METHODOLOGICAL ARTICLES

Validating and Improving the Reliability of the EORTC QLQ-C30 Using a Multidimensional Rasch Model

Ching-Lin Shih, PhD[1], *Chien-Hung Chen, MD, PhD*[2], *Ching-Fan Sheu, PhD*[3], *Hui-Chu Lang, PhD*[4,*], *Ching-Lin Hsieh, PhD*[5]

[1]*Center for Teacher Education, National Sun Yat-sen University, Kaohsiung, Taiwan;* [2]*Department of Internal Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan;* [3]*Institute of Education, National Cheng Kung University, Tainan, Taiwan;* [4]*Institute of Hospital and Health Care Administration, National Yang-Ming University, Taipei, Taiwan;* [5]*School of Occupational Therapy, College of Medicine, National Taiwan University and Department of Physical Medicine and Rehabilitation, National Taiwan University Hospital, Taipei, Taiwan*

ABSTRACT

Objective: The reliability and validity of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) has not been examined while taking into account the correlation between subscales. The reliability of the EORTC QLQ-C30 subscales is modest, thus limiting their utility in both clinical and research settings. The purpose of this study was to validate the factor structure of multiple-item subscales of the EORTC QLQ-C30 and to improve their reliability by means of an item response analysis by using the multidimensional partial credit model. **Methods:** A total of 2295 patients with complete data were used for the analysis. One- and nine-dimensional partial credit models were used to fit the data to validate the construct validity of the multiple-item subscales of the QLQ-C30. **Results:** The model comparison showed that the nine-dimensional factor structure of multiple-item subscales was satisfactory. The multidimensional partial credit model fit data of the multiple-item subscales of the QLQ-C30 reasonably well. The estimated test reliabilities of each domain obtained from the multidimensional approach were higher than those obtained from the unidimensional approach. **Conclusions:** The constructs represented by the multiple-item subscales of the QLQ-C30 were validated. The improved reliability of the multiple-item subscales of the QLQ-C30 under the multidimensional approach can facilitate their applications in clinical and research settings.
Keywords: cancer, quality of life, Rasch model, questionnaire.

Introduction

The European Organization for Research and Treatment of Cancer's Quality of Life Questionnaire (EORTC QLQ-C30) was designed to assess the impact of disease and clinical treatment on cancer patients' daily lives [1]. The QLQ-C30 has been translated into more than 54 languages and is widely used to measure cancer-specific quality of life (QOL). A number of studies have investigated the validity, reliability, and other related measurement properties of the QLQ-C30 [2–7]. The conclusions of these studies have been that the questionnaire is generally an excellent QOL instrument, with good psychometric properties relevant to different cancer-patient populations [2–4,8–11]. However, the internal consistency, as indicated by Cronbach's alpha, of some of the QLQ-C30's subscales has been found to be poor. For example, it was reported that Cronbach's alpha of the cognitive-function subscale ranged from 0.30 to 0.71 [12–23]. For the nausea-and-vomiting [NV] subscale, it was only 0.49 [12]. Poor internal consistency of these subscales can lead to unreliable evaluations regarding patients' psychological discomfort [12] and can mislead clinicians into making incorrect clinical decisions or inappropriate interpretations of assessment results.

The QLQ-C30 is scored on the basis of classical test theory (CTT), and uses the total item score as the scale score. In other words, the main focus of CTT is on test-level information [24]. The correlation between subscales is not taken into consideration by a CTT analysis when several subscales are analyzed together. This causes theoretical difficulties when CTT is applied to analyze a scale that consists of several subscales. However, item response theory, especially the Rasch model, is usually used to analyze item response data and to provide item-level information, regardless of whether it is unidimensional or multidimensional. Through Rasch analysis, the patients' original ordinal responses can be transformed into interval scales [25,26]. Within the interval scale, equal intervals between any two points on a latent trait are equal in value. Relative to the ordinal scale, the interval scale can accurately reflect the true magnitude of the difference between repeated assessments and it is more accurate to show changes over time of a patient or difference between patients. Therefore, the interval scale can truly reflect and

Note: Both Hui-Chu Lang and Ching-Lin Hsieh contributed equally to this work.

* *Address correspondence to:* Hui-Chu Lang, Institute of Hospital and Health Care Administration, School of Medicine, National Yang-Ming University, No.155, Sec. 2, Linong Street, Beitou District, Taipei 11221, Taiwan.

E-mail: hclang@ym.edu.tw.

http://dx.doi.org/10.1016/j.jval.2013.05.004

compare treatment effect [27]. Furthermore, an interval scale can be analyzed by using parametric statistics, which are often more powerful than nonparametric methods [28]. Therefore, the Rasch model is believed to be an objective measurement model and useful for constructing interval scales [25,29,30].

A multidimensional item response model can improve the reliability of a measuring instrument that has subscales by taking into account the correlation among different subscale responses by the same individual [31]. Besides, because measurement error is taken into account in the estimation of between-dimensional correlations in the multidimensional model, the correlation estimates are then free from the attenuation caused by measurement error [32]. Thus, the multidimensional approach allows a patient's responses on subscales to complement one another, thereby producing a more accurate and reliable assessment of the patient's status. This makes it possible to increase the reliability of each subscale without increasing the number of items or the physical burden on patients, which is especially useful for those who are extremely ill [1,33,34].

Although the QLQ-C30 is usually administered as a whole, subscales of the QLQ-C30 have been used separately and independently in clinical studies. For example, three subscales of the QLQ-C30 (emotional function [EF], social function [SF], and global health status/quality of life [QL]) were taken as independent predictors to predict the out- of-hospital length of stay within the first 30 days [35]. In addition, three subscales (pain [PA], NV, and fatigue [FA]) were scored separately, and they were used to measure three symptoms that related to cancer patients' QOL [36]. As mentioned earlier, however, previous studies have found poor internal consistency in some of the subscales (e.g., cognitive function [CF], SF, physical function [PF], role function [RF], NV, constipation, diarrhea, and PA), which might not only have threatened the validity of the results of these studies [12–23] but also hamper the utility of the QLQ-C30 in future studies. The purpose of the present study was to analyze the responses to the QLQ-C30 from a sample of cancer patients by using a multidimensional item response model. Specifically, we examined the improvement in the results when using the multidimensional approach compared with using the unidimensional approach regarding the reliability estimates of the multiple-item subscales.

Methods

Subjects

Patients aged 18 years or older and diagnosed with cervical, breast, lung, liver, or colorectal cancer were recruited through senior-specialist referrals from the Taipei Veterans General Hospital, Taichung Veterans General Hospital, Kaohsiung Veterans General Hospital, and the Koo Foundation Sun Yat-sen Cancer Center from 2003 to 2004. These medical facilities are located in the northern, central, and southern regions of Taiwan. The participants were primarily recruited from outpatients who had received therapy for more than 3 months. Patients with hepatocellular carcinoma, however, were recruited from among inpatients. The other selection criterion was the ability to communicate in Mandarin Chinese or Taiwanese. Patients who were unaware of their own medical conditions were excluded. This study was approved by the institutional review boards of all medical centers involved. All participants gave written informed consent before participating in the face-to-face interviews.

Procedure

At each of the participating medical facilities, a trained research nurse interviewed the participants individually in a secluded interview room after the patient's routine consultations.

Questionnaire

The QLQ-C30 includes nine multiple-item subscales and six single-item subscales, resulting in 15 domains measured by 30 items [20]. Because latent factors cannot be well defined by a single item and at least two or three items are recommended [37], this study focused only on multiple-item subscales. The longest scale, PF, is measured by five items, followed by EF and FA, which are measured by four and three items, respectively. In addition to the three subscales, 6 of the remaining 12 subscales are measured by two items. These six are RF, CF, SF, QL, NV, and PA. Within these subscales, each item has four response categories —"not at all," "a little," "quite a bit," and "very much"—and are scored as 1, 2, 3, and 4 in this study, except for the QL subscale, in which each item is scored from 1 ("very poor") to 7 ("excellent"). To rate all subscales in the same direction, the QL subscale was scored inversely and therefore for all subscales a higher score indicated a lower QOL.

Data Analysis

Because missing responses for any items in a domain could produce a large bias in the parameter estimates of the corresponding domain, 205 patients in the sample with missing observations were excluded from the analysis. A total of 2295 (out of 2500) cancer patients with complete data were included in the analysis. The data set that comprised multiple-item subscales was analyzed.

The unidimensional and between-item multidimensional versions of the partial credit model (PCM) were used to fit the responses for polytomous items in this study. For the unidimensional PCM [38]—an extension of the Rasch model—the subscales of the QLQ-C30 were analyzed individually, and a set of step difficulties that determine the threshold locations on the latent continuum should be estimated. This approach requires that each scale meets the "unidimensionality" assumption; that is, the scale should measure only one domain. Any other domains or sources of variation are considered confounding and are not expected to be included in the PCM analysis because of their influence on the accuracy of the estimations.

The between-item multidimensional PCM (MPCM), a special case of the multidimensional random coefficients multinomial logit model (MRCMLM) [39], considers responses from several subscales simultaneously. The MRCMLM is a member of the family of Rasch models, and it shares the measurement properties of these models [40]. It allows a general model to be written that includes most of the existing Rasch models, such as the unidimensional and multidimensional PCM used in this study. The equation and parameter explanation of MRCMLM are shown in Appendix A in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2013.05.004. To address multiple responses from the same patient, MPCM specifies a correlation structure between domains in its model formulation. In general, this model better fits the data than its unidimensional counterpart and produces parameter estimates that are more accurate [31]. ConQuest computer software [41] was used to perform the analysis.

Two models, unidimensional and between-item multidimensional PCM, were fitted to the data, and their overall fits were compared. Because the two competing models are nested model, the fit of these two models can be compared with the likelihood ratio test [42]. The likelihood ratio test statistics, G^2, defined as -2loglikelihood, is approximately chi-square distributed with the degree of freedom equal to the difference between the number of parameters of two models [43]. The model with smaller deviance, that is, greater likelihood, would be expected to be closer to the true model and therefore was selected in this study [44]. Furthermore, the between-dimensional correlations can be estimated with PCM ability estimates, MPCM ability estimates, and