Measuring Quality of Life After Stroke Using the SF-36

Rob J. de Haan, PhD

The SF-36 is the most widely used generic instrument for measuring quality of life (QOL). The instrument is translated into numerous languages, and the validity of the 8 subscales is confirmed in general populations and in a wide variety of patient groups in more than 2000 articles. In an article published in this issue of *Stroke*, Hobart et al report the psychometric properties of the SF-36 in a sample of ischemic stroke patients. The authors conclude that (1) some subscales, especially the scales for General Health (GH) and Social Functioning (SF), have limited reliability and validity; (2) half of the subscales suffer from floor and/or ceiling effects; and (3) the 2 summary scores inadequately reflect the patient’s physical and mental health. In view of the overwhelming weight of evidence that the subscales of the SF-36 are psychometrically sound to measure QOL in a range of patient populations, the question arises how convincing the arguments of Hobart and his colleagues are.

The authors argue that the GH and SF scales generate low reliability scores and have limited convergent and discriminant validity. However, these conclusions can be challenged. The reliability of only 1 scale (GH) was marginally less (Cronbach’s alpha = 0.68) than the authors’ predefined criteria. Although it is often recommended that coefficient values should be above 0.80, values above 0.70 are generally regarded as acceptable for scales when assessing outcome on a group level. Moreover, it should be noticed that the alpha coefficient not only depends on the correlations of the items but is also related to the number of items in the scale. For example, the relatively low coefficient (0.70) of the SF scale may also be explained by the fact that this scale encompasses only 2 items. Since Cronbach’s coefficients increase as the number of items is increased (or vice versa), one may wonder whether it is sensible to specify criteria for acceptable levels of the alpha coefficient without specifying the number of items in the scale. The authors’ criticism with regard to the convergent and discriminant validity of both subscales is not convincing. In general, the item-total subscale correlation of the GH and SF is above 0.40, indicating that the items in each scale measure a common underlying trait. Moreover, in both scales, all item-other scale correlations are higher than the item-other scale correlations (although less than 2 SE), indicating that the subscale items are best placed in the scales in which they already appear.

In this study, a number of subscales of the SF-36 exhibit floor and/or ceiling effects. However, as the authors noted, the psychometric properties of an instrument are sample dependent. The narrow range of many health scales partly results from the traditional methods used in the development of scales, such as item analyses. Unfortunately, these techniques for item analyses are highly dependent on the average level of the patients in the samples used in the psychometric evaluation of a scale. This means that the resulting scales may exhibit considerable ceiling and floor effects in score distribution when they are used with groups of patients with a lower or higher average level (e.g., mild to moderate stroke patients as in this study) of functional health.

To simplify the statistical analysis and to enhance the interpretation of the SF-36, the developers of the instrument recently made available scoring algorithms for aggregating subscale scores in 2 distinct summary scores: Physical Component Summary and Mental Component Summary. An important finding of Hobart et al is that the study results do not support the computation of summarized scores in stroke. These findings are in line with recent studies, which also demonstrate shortcomings of the summary scores in accurately reflecting patients’ physical and mental health on the basis of subscale scores. Taft et al showed that the discrepancies between subscale profile and component scores of the SF-36 are attributable to the way in which these summary scores are calculated. The main problem in the scoring algorithm derives from the use of negatively weighted subscale factor score coefficients, leading to inaccurately summarize profile scores and, sometimes, clinically counterintuitive study results.

To summarize, when the results of the relatively small study of Hobart et al are taken in conjunction with the findings of previous research, there is at present insufficient evidence to question the reliability and validity of the SF-36 subscales in stroke. The finding that the SF-36 is suffering from floor and/or ceiling effects can largely be explained by the specific characteristics of the mild to moderate stroke patients studied. However, a point in case is their finding that the assumptions for generating 2 summary scores could not be supported. Until the current scoring method is statistically revised, it is advisable not to use the component scores in stroke research.

In the light of the study results presented by Hobart et al, some additional remarks should be made with regard to the measurement of functional outcome using traditional multi-item instruments such as the SF-36. Typically, these scales...
calculate the total scale score for each patient using a (weighted) sum of the responses to each item. However, some serious problems are associated with this approach. Firstly, all items on a scale have to be presented to patients in order to obtain a summed score. This inefficiency has led researchers to shorten health instruments, resulting in more practical but less precise scales. Secondly, since summed scores are dependent on the number of, and precisely which, items are included in the instrument, it is impossible to compare scores obtained on different instruments, even if they measure the same health concept: 10 points on the Barthel are not the same as 10 points on the physical dimension of the Sickness Impact Profile.

Thirdly, the clinical interpretation of summated scores is not as straightforward as it may seem. For example, in the study of Hobart et al, stroke patients had a mean score of 47.6 on the subscale Physical Functioning. The clinical meaning of this SF-36 score would be unclear for most neurologists. This problem is amplified by the ordinal nature of summated scores, meaning that a given difference in scores at one point on the scale does not necessarily represent the same amount of functional change as an identical difference at another point on the scale. Following growing dissatisfaction with the classical approach, an alternative method has been introduced: item response theory (IRT).* This statistical paradigm uses a logistic regression-type analysis to model the responses of the patients to the individual items. Using this technique, both patients and items can be placed on the same hierarchical continuous scale. There are a number of advantages to the use of IRT techniques in clinical measurement. Firstly, not all items in an instrument have to be presented to all patients to assess their functional health. Thus, more difficult items (eg, vigorous activities) can be presented to less disabled patients and easier ones (eg, bathing or dressing) to more severely impaired patients. This approach leads to a more efficient data collection method known as adaptive testing and results in a considerable reduction of floor or ceiling effects. Secondly, even if different subsets of items are presented to subgroups of patients, the measurements of their functional level remain completely comparable. This is because the difficulty of each item (its position on the linear scale) has been estimated beforehand. Thirdly, the clinical interpretation of functional measurements is straightforward because the patients’ level of functional health can be directly compared with the hierarchically ordered items on the linear scale. In spite of recent interest in IRT in clinical outcome measurement, these methods have not yet been developed in stroke research. Perhaps IRT is not suitable for the development of scales measuring a subjective and multidimensional construct such as QOL. However, with regard to more direct and tangible manifestations of disease such as physical disability, IRT is probably a useful supplement to the traditional approach.

References
Measuring Quality of Life After Stroke Using the SF-36
Rob J. de Haan

Stroke. 2002;33:1176-1177
doi: 10.1161/01.STR.000015223.98452.97

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/33/5/1176

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/