Volume 7 • Supplement 1 • 2004
VALUE IN HEALTH

Requirements for Quality of Life Instruments in Clinical Research

Lynda C. Doward, BSc,[1] David M. Meads, BA,[1] Hanne Thorsen, MD[2]

[1]Galen Research, Manchester, UK; [2]Department of General Practice, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

The ability to produce high quality instruments for the assessment of quality of life has advanced considerably in recent years. As the science progresses it has become clear that certain standards must be met if outcome measures are to be capable of providing useful, reliable, and valid information within the context of clinical studies and trials. This paper specifies what these standards are with particular reference to theoretical basis, practicality, acceptability to respondents, unidimensionality, scaling and psychometric properties, and cultural validity and equivalence. The paper also indicates how failure to achieve such standards results in measures that are inaccurate and insensitive to true changes in outcome.

Keywords: construct validity, cultural equivalence, dual panel translation, quality of life.

Introduction

> "All this he knows but will not tell
> To those who cannot question well"
> Percy Bysshe Shelley [1]

There are a number of requirements that any instrument intended to collect information from patients should meet. As the importance of a number of these has only recently been recognized, it is unlikely that many existing measures will meet them all. Although failure on some of the requirements may not be serious enough to preclude their use in a clinical trial, it may affect the way data are analyzed and limit inferences that can be drawn from these analyses.

Instrument Development

Theoretical Basis

All instruments should be based on a stated model or theory of the construct being measured; a requirement usually neglected in the development of patient-reported outcome (PRO) measures. There is a tendency for investigators to refer to any PRO measure as a quality of life (QoL) instrument, even where it was designed for a different purpose. In their critical appraisal of the quality of QoL measures, Gill and Feinstein [2] found that investigators had defined QoL conceptually in only 15% of 75 articles they reviewed. This raises two related problems. First, unless investigators explicitly state what they mean by QoL it is not possible to determine whether their definition is reasonable. Second, without such information it is not possible to establish that an instrument has construct validity [3].

Content Derivation

It is now generally agreed that the content of any QoL measure must be derived directly from relevant respondents [4,5]. This applies both to the issues covered by the instrument and (as far as possible) to the actual wording of its items. By deriving the items from relevant individuals it is possible to ensure that appropriate QoL concerns are included [6,7]. Measures consisting of questions written by the authors or drawn from the literature reflect professional views of the experience of disease, rather than those of potential respondents.

The content of an instrument should also be appropriate to the culture and lifestyle of the countries in which it will be used. Instruments developed in a single country reflect the language and culture of that particular society, resulting in problems when adapting the measure into other languages. Deriving the content from patient interviews conducted in different countries maximizes the cultural relevance of the measure [8].

Address correspondence to: Lynda Doward, Galen Research, Enterprise House, Manchester Science Park, Lloyd Street North, Manchester M15 6SE, UK. E-mail: ldoward@galen-research.com

 1098-3015/04/$15.00/S13 S13–S16

Acceptability to Respondents

The respondents should perceive the content of a measure as relevant. One of the major disadvantages of the generic health status instruments is that, by definition, they will include some items that are inappropriate for the specific health problem and miss some areas of importance. By deriving the content of the instrument from relevant sources, problems of nonapplicability are avoided and face and content validity maximized.

Psychometric Properties

Unidimensionality

Any scale must be unidimensional if it is to provide valid change scores. This applies to both single-scale instruments and to individual scales within multidimensional instruments. Until recently, test-developers have relied on internal consistency coefficients to indicate unidimensionality but this statistic merely indicates the degree of interrelation between the items in a scale [9]. Different scales can be added together and still have relatively high internal consistency [3]. Factor-analytic methods have also been used to assess the dimensional structure of scales. However, these are parametric methods requiring interval level data and so their use may not always be scientifically valid. In addition, it has been known for some time that factor analysis, particularly with dichotomous data, produces spurious factors [10]. Rasch analysis is now considered to be the most efficient means of establishing unidimensionality [11]. However, such analyses have not been applied to most existing patient-completed outcome measures. Where such analyses have been applied, the results tend to have been reported in inadequate detail [12,13].

Reliability (Reproducibility)

Another essential quality of scales is that they should have good reliability (that is, low random measurement error). The appropriate method of determining the reliability of an instrument intended for use in a clinical trial is to apply it twice to the same population and correlate the scores, as stability over time (or reproducibility) is the crucial variable. Where reliability is low; that is, where the correlation coefficient is below 0.85 [14] an instrument will be unlikely to show changes in a patient's QoL.

Internal Consistency

Internal consistency is usually assessed through Cronbach's alpha coefficients and informs on the degree of interrelatedness of items. This property, although not essential in itself, is a prerequisite for unidimensionality. Items have to be interrelated if they are to measure the same underlying construct. At the same time Streiner and Norman [3] warn against scales with very high alpha coefficients, as this indicates item redundancy.

Face and Content Validity

Potential respondents should perceive the content of any instrument as relevant. Too often test-developers check the content of measures with clinical experts only, neglecting the views of patients. Patient interviews are required covering the suitability and completeness of the content of the questionnaire and whether it is easy to understand and complete [6,8].

Construct Validity

It is essential to establish that a new instrument has construct validity, that is, that it is measuring the intended construct. Three prerequisites of this are that the instrument is based on a coherent theory or definition [3], that the scale is unidimensional [15], and that it has good reproducibility [3]. Where these criteria are met it is possible to infer that the instrument provides a valid assessment of the construct defined in the model. However, it is still necessary to assess construct validity formally. Assessing the construct validity of a health outcome measure requires the specification of hypotheses and testing of relationships among clinical and other PRO variables. There are many ways to do this, the most common being through association with instruments measuring related constructs (convergent validity) and unrelated constructs (divergent validity). The instrument employed as the comparator measure should itself be of proven reliability and validity. Thus, interscale correlations, essentially using a new instrument to validate itself, are not an adequate assessment of validity. Another valuable approach is known groups' validity. This approach includes distinguishing the status of patients at different stages of their disease or with different disease severity. If hypotheses are confirmed with observed data, then there is evidence-supporting validity. It is the accumulation of this evidence that provides greater confidence in the validity of a measure.

Responsiveness

Any instrument intended for use in clinical trials or for monitoring individual patient care should be responsive, that is, able to detect real changes in the

measured construct. Measures with poor reproducibility have poor responsiveness, as in the case of generic health status measures, such as the Nottingham Health Profile and the SF-36 [16–19]. The ability of an instrument to detect changes within a single group over time is necessary but not sufficient to indicate responsiveness. In order for an instrument to have satisfactory responsiveness, it is also necessary for it to have an adequate range of coverage of the construct. For example, it should be able to assess minor QoL impairment where the instrument is intended for use in preventive studies or where asymptomatic patients are included in the trials.

Practicality for Use in Clinical Trial

Instruments should be practical if they are to be included in clinical trials. They should be short and easy to answer and administer. Thus, complex response systems should be avoided, as should the requirement for trained administrators. In the latter case, this increases resource consumption and may reduce the willingness of respondents to acknowledge problems [20]. The optimal mode of administration for instruments included in a clinical trial is self-administration. Here, resource input is minimal and the likelihood that respondents will acknowledge problems is maximized. Where an instrument is well developed and tested, potential problems of missing data or misunderstanding will be minimized. Where repeated measurements are needed over a period of time the instruments should be administered in the same way and under the same conditions on each occasion.

Cultural Considerations

Language Availability

As most clinical trials are now international, instruments should be available in a large number of languages. However, problems can arise where attempts are made to adapt an instrument developed in one language for use in others. Given that the instrument is in a finished format, cultural differences may be impossible to overcome [8]. All new versions produced should be shown to have equivalent psychometric properties to those of the original language version. This requires proper validation studies to establish face and content validity, reproducibility, and construct validity. To a large extent cultural problems can be avoided where instruments are developed simultaneously in the countries for which they are required [21].

Given the expense of, and time required for adapting instruments for use in additional countries, careful consideration should be given to selecting the countries in which a trial will be run. Rather than recruiting a few patients from each of a large number of countries it is more efficient to select larger numbers of participants from fewer countries where validated versions of the outcome measures are already available.

Language Equivalence

A number of groups have developed several language versions of instruments without adequate consideration of issues of cultural equivalence (see for example [22–25]). It is always possible to translate a questionnaire into a new language but it does not follow automatically that the different versions are directly comparable. Where it is intended to combine QoL data from different countries, the requirements of the scaling properties of the instrument are increased. It is necessary to establish that the items in the scale, or in the individual subscales, have the same ordering, in terms of amount of the measured construct represented, in each language version. This assumes that there will be little or no cultural bias in the individual items [26].

Summary

While some of these requirements have developed or become more specific over recent years, they have formed the framework around which all needs-based measures have been built. By setting high psychometric standards it is possible to ensure that new instruments have the quality to be effective outcome measures in clinical practice and trials.

References

1 Shelly PB (Hutchinson T, ed.). The Complete Poetical Works of Percy Bysshe Shelley. London: Oxford University Press, 1927.
2 Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. JAMA 1994;272:619–25.
3 Streiner D, Norman G. Health Measurement Scales. Oxford: Oxford University Press, 1989.
4 Berzon R, Hays RD, Shumaker SA. International use, application and performance of health-related quality of life instruments. Qual Life Res 1993;2:367–8.
5 Jachuck SJ, Brierly H, Jachuck S, Wilcox PM. The effect of hypotensive drugs on the quality of life. J Coll Gen Pract 1982;32:103–5.

6 Hunt SM, McKenna SP. The QLDS: a scale for the measurement of quality of life in depression. Health Policy 1992;22:307–19.
7 Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622–9.
8 McKenna SP, Cook SA, Whalley D, et al. Development of the PSORIQoL, a psoriasis-specific measure of quality of life designed for use in clinical practice and trials. Br J Dermatol 2003; 149:323–31.
9 Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol 1993;78:98–104.
10 Stout WF. A nonparametric approach for assessing latent trait dimensionality. Psycometrika 1987; 55:293–326.
11 Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980.
12 Cella DF, Dineen K, Arnason B, et al. Validation of the functional assessment of multiple sclerosis quality of life instrument. Neurology 1996; 47:129–39.
13 Raczek AE, Ware JE, Bjorner JB, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. International quality of life assessment. J Clin Epidemiol 1998;51:1203–14.
14 Weiner EA, Stewart BJ. Assessing Individuals. Boston: Little Brown, 1984.
15 Kreiner S. Validation of index scales for analysis of survey data. In: Dean K, ed., Population Health Research—Linking Theory and Methods. Beverly Hills: Sage, 1993.
16 Hunt SM, McEwen J, McKenna SP. Measuring Health Status. London: Croom-Helm, 1986.
17 Brazier J, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. BMJ 1992; 305:160–4.
18 Nerenz DR, Repasky DP, Whitehouse FW, Kahkonen M. Ongoing assessment of health status in patients with diabetes mellitus. Med Care 1992;30 (Suppl 5):S112–23.
19 Dorman P, Slattery J, Farrell B, et al. Qualitative comparison of the reliability of health status assessments with the EuroQol and SF-36 questionnaires after stroke. United Kingdom. Collaborators in the international stroke trial. Stroke 1998; 29:63–8.
20 Guyatt GH, Jaeschke R, Feeny DH, Patrick DP. Measurements in clinical trials: choosing the right approach. In: Spilker B, ed., Quality of Life and Pharmacoeconomics in Clinical Trials, 2nd ed. Philadelphia: Lippincott-Raven, 1996.
21 Doward LC, McKenna SP, Kohlmann T, et al. The international development of the RGHQoL: a quality of life measure for recurrent genital herpes. Qual Life Res 1998;7:143–53.
22 Fukuhara S, Bito S, Green J, et al. Translation, adaptation and validation of the SF-36 health survey for use in Japan. J Clin Epidemiol 1998; 51:1037–44.
23 Aaronson NK, Ahmedzai S, Bergman B, et al. 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 1993;85:365–76.
24 Anderson RT, Aaronson NK, Bullinger M, McBee WL. A review of the progress towards developing health-related quality-of-life instruments for international clinical studies and outcomes research. Pharmacoeconomics 1996;10:336–55.
25 Bullinger M, Alonso J, Apolone G, et al. Translating health status questionnaires and evaluating their quality: the IQOLA project approach. International quality of life assessment. J Clin Epidemiol 1998;51:913–23.
26 Angoff WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H, eds., Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993.