Osteoporos Int (2004) 15: 575–582
DOI 10.1007/s00198-003-1580-3

ORIGINAL ARTICLE

Evaluation of Medical Outcomes Study Short Form-36 Taiwan version in assessing elderly patients with hip fracture

Yea-Ing Lotus Shyu · Jui-fen Rachel Lu · Jersey Liang

Received: 9 September 2003 / Accepted: 15 December 2003 / Published online: 5 February 2004

Abstract The Medical Outcomes Study Short Form-36 (SF-36) is a widely used measure of generic health related quality of life. The purpose of this study is to establish the validity and reliability of the SF-36, Taiwan Version, when applied to a sample of elderly patients with hip fracture in Taiwan. Data from two samples were used, the first sample ($n = 87$) from a prospective descriptive study for testing psychometric scaling assumptions, scale responsiveness and criterion validity, and the second sample ($n = 69$) from a clinical trial for examining the validity of the differences in the group. The SF-36 Taiwan version demonstrated good evidence of supporting the scaling assumption. Cronbach's alpha coefficients above 0.70 for all scales support the internal consistency. The Physical Function (PF) scale had an effect size of 0.88 from months 1 to 3, and 0.59 from months 3 to 6 after discharge, which appears to have the best responsiveness to clinical changes. Notable floor and ceiling effects ($> 15\%$) for Role Emotion (RE), Role-Physical (RP) and PF scales were found. High correlation of 0.62 between the PF and measures of activities of daily living (ADLs), and between RP and instrumental activities of daily living (IADLs) (0.63) supports the construct validity. Significantly higher performance in most SF-36 scales in elders without risk for depression than those who were at risk supported the validity of the group differences. In its current form, the SF-36 Taiwan version demonstrated good reliability and validity as applied to patients with hip fracture.

Keywords Elderly patients · Health status · Hip fracture · SF-36 health survey

Y.-I.L. Shyu (✉)
Center for Gerontological Research and School of Nursing,
Chang Gung University, 259 Wen-Hua 1st Road,
Kwei-Shan, Tao-Yuan 333, Taiwan
E-mail: yeaing@mail.cgu.edu.tw
Tel.: +886-3-2118800 ext. 5275
Fax: +886-3-2118400

J.-R. Lu
Center for Health Industry Management and Public Policy,
Department of Health Care Management,
Chang Gung University, 259 Wen-Hua 1st Road,
Kwei-Shan, Tao-Yuan 333, Taiwan

J. Liang
Health Management and Policy, School of Public Health,
University of Michigan, M3234 School of Public Health II,
Ann Arbor, MI 48109-2029, USA

Introduction

Hip fracture is one of the consequences of low bone density. The high incidence of hip fracture, and the significant morbidity and mortality that occur following hip fracture, make it a major worldwide health care problem, no less so in Taiwan [1,2,3]. As a result, clinicians and researchers in Taiwan have need of accurate and reproducible assessments of outcome following these fractures in order to assess the impact of the injury and the results of interventions. The clinical assessment instrument must be practical, comprehensive, being able to capture the changes caused by course of diseases or interventions while remaining reliable and valid for elders with hip fracture in Taiwan. It should also contribute pertinent information without having any unnecessary features and can provide comparisons to different countries to enable further collaborative development of intervention strategies [4,5].

Health-related quality of life (HRQoL) has been used as a measure, to supplement mortality and objective clinical parameters, in assessing the effects of illness and the outcomes of treatment [6,7]. The SF-36 is a widely used measure of generic HRQoL, including eight health concepts: physical functioning; role limitations—physical; bodily pain; vitality; general health; social functioning; role limitations—emotional; and mental health [8,9]. The SF-36 has also been used in Western countries to measure the impact of hip fracture [10,11,12,13,14]. However, the reliability and validity of the SF-36 for Asian and Chinese elders with hip fracture was unknown.

Because psychometric properties are sample dependent, the performance of a measure in a specific application is especially important [15]. The purpose of this study is, therefore, to describe the validity and psychometric properties of the SF-36 Taiwan Version when applied to an elderly hip fracture patient sample during the first 6-month recovery period after hospital discharge. Previous studies have found that the SF-36 results differ significantly between patients with and without depression [16]. Thus, the group differences on each scale of the SF-36 between hip fractured elders who were likely and those who were not depressed were also explored to examine the construct validity of the SF-36.

Materials and methods

Study setting and sample

Data from two samples were compiled for this study. The first sample ($n = 87$) is from a prospective descriptive study on the outcomes of elderly hip fracture patients in Taiwan. The 87 subjects participating in this study completed the SF-36 at months 1, 3 and 6 after hospital discharge. This sample was used for testing the psychometric scaling assumptions, scale responsiveness and criterion validity. The second sample ($n = 69$) was from a clinical trial on a care model for elderly hip fracture patients in Taiwan. The validity of group differences was examined in the second sample. Although this sample was drawn from a clinical trial, depression management was not a part of the experimenting interventions, thus the percentage of subjects at risk for depression was comparable in the experimental and control groups. In addition, gender, age, marital status, education background and SF-36 performance were comparable between the experimental and control group. Therefore, the experimental and control groups were combined for analysis purposes. Eighty subjects, who completed the SF-36 at the end of the first month after discharge from hospital, were used for this study. The differences of SF-36 scores for two groups, having a chance of having depression or not, were explored in the second sample.

Both samples were collected from the trauma wards of a 3800-bed teaching medical center in northern Taiwan. Sample inclusion criteria were as follows: 1) to expand the potential subjects, age 60 years instead of 65 years or older; 2) hospitalized due to hip fracture caused by accident and received operation for either internal fixation or arthroplasty; 3) living in the geographical area of northern Taiwan; 4) having been assessed by a physician as being alert and without cognitive impairment; and 5) being able to respond to SF-36 items.

Procedures

Human Subjects Approval was obtained from the medical centers for the studies, after which research assistants identified potential subjects. These patients were invited to participate in the study after they had received surgery but before being discharged from the hospital. Because a large percentage of the subjects in both samples were illiterate, interviews were used to collect the data. For the first sample, the elderly participants, after being discharged from hospital, were interviewed face-to-face at outpatient clinics at the end of months 1 and 3 and received telephone interviews at the end of month 6.

Instruments

Several existing instruments, including specific clinical measurements for patients with hip fracture and for self-care capability, were used to examine the criterion validity of the SF-36 items. The SF-36 scores were predicted to be significantly associated with the criterion measures for subscales. Training programs for the interviewers and regular research team meetings were set up to control the quality of the data obtained. Interviewers received the interviewer training in two sessions for both samples. The first training session covered forms to be used in the interviews and an interview demonstration. The second training session consisted of actual interview practice. Any problems arising during the data collection were discussed and resolved on a timely basis.

The SF-36 Taiwan version

The SF-36 Taiwan version was translated, back translated, and judged on similar meaning and demonstrated good reliability and validity in a healthy adult sample [17,18]. The SF-36 Taiwan version is identical to the original SF-36. It measures concepts of physical functioning (10 items), role limitations due to physical health problems (4 items), bodily pain (BP, 2 items), general health (GH, 5 items), vitality (VT, 4 items), social functioning (SF, 2 items), role limitations due to emotional problems (3 items) and mental health (MH, 5 items). For each scale, reverse items were recoded, simple algebraic sums were computed then the raw scale scores were transformed into a scale of 0–100. The higher the score, the better the implied health related quality of life.

Self-care ability

Measures of the ADLs and IADLs were used to assess the self–care ability of the subjects. The Chinese Barthel Index (CBI) was used to measure the ADLs and Lawton's IADL to measure the self–care ability of the subjects. The CBI, translated by Chen, Dai, Yang, Wang and Teng [19], has been established as reliable, valid and appropriate for assessing frail elders in Taiwan. Self-care ability measurements at month 3 after hospital discharge were used to examine the construct validity for the first sample. The Cronbach's alpha of the CBI in this sample was 0.84.

Clinical outcomes

The instruments used by Shih and his colleagues [20], to measure the clinical outcomes of patients treated with total hip arthroplasty, were selected to assess the clinical outcomes of our subjects in order to test the criterion validity of the SF-36 Taiwan version. Pain, range of motion (ROM), and walking ability were each assessed once for each subject and recorded by the attending surgeon during the clinical visit at the end of the third month after discharge. The score for each item went from 1 to 3, with 3 representing better health (less pain, more ROM, better walking ability) and 1 representing worse health (more pain, less ROM, worse walking ability).

Chinese version of Geriatric Depression Scale (GDS) short form

The GDS short form is used to screen and assess the severity of symptoms of depression in the elderly. The total score on the GDS short form is 15; the higher the score the more severe the depressive symptoms. A score of ≥5 on the GDS is considered to be indicative of depression [21]. The reliability (internal consistency) and the construct validity of the GDS short form have been established among the Taiwanese elderly [22]. Elderly hip fracture patients with a GDS short form score equal to or greater than 5 were classified as likely being depressed, and with a score of less than 5 as not likely being depressed.

Data analysis

Formal psychometric tests, of the assumptions underlying the item scoring and construction of multi-item scales, at months 1, 3 and 6

after discharge were conducted. Mean, standard deviation, item-scale correlation, Cronbach's alpha and inter-scale correlations at each time point were used to examine whether the scores satisfied the scaling assumptions of the original SF-36 [23]. The responsiveness, which refers to the ability of an instrument to measure clinical changes [24,25], was examined. Effect size and standardized response mean were used to quantify the responsiveness of the SF-36 to clinical changes. Effect size is calculated as the difference between the mean 1st and 3rd month score divided by the standard deviation of 1st month score; and the difference between the mean 3rd and 6th month score divided by the standard deviation of 3rd month score. Effect sizes of 0.2, 0.5 and 0.8 are typically considered as a small, median and large change, respectively [26]. The standardized response mean, which was defined by the mean change score divided by the standard deviation of the change score of the SF-36, was also used to examine the responsiveness of the SF-36 to clinical changes [27]. The relationship between the SF-36 and criterion measures including the surgeon's clinical assessment of pain, ROM and walking ability, and well established measures of ADLs and IADLs was explored to examine its criterion validity. We expected that physical related SF-36 scales would have stronger associations with these clinical and physical functioning criterion measures. The data were normally distributed for each variable; therefore, the Pearson product moment correlation of the SF-36 with the CBI, Lawton's IADL and the surgeon's rating of pain, ROM and walking ability, was examined [28]. In addition to the relationship between SF-36 and clinical and physical functioning measures, we wished to examine the relationship of SF-36 and mental/emotional condition. Thus, the group differences on each scale of the SF-36, at month 1 following discharge, between elders who were or were not likely depressed were examined for the second sample. We expected that the group of likely depressed patients would perform worse on all subscales. Due to data were not normally distributed for both groups, the Mann-Whitney *U*-test was used to examine group differences. All of the data were coded and entered into a computer and analyzed by SPSS Windows 10.0.

Results

Characteristics of the two samples are listed in Table 1. The demographic variables and SF-36 scores were comparable between the samples. For the first sample ($n=87$), information on SF-36 as well as the clinical outcomes of pain, ROM, and walking ability, the CBI and the Lawton IADL. Data were collected via face-to-face interviews from all 87 subjects at the end of month 1, from 80 subjects at the end of month 3 and via telephone interviews from 72 subjects at the end of month 6 after discharge from hospital. Subject losses from months 1 to 3 were due to mortality ($n=3$) and refusal to participate ($n=4$). Further subject losses from months 3 to 6 were primarily due to refusal ($n=6$), and mortality ($n=2$). The second sample ($n=69$) was interviewed face-to-face 1 month after hospital discharge using both the SF-36 and the GDS.

Table 1 Characteristics of two samples

Characteristic	Sample 1 ($n=87$)	Sample 2 ($n=69$)
Gender		
Male	31 (35.6%)	20 (29%)
Female	56 (64.4%)	49 (71%)
Age (mean ± SD)	79.8 ± 7.2	78.51 ± 8.27
Types of surgery		
Internal fixation	53 (60.9%)	40 (58%)
Arthroplasty	34 (39.1%)	29 (42%)
Educational background		
Illiterate	63 (72.4%)	42 (60.9%)
Primary school	12 (13.8%)	12 (17.4%)
High school	10 (11.5%)	10 (14.5%)
College or above	2 (2.3%)	5 (7.2%)

Psychometric analysis of the scaling assumptions

Psychometric analysis of the scaling assumptions was carried out in sample 1 on data collected at months 1, 3 and 6 after discharge. The percentage of missing data, response option frequency distribution, the mean, standard deviation and skewness of each item were grouped under each scale for the first month following discharge (Table 2). The missing-value rates ranged from 0% to 3.4% ($n=87$). Item response-option frequency distributions were relatively symmetrical for six of the eight scales (BP, GH, VT, SF, RE and MH). Both the PF and the RP scale were positively skewed (worse health). The items within each scale had similar standard deviations, which supports the scaling assumptions on the equal item variance when measuring the same concept. After deleting the patients who had been lost through death or for other reasons, the missing data in most items ranged from 1.3% to 6.3% ($n=80$) at the end of month 3; and from 4.2% to 11.1% ($n=72$) at month 6 after discharge. At the same time, the skewness of items in PF and RE remained, but decreased as the length of time after discharge increased; and RE was negatively skewed (better health) at the end of months 3 and 6 after discharge.

The Pearson item-scale correlation between each item and scale of the SF-36 for the data at month 1 following discharge is presented in Table 3. The correlation between an item and its postulated scale was mostly 0.7 or above with PF2 and MH1 close to 0.6, and PF1 being 0.36. This result indicates that most items met the scaling assumption of internal consistency. The item discriminant validity was demonstrated by item-own scale correlation being higher than item-other scale correlation for all items. Scale correlations within the same scale were generally similar except for item PF1. This result supported the assumption that most items in a given scale contain approximately the same proportion of information about a concept. A similar pattern was found for data at the end of months 3 and 6 after discharge, except that the item-own correlation of PF1 at both the 3rd and 6th month stages, were above 0.50. Item discriminant validity also succeeded in all items at both the 3rd and the 6th month following discharge. Scale correlations within the same scale were also generally similar.

Table 4 shows that the Cronbach's alpha coefficients of internal reliability were above or close to 0.8 for all subscales. The Cronbach's alpha coefficients were all above 0.75 for months 3 and 6 after discharge except

Table 2 The percentage of missing data, response option frequency distribution, the mean, standard deviation and skewness of each item. *PF* physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health

		Missing		Response option frequency Distribution (%)						Item score		
Scale	Item	*n*	%	1	2	3	4	5	6	Mean	SD	Skew
PF	3a	0	0	96.6	3.4	0	–	–	–	1.03	0.18	5.19
	3b	0	0	95.4	4.6	0	–	–	–	1.04	0.21	4.41
	3c	0	0	93.1	4.6	2.3	–	–	–	1.09	0.36	4.22
	3d	0	0	90.8	6.9	2.3	–	–	–	1.11	0.38	3.59
	3e	0	0	81.6	10.3	8.0	–	–	–	1.26	0.59	2.14
	3f	0	0	93.1	5.7	1.1	–	–	–	1.08	0.31	4.22
	3g	0	0	88.5	10.3	1.1	–	–	–	1.12	0.36	2.98
	3h	0	0	70.1	25.3	4.6	–	–	–	1.34	0.56	1.42
	3i	0	0	48.3	32.2	19.5	–	–	–	1.71	0.77	0.55
	3j	0	0	70.1	21.8	8.0	–	–	–	1.37	0.63	1.45
RP	4a	0	0	94.3	5.7	–	–	–	–	1.05	0.23	3.87
	4b	0	0	89.7	10.3	–	–	–	–	1.10	0.30	2.65
	4c	0	0	94.3	5.7	–	–	–	–	1.05	0.23	3.87
	4d	0	0	96.6	3.4	–	–	–	–	1.03	0.18	5.19
BP	7	0	0	1.1	14.9	25.3	29.9	18.4	10.3	3.99	1.26	0.001
	8	0	0	1.1	13.8	10.3	50.6	14.9	9.2	3.91	1.13	−0.23
GH	1	0	0	18.4	35.6	25.3	19.1	1.1	–	2.67	1.21	0.16
	11a	0	0	4.6	10.3	18.4	42.5	24.1	–	3.71	1.08	−0.78
	11b	0	0	2.3	11.5	31.0	42.5	12.6	–	3.51	0.93	−0.44
	11c	0	0	9.2	26.4	24.1	23.0	17.2	–	3.12	1.24	0.01
	11d	0	0	6.9	18.4	19.5	35.6	19.5	–	3.42	1.19	−0.42
VT	9a	1	1.1	4.6	57.5	10.3	5.7	14.9	5.7	2.86	1.39	1.02
	9e	1	1.1	6.9	51.7	19.5	4.6	13.8	2.3	2.73	1.25	1.04
	9g	1	1.1	4.6	26.4	11.5	14.9	28.7	12.6	3.75	1.54	−0.13
	9i	0	0	6.9	20.7	10.3	23.0	23.0	16.1	3.82	1.55	−0.23
SF	6	0	0	8.0	31.0	18.4	27.6	14.9	–	3.10	1.22	0.02
	10	0	0	6.9	29.9	16.1	39.1	8.0	–	3.11	1.13	−0.18
RE	5a	0	0	54.0	46.0	–	–	–	–	1.45	0.50	0.16
	5b	0	0	52.9	47.1	–	–	–	–	1.47	0.50	0.11
	5c	0	0	67.8	32.2	–	–	–	–	1.32	0.46	0.77
MH	9b	1	1.1	5.7	11.5	5.7	14.9	46.0	14.9	4.30	1.43	−0.97
	9c	3	3.4	4.6	10.3	10.3	31.0	33.3	6.9	4.02	1.27	−0.73
	9d	0	0	0	23.0	33.0	17.2	21.8	4.6	3.51	1.19	0.35
	9f	2	2.3	6.9	19.5	17.2	18.4	31.0	4.6	3.62	1.41	−0.25
	9h	2	2.3	3.4	37.9	24.1	10.3	18.4	3.4	3.12	1.30	0.57

that at the end of month 3, SF was at 0.68. Inter-scale correlations were less than the scale internal reliability coefficient for all scales, consistently, from months 1 to 6 after discharge. This result indicated that all scales were able to meet the standard of 0.70 consistently during the first 6 months after discharge and demonstrated the success of measuring a unique concept during the recovery period for a hip fracture.

The ranges of scores for the 8 SF-36 scales demonstrate good variability. The skewness of most of the scales except for RP and PF was within the recommended range of –1 to +1 at month 1 after discharge. Scores for both RP (3.70) and PF (2.35) were positively skewed, indicating that respondents had relatively poor physical functioning and higher role limitations due to physical problems at month 1 after discharge. At months 3 and 6 after discharge, the scores of RE (month 3: −1.13; month 6: −1.65) were negatively skewed, indicating better health resulting from relatively lower role limitation due to emotional problems. The scores of RP (month 3: 1.58; month 6: 1.79) were positively skewed, indicating relatively poorer health from higher role limitation due to physical problems. The skewness of other scales at months 3 and 6 after discharge was within the recommended range of –1 to +1. Notable floor effects greater than 15% for RE (floor effect 1, 3 = 50.6%, 23.8%), PF (floor effect 1,3 = 46.0%, 22.5%), RP (floor effect 1,3,6 = 88.5%, 71.3%, 65.3%) were found. Notable ceiling effects greater than 15% for RE (ceiling effect 1,3,6 = 29.9%, 68.8%, 72.2%) and BP (ceiling effect 3,6 = 26.3%, 31.9%) were also found. In addition, an acceptable to good inter-rater reliability was found, with intra-class correlation coefficients (ICC) ranging from 0.73 to 0.92 in the first sample.

Responsiveness to clinical changes

Using paired *t*-tests, a significant increase in scores from the first to the third month was found for all measurements except GH, while a significant increase in scores for only PF along with a significant decrease in GH scores was found at months 3–6 after discharge for sample 1. The extent of the changes in the SF-36, the CBI, and Lawton's IADL between months 1 and 3 and between months 3 and 6 after discharge was examined by the effect size and the standardized response means (Table 5). From months 1 to 3 after discharge, most

Table 3 Pearson item-scale correlation between each item and its own scale of the SF-36

Scale	Item	Item scale correlation
Physical functioning (PF)	3a	0.36*
	3b	0.57**
	3c	0.77**
	3d	0.75**
	3e	0.74**
	3f	0.70**
	3g	0.76**
	3h	0.86**
	3i	0.75**
	3j	0.75**
Role physical (RP)	4a	0.89**
	4b	0.86**
	4c	0.83**
	4d	0.86**
Bodily pain (BP)	7	0.95**
	8	0.94**
General health (GH)	1	0.71**
	11a	0.80**
	11b	0.80**
	11c	0.76**
	11d	0.91**
Vitality (VT)	9a	0.84**
	9e	0.85**
	9g	0.82**
	9i	0.77**
Social functioning (SF)	6	0.93**
	10	0.92**
Role emotional (RE)	5a	0.96**
	5b	0.96**
	5c	0.82**
Mental health (MH)	9b	0.60**
	9c	0.79**
	9d	0.77
	9f	0.83**
	9h	0.68**

*$P < 0.01$; **$P < 0.001$

scales (BP, VT, SF, MH and RP) showed a median degree of effect size and standardized response mean. A large degree of effect size and standardized response mean were observed in PF, RE and ADL and a small degree appeared in GH and IADL. From months 3–6 after discharge, the extent of all the changes for all scales was small, except for PF. PF appeared to have a median degree of effect size and standardized response mean from months 3–6 after discharge.

Criterion validity

The relationship between the SF-36 and the clinical measures of ROM, walking ability, and pain, and between the SF-36 and self-care ability measurements of ADL and IADL at month 3 after discharge for sample 1 was explored to examine its criterion validity. Table 6 lists the Pearson correlation between each subscale and each clinical and self-care measurement for month 3 after hospital discharge. Physical scales (GH, VT, PF, RP) were moderately related to ADL, IADL, ROM, and walking ability. A stronger correlation (0.50 or above) was noted between PF and ADL, IADL, ROM and walking ability, and between RP and IADL (0.63). BP had a stronger correlation to clinical pain assessment ($P = 0.56$) than any of the other scales. SF, MH and RE had an insignificant or mild correlation to ADL, IADL, ROM and walking ability.

Table 4 Cronbach's alpha and inter-scale correlations of each SF-36 scale. *PF*physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health; *ADL* activities of daily living; *IADL* instrumental activities of daily living. Bolded values are Cronbach's alpha. Non-bolded values are inter-scale correlations

	SF36 scale							
Scale	PF	RP	BP	GH	VT	SF	RE	MH
PF	**0.87**	–	–	–	–	–	–	–
RP	0.31**	**0.87**	–	–	–	–	–	–
BP	0.31**	0.20	**0.88**	–	–	–	–	–
GH	0.30**	−0.02	0.13	**0.85**	–	–	–	–
VT	0.34**	−0.04	0.13	0.58***	**0.83**	–	–	–
SF	0.19	0.14	0.33**	0.15	0.12	**0.82**	–	–
RE	0.41***	0.37***	0.36**	0.14	0.27	0.27*	**0.90**	–
MH	0.34**	0.08	0.08	0.39***	0.61***	0.17	0.54***	**0.80**
ADL	0.56***	0.17	0.29**	0.36**	0.39***	0.14	0.34**	0.26*
IADL	0.56***	0.28**	0.16	0.14	0.22*	0.11	0.32**	0.26*

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$

Construct validity

In the second sample, 27 subjects scored 5 or above in GDS and were classified as the group who were likely depressed. Forty-two subjects scored less than 5 and were classified as the group who were not likely depressed. Construct validity was demonstrated by consistently better health (a higher score) across all SF-36 scales for the group who were not likely depressed than for the group who were likely depressed (Fig. 1). By using Mann-Whitney *U*-tests, patients with hip fracture, not being likely depressed, appeared to have significantly better health according to most of the SF-36 scales, except for RE and RP, than those likely depressed patients.

Discussion

In this study, the sample of elderly patients who had hip fracture was taken from a teaching hospital and is not necessarily representative of patients suffering from hip fracture in the general population. However, the similarity between the profile of this sample and samples used in other studies of patients with hip fracture in Taiwan [29] may lessen the influence of the convenience sample effect on the generalizability of the results. At the same time, the psychometric characteristics of the SF-36 Taiwan version including mean scores, standard deviations, internal consistency and inter-scale correlations appeared to be stable over different points of time. This

Table 5 Changes in score, effect size and standard response means of the SF-36 scale, ADL and IADL between months 1 and 3 and between months 3 and 6. *PF* physical functioning; *RP* role physical; *BP* bodily pain; *GH* general health; *VT* vitality; *SF* social functioning; *RE* role emotional; *MH* mental health; *ADL* activities of daily living; *IADL* instrumental activities of daily living

	Months 1–3			Months 3–6		
	Change score	Effect size	SD response mean	Change score	Effect size	SD response mean
PF	14.62 ± 18.39	0.88	0.79	12.21 ± 23.26	0.59	0.52
RP	13.14 ± 37.95	0.59	0.34	−2.17 ± 37.31	−0.04	−0.05
BP	15.44 ± 24.30	0.66	0.63	−1.52 ± 30.76	−0.07	−0.04
GH	3.67 ± 25.84	0.17	0.14	−4.30 ± 23.21	−0.15	−0.18
VT	11.92 ± 26.37	0.50	0.45	2.34 ± 22.41	0.14	0.10
SF	11.07 ± 29.21	0.38	0.37	−1.25 ± 29.65	−0.05	−0.04
RE	34.61 ± 49.07	0.72	0.70	2.41 ± 40.55	0.13	0.05
MH	10.71 ± 23.40	0.51	0.45	2.24 ± 18.43	0.14	0.12
ADL	10.27 ± 13.50	0.48	0.76	1.62 ± 16.67	0.12	0.09
IADL	0.29 ± 1.27	0.20	0.22	0.25 ± 1.42	0.20	0.17

reason increases our confidence in the validity and reliability of the findings.

The SF-36 survey was designed for self-administration or telephone and interviewer administration. A telephone survey is suggested when a target sample consists of the elderly or people with less than 9 years of education [15]. Due to the large percentage of illiterate participants in our sample, face-to-face interviews were conducted in the clinic at the end of months 1 and 3 after discharge from hospital and telephone interviews were conducted at the end of month 6 after discharge. In comparison to the study of McHorney and colleagues [15], the completeness of data collection for the SF-36 items in this sample is good for the face-to-face interviews at the end of months 1 and 3. The completion rate decreased a little, however, for the telephone interviews at the end of month 6.

The evaluation of scaling assumptions and score reliability are very important as applied to a particular sample [15]. In this study, the SF-36 Taiwan version demonstrated good evidence to support the scaling assumptions when administered to elderly patients with hip fracture. Equal item variance in measuring the same concept was supported by a similar standard deviation of items within each scale. With very few exceptions, discriminant validity of the items was supported by a higher correlation between the item and its postulated scale, than with scales measuring other concepts. Internal consistency was supported by Cronbach's alpha coefficients above 0.70 for all scales at all time points [30].

Substantial floor effects (percent sample at lowest score) for PF, RE, and RP were found at the end of the 1st month after discharge. The floor effect of RP remained substantial for measurements at the end of months 3 and 6, but the floor effect of PF and RE decreased to modest at the end of month 3 and to trivial at the end of month 6 after discharge. As supported by a previous study [15], the floor effect was most common among the most seriously ill patients who were in the worst health during the 1st month after leaving hospital, but later improved. In general, ceiling and floor effects are rare for the three bipolar scales (GH, VT and MH), but often occur in the two role disability scales, while the ceiling effect for BP is modest [15]. The ceiling effect of RE, even from early in the 1st month after discharge,

Table 6 Pearson correlation between each subscale and each clinical and self-care measure for month 3 after hospital discharge. *ADL* activities of daily living; *IADL* instrumental activities of daily living; *ROM* range of motion

	SF36 scale							
Scale	PF	RP	BP	GH	VT	SF	RE	MH
ADL	0.62^{***}	0.41^{***}	0.07	0.44^{***}	0.38^{**}	0.10	0.22	0.23^{*}
IADL	0.59^{***}	0.63^{***}	0.09	0.40^{***}	0.26^{*}	0.19	0.22	0.15
Pain	0.28^{*}	0.22^{*}	0.56^{***}	0.17	0.08	0.38^{**}	0.26^{*}	0.24^{*}
ROM	0.50^{***}	0.32^{**}	0.07	0.37^{**}	0.45^{***}	0.13	0.23^{*}	0.33^{**}
Walk	0.58^{***}	0.34^{**}	0.09	0.38^{**}	0.47^{***}	0.23^{*}	0.27^{*}	0.32^{**}

$^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$

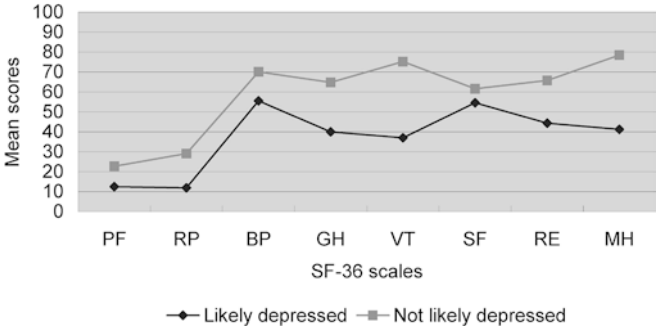

Fig. 1 SF-36 scales for elderly patients with hip fracture were likely to be depressed and for those who were not likely to be depressed

indicates that regardless of any clinical improvement in role limitation due to emotional problems, RE cannot capture the score changes in a certain percentage of individuals with hip fracture. Similarly, the floor effect of RE, PF, and RP soon after discharge, indicating a deterioration of physical function, physical role and emotional role, could not be detected for a subgroup of patients. Therefore, while using the SF-36 Taiwan version in clinical studies, the impact of hip fracture and the intervention effects may be underestimated and require cautious interpretation.

The responsiveness of the SF-36 Taiwan version to clinical changes appeared to be higher for changes between months 1 and 3 than between months 3 and 6. This finding is supported by a previous study that

showed most of the recovery occurring within the first 3 months after discharge [31,32]. Similar effects in PF and ADL were also found in a previous study [33]. A higher responsiveness was shown with PF than with ADL for measuring changes in elderly patients with hip fracture later in the recovery period (from months 3 to 6). The PF in SF-36 Taiwan version seemed to be better than CBI in capturing changes during the later recovery period after receiving surgery for hip fracture for elderly patients in Taiwan.

Criterion validity of the SF-36 for elderly hip fracture patients in Taiwan is supported by the relationships between the SF-36 Taiwan version and the clinical and self-care ability measures. The physical function related scales (PF, RP, VT, GH), had higher correlations than the mental function related scales (SF, RE, MH), with the existing clinical measurements of ROM, walking ability and ADL/IADL which supports the criterion validity of the SF-36 Taiwan version. The strong correlations found between PF and the clinical functional measures were supported by Jaglal and colleagues [33]. Instrumental behaviors were viewed as indicators of role function [34] and support the validity of the RP with the strongest correlation to IADL. This finding also indicates that the SF-36 Taiwan version provides information that cannot be obtained by standard clinical and functional scales. Supplementary information includes the effects of physical problems and pain on normal social and work activities, reduced levels of energy, self-perceived health, limitations in community involvement and normal activities due to emotional problems and perceived mental functioning. Construct validity was supported by the consistently poorer performance of HRQoL for subjects who had a chance of having depression than those who did not. Similar results were found in a sample of stroke survivors in Australia [16].

In summary, these findings provide a preliminary base for the use of the SF-36 Taiwan version in elderly people with hip fracture. The small convenience sample limits the generalizability of the results. A larger sample using structural equation modeling to further explore the underlying concepts in relation to related concepts is suggested for the future studies. In its current form, the SF-36 Taiwan version demonstrated good reliability and validity as applied to patients with hip fracture in this study. The SF-36 Taiwan version is easily administered to illiterate elderly patients suffering from hip fracture, via face-to-face or telephone interviews. Interviewer training does not take long, the interviews are not burdensome and hence do not wear out the elderly patients. However, although several activity and role performance related scales (PF, RE, RP) appeared to be able to capture changes during the recovery period, they seemed to have a tendency to underestimate the impact of interventions in clinical trials and hence, must be given careful consideration.

Acknowledgments This work was supported by the Chang Gung Memorial Hospital (CMRP819) and the National Health Research Institute, Republic of China. The authors would also like to thank research assistants Grace Wu, Shiu-shin Tsai, Shu-Chuan Chou, and Hsiao-Chin Lee for their help in data collection.

References

1. Cumming RG, Klineberg RJ (1994) Fall frequency and characteristics and the risk of hip fracture. J Am Geriatr Soc 42:774–778
2. Aharonoff GB, Koval KJ, Skovron ML, Zuckerman JD (1997) Hip fractures in the elderly predictors of one-year mortality. J Orthop Trauma 11:162–165
3. Tsai YJ, Lin YS, Chen HC et al. (1998) (in Chinese) Conditions of elderly patients with hip fracture in Taiwan]. Taiwan Provincial Institute of Family Planning, Taichung, Taiwan
4. Ware JE, Brook RH, Davies AR (1981) Choosing measures of health status for individuals in general populations. Am J Publ Health 71:620–625
5. Hays RD, Anderson R, Revicki D (1993) Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 2:441–249
6. Greenfield S, Nelson EC (1992) Recent developments and future issues in the use of health status assessment measures in clinical settings. Med Care:MS23–MS41
7. Wilson IB, Cleary PD (1995) Linking clinical variables with health relater quality of life. JAMA 273:59–65
8. Weinberger M, Samsa GP, Hanlon JT et al. (1991) An evaluation of a brief health status measure in elderly veterans. J Am Geriatr Soc 39:691–694
9. Lyons RA, Perry HM, Littlepage BNC (1994) Evidence for the validity of the Short Form 36 questionnaire (SF-36) in an elderly population. Age Aging 23:182–184
10. Adachi JD, Loannidis G, Berger C et al. (2001) The influence of osteoporotic fractures on health-related quality of life in community-dwelling men and women across Canada. Osteoporos Int 12:903–908
11. Gabriel SE, Kneeland TS, Melton LJ 3rd, Moncur MM, Ettinger B, Tosteson AN (1999) Health-related quality of life in economic evaluations for osteoporosis: whose values should we use? Med Decision Making 19:141–148
12. Peterson MGE, Allegrante JP, Cornell CN et al. (2002) Measuring recovery after a hip fracture using the SF-36 and Cummings scales. Osteoporos Int 13:296–302
13. Randell AG, Nguyen TV, Bhalerao N, Silverman SL, Sambrook PN, Eisman JA (2000) Deterioration in quality of life following hip fracture: a prospective study. Osteoporos Int 11:460–466
14. Tosteson AN, Gabriel SE, Grove MR, Moncur MM, Kneeland TS, Melton LJ 3rd (2001) Impact of hip and vertebral fractures on quality-adjusted life years. Osteoporos Int 12:1042–1049
15. McHorney CA, Ware JE Jr, Lu JFR, Sherbourne CD (1994) The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 32:40–66
16. Anderson C, Laubscher S, Burns R. (1996) Validation of the Short Form 36 (SF-36) health survey questionnaire among stroke patients. Stroke 27:1812–1816
17. Lu JFR, Tseng HM, Tsai YJ (2004) Assessment of health–related quality of life in Taiwan (I): development and psychometric testing of SF-36 Taiwan Version. Chin J Public Health (in press)
18. Tseng HM, Lu JFR, Tsai YJ (2004) Assessment of health-related quality of life (II): norming and validation of SF-36 Taiwan version. Chin J Public Health (in press)
19. Chen YJ, Dai YT, Yang CT, Wang TJ, Teng YH (1995) A review and proposal on patient classification in long-term care system. Department of Health, Taipei, Taiwan, Republic of China
20. Shih CH, Lee ZL, Hsu WW (1987) Five years follow-up study of cemented total hip arthroplasty. J Orthop Surg ROC 4:133–138

21. Burke WJ, Roccaforte WH, Wengel SP (1991) The short form of the Geriatric Depression Scale: a comparison with the 30-item form. J Geriatr Psychiatr Neurol 4:173–178
22. Liu CY, Lu CH, Yu S, Yang YY (1998) Correlations between scores on Chinese versions of long and short forms of the Geriatric Depression Scale among elderly Chinese. Psychol Rep 82:211–214
23. Lam CLK, Gandek B, Ren XS, Chan MS (1998) Tests of scaling assumptions and construct validity of the Chinese (HK) version of the SF–36 Health Survey. J Clin Epidemiol 51:1139–1147
24. Guyatt G, Walter S, Norman G (1987) Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 40:171–178
25. Liang MH (1995) Evaluating measurement responsiveness. J Rheumatol 22:1191–1192
26. Kazis LE, Anderson JJ, Meenan RF (1989) Effect sizes for interpreting changes in health status. Med Care 27:S178–189
27. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH (1992) Comparative measurement sensitivity of short and longer health status instruments. Med Care 30:917–925
28. Crocker L, Algina J (1986) Introduction to classical and modern test theory. Harcourt Brace Jovanovich, Orlando, Fla.
29. Tsai YJ, Lin HS, Chen HC, Chou LP, Chang MC, Liu TK (1995) (in Chinese). A pilot study on health sector priority review by retrospective study on the elderly hip fracture in Taiwan. Taiwan Provincial Institute of Family Planning, Taichung, Taiwan
30. Nunnally JC (1994) Psychometric theory, 3rd edn. McGraw Hill, New York
31. Young Y, Brant L, German P et al (1997) A longitudinal examination of functional recovery among older people with subcapital hip fractures. J Am Geriatr Soc 45:288–294
32. Peterson MGE, Allegrante JP, Cornell CN, et al (2002) Measuring recovery after a hip fracture using the SF-36 and Cummings Scales. Osteoporos Int 13:296–302
33. Jaglal S, Lakhani Z, Schatzker J (2000) Reliability, validity, and responsiveness of the lower extremity measure for patients with a hip fracture. J Bone Jt Surg [Am] 82A:955–962
34. Roy SC, Andrews HA (1999) The Roy adaptation model, 2nd edn. Appletion & Lange, Stamford, Conn.