Validity and reliability of the Medical Outcomes Study Short Form-20 questionnaire as a measure of quality of life in elderly people living at home

DANIEL J. CARVER, CHERE A. CHAPMAN, VINCE SALAZAR THOMAS, KAREN J. STADNYK, KENNETH ROCKWOOD

Division of Geriatric Medicine, Dalhousie University, Halifax, Nova Scotia, Canada

Address correspondence to: D. J. Carver, Centre for Health Care for the Elderly, Queen Elizabeth II Health Sciences Centre, 1763 Robie Street, Halifax, Nova Scotia, Canada B3H 3G2. Fax: (+1) 902 473 7133

Abstract

Background: the Medical Outcomes Study Short Form-20 (SF-20) questionnaire is recommended for health-related quality of life research, but there is little information on its utility in older people. We assessed the validity, reliability and feasibility of using the SF-20 in an elderly community-dwelling population.

Methods: the SF-20 was administered to a stratified, random sample of 333 elderly subjects.

Findings: assessment of content validity revealed that important domains were lacking, while others appeared to be inappropriately combined. Using Spearman correlation coefficients, the SF-20 had acceptable convergent and discriminant validity. A principal components analysis provided evidence for internal consistency for some of the subscales. Evidence for test–retest reliability was good.

Interpretation: while the reliability and feasibility of the SF-20 appear satisfactory, concerns about validity and responsiveness should temper enthusiasm for its use with elderly people living at home.

Keywords: elderly, Medical Outcomes Study Short Form-20, quality of life, validation

Introduction

There is a need to address health-related quality of life in elderly people but no standard outcome measure exists [1–3]. Important psychometric properties of validity, reliability and responsiveness to change of many measures are inadequate or undocumented for use in elderly subjects [4]. The Medical Outcomes Study Short Form-20 (SF-20) has been proposed for health-related quality of life research although it is conceded that "...there is little information on its utility for frail older patients" [5]. A related instrument, the SF-36, has been proposed for the National Health Service [6, 7].

To evaluate the SF-20, we studied its measurement properties in a representative sample of community-dwelling elderly people contacted in an interim follow-up study of the Canadian Study of Health and Aging (CSHA) [8]. In 1991 the CSHA made a nationwide survey of the prevalence of dementia and health status of elderly Canadians. An interim follow-up was conducted in 1993 to maintain contact with the cohort, and to assess limited health outcomes before a more comprehensive reassessment in 1996. The objectives of the current study were to examine the validity, reliability and feasibility of using the SF-20 to assess elderly subjects living at home.

Methods

Sample

Dalhousie University was one of 18 centres that participated in the CSHA [8]. The first phase of the CSHA was a cross-sectional survey of community-dwelling elderly Canadians which helped form a cohort for future studies. Subjects were aged 65 or over living in a non-institutional setting (i.e. not living in a nursing home, chronic care facility or collective dwelling such as a convent). In Nova Scotia, a random
clustered sample was drawn from the Medical Services Incorporated registration of beneficiaries of the provincial health insurance plan. The 1991 Nova Scotia subsample from the CSHA consisted of 555 subjects, 37 of whom refused further interviews. Of the remaining 498 potential subjects available for interview at the beginning of the interim recontact study in 1993 (91%), 395 agreed to be interviewed. Of these, 63 had a Modified Mini-Mental State Examination (3MS) score of ≤50 [9] and were therefore considered too cognitively impaired to complete the SF-20 questionnaire. The final sample consisted of 355 subjects (67% of the original sample) [8].

The project was approved by the research and ethics committee of the Queen Elizabeth II Health Sciences Centre. Informed consent was obtained from each subject or their legal guardian.

Measures
The SF-20 is a short multi-dimensional instrument which measures six aspects of health status: physical functioning (six items), role functioning (two items), social functioning (one item), mental health (five items), general health perceptions (five items), and bodily pain (one item). Scores for this instrument are specifically coded and calibrated so that each of the six dimensions is equally weighted. Each parameter score is transformed to a scale from 0 to 100, higher values indicating better status [10–12].

The SF-20 was incorporated into the larger CSHA questionnaire which also consisted of the Spitzer quality of life index [13], the Older Americans' Resources and Services (OARS) [14, 15], the Barthel index [16–18] and the 3MS [9].

The questionnaire was administered in face-to-face interviews at home. To assess reliability, a randomly chosen group (n = 19) was administered the SF-20 again by telephone. Three experienced and trained interviewers were employed for data collection.

Performance of the instruments
The six health dimension scores of the SF-20 were calculated for all subjects. Means and standard deviations were calculated by age group and sex.

Validity
This was assessed by addressing content and construct validity and by performing a principal components analysis. Content validity is the judgement of whether all relevant concepts and domains are represented in a measure [19], and was assessed in two ways. First, the parameters that make up the SF-20 were compared with Lawton’s theoretic constructs of quality of life in elderly people [20]. Lawton’s conceptual framework for the ‘good life’ consists of four interacting factors: psychological well-being, behavioural competence, perceived quality of life and objective environment. Secondly, face validity—whether the instrument appears to adequately assess the desired qualities it professes to measure—was assessed by an expert panel [19]. This panel consisted of two geriatricians, an epidemiologist, and a multidisciplinary team of health professionals working on a geriatric rehabilitation ward.

Construct validity is a measure of whether an instrument’s scores correlate with measures of other parameters in hypothesized ways [19]. This was assessed by addressing convergent and discriminant construct validity. Convergent construct validity was evaluated by comparing SF-20 total scores and subscales with the Spitzer quality of life index, Barthel index, OARS and OARS instrumental activities of daily living (IADL) total scores using Spearman correlation coefficients. Discriminant construct validity was assessed by correlating the SF-20 total score with that of the 3MS.

The 20 items of the SF-20 were subjected to a principal components analysis. The initial factor matrix was rotated orthogonally as we expected the dimensions to be independent, as claimed. We attempted to identify factors by grouping items (observed variables) with high loadings for a particular factor. We used a rigorous criterion of 0.5. Items with acceptable loadings but considerably lower than other items on the factor were carefully evaluated because they may not enhance the internal consistency of the scale derived from the group of variables.

Reliability
Informed by the factor analysis as well as the theoretical concepts behind the SF-20, we derived estimates of internal reliability for each of the cluster of variables identified in the factor matrix. Internal reliability was not reported for the two single-item indices. Internal reliability was estimated using Cronbach’s α coefficient. Higher values result from a strong correlation of scale items and suggest that the scale is measuring a single underlying dimension [21].

Potential item bias in each subscale of the SF-20 as well as the complete SF-20 scale was assessed by performing one-way analyses of variance to contrast the mean score of each scale across gender and age groups. t-tests were calculated to indicate any significant difference in scale score between subgroups. Test–retest reliability was assessed using an intra-class correlation coefficient with a subset of 19 subjects.

Analysis
Analyses were performed using the Statistical Package for the Social Sciences (SPSS) software program [22]. The intra-class correlation coefficient was calculated using the method described by Streiner and Norman [19].
Performance of the instruments

The mean values by age group and sex of the six health dimension scores of the SF-20 are shown in Table 1.

Descriptive statistics

The mean age of the subjects was 76 years, 47% were in the 66–75 age range and 10.6% were 86 years or older. More than half (58%) were female, 231 subjects (31%) lived alone and 90 (27%) had spent a night in the hospital in the past year. Most subjects lived in urban regions (96.7%). In general, our sample was functionally independent. Over 85% were in the top quartile of the Barthel index. Most subjects (84%) rated their health status as very good or pretty good and 72% rated their health as about the same or better than it was 2 years previously.

Older age groups tended to score lower on all subscales of the SF-20 except mental health. Men tended to score higher on the physical functioning subscale and women reported higher social functioning.

Floor effects were observed for three of the six subscales of the SF-20 (social functioning 5.1%, bodily pain 13.2% and role functioning 27.6%). Ceiling effects were noted for all six subscales (mental health 8.1%, health perceptions 13.8%, physical functioning 25.5%, bodily pain 47.1%, role functioning 55.9% and social functioning 65.5%).

In general, the scores of the Spitzer quality of life index were stable across all age groups. In contrast, the scores of the Barthel index, the OARS and the 3MS tended to decline with age.

Validity

The panel of experts’ assessment of the content validity of the SF-20 concluded that the items used in the instrument appeared valid. Compared with Lawton’s constructs, the SF-20 addressed all but the objective environment.

The expert panel did, however, raise concerns about the face validity of the SF-20 in an elderly population. Question 16F, for example, combines four personal activities of daily living: eating, dressing, bathing and using the toilet. Each of these activities have different functional considerations.

The panel identified several omissions in the SF-20’s ability to assess the elderly population. It does not assess memory, other aspects of cognitive function or self-administration of medication. The SF-20 also fails to consider environmental adaptation.

The SF-20 physical functioning subscale score correlated moderately with the Barthel index, OARS and OARS-IADL (Spearman correlation coefficients $r = 0.63, 0.65, 0.67$ respectively; Table 2). The SF-20 role function subscale correlated less well with these

### Table 1. Mean values (and standard deviations) of the six subscales of the Medical Outcomes Study SF-20 health survey administered to 333 elderly community-dwelling subjects ($n = 333$)

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Sex</th>
<th>$n^a$</th>
<th>Physical function</th>
<th>Role function</th>
<th>Social</th>
<th>Mental health</th>
<th>Health perception</th>
<th>Pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>66–70</td>
<td>M</td>
<td>37</td>
<td>77 (20)</td>
<td>78 (38)</td>
<td>86 (28)</td>
<td>85 (10)</td>
<td>77 (23)</td>
<td>65 (41)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>38</td>
<td>71 (26)</td>
<td>62 (44)</td>
<td>84 (26)</td>
<td>80 (18)</td>
<td>76 (24)</td>
<td>61 (39)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>74 (23)</td>
<td>70 (42)</td>
<td>85 (27)</td>
<td>85 (15)</td>
<td>76 (23)</td>
<td>65 (39)</td>
<td></td>
</tr>
<tr>
<td>71–75</td>
<td>M</td>
<td>32</td>
<td>70 (29)</td>
<td>70 (41)</td>
<td>86 (25)</td>
<td>88 (11)</td>
<td>73 (23)</td>
<td>73 (35)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>47</td>
<td>72 (25)</td>
<td>73 (39)</td>
<td>80 (31)</td>
<td>78 (17)</td>
<td>68 (27)</td>
<td>55 (41)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>72 (26)</td>
<td>72 (40)</td>
<td>82 (29)</td>
<td>82 (15)</td>
<td>70 (25)</td>
<td>62 (39)</td>
<td></td>
</tr>
<tr>
<td>76–80</td>
<td>M</td>
<td>42</td>
<td>69 (27)</td>
<td>64 (43)</td>
<td>80 (34)</td>
<td>83 (19)</td>
<td>66 (28)</td>
<td>76 (38)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>55</td>
<td>62 (27)</td>
<td>68 (43)</td>
<td>83 (30)</td>
<td>80 (15)</td>
<td>70 (24)</td>
<td>63 (38)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>65 (27)</td>
<td>67 (42)</td>
<td>82 (32)</td>
<td>82 (17)</td>
<td>68 (26)</td>
<td>69 (38)</td>
<td></td>
</tr>
<tr>
<td>81–85</td>
<td>M</td>
<td>18</td>
<td>72 (25)</td>
<td>65 (46)</td>
<td>67 (41)</td>
<td>80 (20)</td>
<td>63 (34)</td>
<td>71 (37)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>22</td>
<td>65 (23)</td>
<td>62 (46)</td>
<td>82 (30)</td>
<td>79 (15)</td>
<td>68 (23)</td>
<td>65 (32)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>68 (24)</td>
<td>62 (45)</td>
<td>86 (36)</td>
<td>80 (17)</td>
<td>66 (28)</td>
<td>67 (34)</td>
<td></td>
</tr>
<tr>
<td>86–90</td>
<td>M</td>
<td>5</td>
<td>65 (25)</td>
<td>83 (26)</td>
<td>87 (35)</td>
<td>83 (20)</td>
<td>84 (19)</td>
<td>77 (41)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>17</td>
<td>42 (29)</td>
<td>12 (28)</td>
<td>69 (39)</td>
<td>75 (18)</td>
<td>70 (28)</td>
<td>67 (41)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>48 (29)</td>
<td>30 (42)</td>
<td>74 (57)</td>
<td>77 (18)</td>
<td>73 (26)</td>
<td>70 (40)</td>
<td></td>
</tr>
<tr>
<td>91+</td>
<td>M</td>
<td>1</td>
<td>75 Missing</td>
<td>100</td>
<td>92</td>
<td>75</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>8</td>
<td>50 (32)</td>
<td>31 (46)</td>
<td>73 (35)</td>
<td>84 (7)</td>
<td>71 (21)</td>
<td>81 (29)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>53 (31)</td>
<td>28 (44)</td>
<td>76 (34)</td>
<td>84 (7)</td>
<td>71 (20)</td>
<td>83 (28)</td>
<td></td>
</tr>
</tbody>
</table>

*Number of valid observations.
measures (Spearman correlation coefficients of $r = 0.48$, 0.59 and 0.56 respectively). In terms of discriminant construct validity, the SF-20 physical functioning and role function subscale scores correlated poorly with the 3MS (Spearman correlation coefficient of $r = 0.2$ and 0.27; Table 2).

In our validation by principal components analysis we identified four factors. The five items from the mental health subscale hung together on factor 1 (loadings of 0.57–0.74). The health perceptions subscale item “I have been feeling bad lately” also hung on factor 1 (0.57) and loaded almost equivalently on factor 2. The other four items loading on factor 2 are consistent with the SF-20 health perceptions subscale. We thus consigned this item to factor 2, thereby fulfilling the standard items in the health perceptions subscale. Factors 3 and 4 identify eight items which load highly. These last two factors are not entirely consistent with established SF-20 subscales. Instead they blend items from the role functioning and physical functioning subscales.

Internal consistency reliability

Based on the above results, Cronbach’s $\alpha$ statistics were estimated for the established subscales of the SF-20 including health perceptions, mental health, role functioning, physical functioning, and also for the subscales suggested by factors 3 and 4 (Table 5). All subscales have internal reliability based on the typical threshold of $\alpha = 0.70$. Measures of internal consistency are not appropriately determined for the remaining two parameters as each had a single item only.

Test–retest reliability was 0.96 using an intra-class correlation coefficient. On average, the recontact time was 22 days (range 11–44 days). The average time for administration of the SF-20 was 5–7 min. Completion rates for all questionnaires were over 95%.

Discussion

Several concerns arise in interpreting this work. Two potential sources of bias operate. The first is survivor bias: although the sample was reasonably representative in 1991, the least healthy were selected out through death or institutionalization. The second is response bias, which also tends to select more healthy

Table 2. Spearman correlation coefficients (and two-tailed significance) of several measurement instruments used to validate the subscales of the Medical Outcomes Study SF-20 ($n = 333$)

<table>
<thead>
<tr>
<th>Instrument/domin</th>
<th>SF-20 subscale</th>
<th>Barthel index</th>
<th>OARS</th>
<th>OARS-IADL</th>
<th>3MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-20 subscale</td>
<td>SF-20 subscale</td>
<td>SF-20 subscale</td>
<td>SF-20 subscale</td>
<td>SF-20 subscale</td>
<td>SF-20 subscale</td>
</tr>
<tr>
<td>Physical function</td>
<td>Physical function</td>
<td>Physical function</td>
<td>Physical function</td>
<td>Physical function</td>
<td>Physical function</td>
</tr>
<tr>
<td>Role function</td>
<td>Role function</td>
<td>Role function</td>
<td>Role function</td>
<td>Role function</td>
<td>Role function</td>
</tr>
<tr>
<td>Spitzer QLI</td>
<td>Spitzer QLI</td>
<td>Spitzer QLI</td>
<td>Spitzer QLI</td>
<td>Spitzer QLI</td>
<td>Spitzer QLI</td>
</tr>
<tr>
<td>Barthel index</td>
<td>Barthel index</td>
<td>Barthel index</td>
<td>Barthel index</td>
<td>Barthel index</td>
<td>Barthel index</td>
</tr>
<tr>
<td>OARS</td>
<td>OARS</td>
<td>OARS</td>
<td>OARS</td>
<td>OARS</td>
<td>OARS</td>
</tr>
<tr>
<td>OARS-IADL</td>
<td>OARS-IADL</td>
<td>OARS-IADL</td>
<td>OARS-IADL</td>
<td>OARS-IADL</td>
<td>OARS-IADL</td>
</tr>
<tr>
<td>3MS</td>
<td>3MS</td>
<td>3MS</td>
<td>3MS</td>
<td>3MS</td>
<td>3MS</td>
</tr>
</tbody>
</table>

QLI, quality of life index; OARS, Older Americans’ Resources and Services; IADL, instrumental activities of daily living; 3MS, Modified Mini-Mental State Examination.

Table 3. Estimates of internal reliability of the subscales of the Medical Outcomes Study SF-20

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health perception</td>
<td>0.85</td>
</tr>
<tr>
<td>Physical function</td>
<td>0.83</td>
</tr>
<tr>
<td>Scale 5</td>
<td>0.76</td>
</tr>
<tr>
<td>Scale 4</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 4. Mean contrasts of the subscales of the Medical Outcomes Study SF-20 by gender

<table>
<thead>
<tr>
<th>Score</th>
<th>$F$-value (significance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodily pain</td>
<td>72.0 61.7 6.05 (0.014)</td>
</tr>
<tr>
<td>Health perception</td>
<td>70.9 70.4 0.034 (0.854)</td>
</tr>
<tr>
<td>Mental health</td>
<td>84.5 79.2 9.12 (0.003)</td>
</tr>
<tr>
<td>Physical function</td>
<td>71.9 64.5 6.35 (0.012)</td>
</tr>
<tr>
<td>Role function</td>
<td>65.5 60.8 5.28 (0.071)</td>
</tr>
<tr>
<td>Social function</td>
<td>81.7 80.7 0.08 (0.774)</td>
</tr>
</tbody>
</table>
Use of SF-20 in a community-dwelling population

people [24]. The sample may therefore consist of an unusually homogeneous group of relatively healthy elderly people. Unfortunately, 16% of those agreeing and eligible were judged too cognitively impaired to complete the questionnaire. This raises questions about the generalizability of our results to elderly subjects with cognitive impairment. Moreover, this study does not give any further information about the construct validity of four of the six subscales (social functioning, health perceptions, mental health, and bodily pain).

In both well and chronically ill people, the SF-20 takes 3–4 min to administer [25, 26]. In this elderly population, however, it took between 5 and 7 min, suggesting that age and frailty may affect the time to administer the questionnaire.

Construct validity of the SF-20 is supported by reasonable correlation coefficients found with the Barthel Index and the OARS. There was evidence of discriminant construct validity as correlation with the 3MS was poor.

Carmines and Zeller suggest that frequently used scales should have α values over 0.80 [27]. Our results are consistent with these accepted values and compare favourably with those reported with the SF-36 [7].

Measurement instruments designed for screening or categorization purposes may not be responsive [28]. As the ultimate goal of most clinical interventions is to induce change in patients’ status, sensitivity to change is an important property. Unfortunately, there is limited information about the responsiveness of many health status measures, including the SF-20 [29]. While we did not test responsiveness, the ceiling and floor effects raise concerns about the sensitivity to change of the SF-20. These important floor and ceiling effects have previously been reported [30].

The SF-20 remains subject to methodological concerns that have been raised about the nature of other quality of life measures. The SF-20 allows for the calculation of a global rating, but does not allow for an individual to rate the severity or importance of problems, nor does it allow for subjects to add items [31–33].

We need valid, reliable and responsive measures that encapsulate the multi-dimensional problems of frail elderly people [19, 23, 28, 29]. This work provides limited support for the SF-20 as a health-related quality of life measure in an elderly community-dwelling population; however, the SF-20 does not measure several important factors that may be relevant to quality of life. There are insufficient data to recommend the routine use of the SF-20 as a measure of the outcome of care in elderly subjects.

Key points

- There is a need for valid, reliable and responsive measures that address health-related quality of life in elderly people. The Medical Outcomes Study Short Form-20 (SF-20) has been proposed for such research.
- There is evidence of construct validity and reliability of the SF-20 in elderly people living at home.
- The SF-20 does not measure several factors that may be relevant to health-related quality of life in an elderly community-dwelling population. The SF-20 shares the methodological shortcomings of other quality of life measures.

Acknowledgements

This project was supported, in part, by the National Health Research and Development Program [grant no. 6606-3954-MC(S)], and by a National Health Scholar Award to K.R.

References

D. J. Carver et al.


Received 15 January 1998; accepted 12 February 1998