International Journal of Technology Assessment in Health Care, **21:3** (2005), 393–398.

Test–retest reliability of health utilities index scores: Evidence from hip fracture

C. Allyson Jones
University of Alberta and *Institute of Health Economics*

David Feeny
Institute of Health Economics and *University of Alberta* and *Health Utilities Incorporated*

Ken Eng
Institute of Health Economics, Alberta, Canada

Objectives: There is relatively little evidence on the test–retest reliability of utility scores derived from multiattribute measures. The objective was to estimate test–retest reliability for Health Utilities Index Mark 2 (HUI2) and Mark 3 (HUI3) utility scores in patients recovering from hip fracture.
Methods: We enrolled an inception cohort of hip fracture patients within 3 to 5 days of surgery. Baseline assessments included the Functional Independence Measure (FIM™), Folstein Mini-Mental State Examinations, and the HUI2 and HUI3 questionnaire. Follow-up assessments at 1, 3, and 6 months also included a global change question. Test–retest reliability was assessed as agreement between 3- and 6-month scores using the intraclass correlation coefficient (ICC). Two approaches were used to classify patients as stable; a third approach based on the generalizability theory was also used. Patients were classified as stable if their FIM™ overall scores changed by 10 points or fewer and if they classified themselves as having experienced no or only a little change according to their global change question.
Results: Complete data at both the 3- and 6-month assessments based on self-report were available for 196 patients; 141 patients with complete data were classified as stable. The ICCs for HUI2 and HUI3 for stable patients were 0.71 and 0.72; the ICCs derived from the generalizability theory were 0.76 and 0.77.
Conclusions: Test–retest reliability for HUI in this cohort was similar to reliability estimates for other preference-based multiattribute and generic health-profile measures—in the acceptable range for making valid group-level comparisons.

Keywords: Test–retest reliability, Health Utilities Index, HUI, Hip fracture

The study, "Measurement of Health Status and Health-Related Quality of Life in Patients with Hip Fractures" was supported by a grants to Drs. C. Allyson Jones, David Feeny, Finlay McAlister, Cheryl Wiens, and John Cinats from the Institute of Health Economics, University Hospital Foundation, Edmonton Orthopaedic Research Trust, and Royal Alexandra Foundation. The analyses reported in this paper were supported by grants from the Alberta Heritage Foundation for Medical Research (AHFMR; #199909) and the Institute of Health Economics (IHE) to C. Allyson Jones and David Feeny. IHE, AHFMR, the University Hospital Foundation, Edmonton Orthopaedic Research Trust, and Royal Alexandra Foundation played no role in the design, interpretation, or analysis of the project and have not reviewed or approved of this manuscript. Support for the postdoctoral fellowship held by Dr. C. Allyson Jones was provided by the Alberta Heritage Foundation for Medical Research and Canadian Institute of Health Research. An earlier version of the paper was presented as a poster at the 11th Annual Meeting of the International Society for Quality of Life Research, October 16–19, 2004, in Hong Kong. We thank the patients and family caregivers for their participation in the study. We also thank the staff of the University of Alberta Hospital Orthopaedics Research Office for their assistance in patient recruitment and data collection. It should be noted that David Feeny has a proprietary interest in Health Utilities Incorporated, Dundas, Ontario, Canada. HUInc. owns the copyright to and distributes HUI materials.

Although there is evidence on the test–retest reliability of directly measured utilities, there is relatively little published evidence on the test–retest reliability of utility scores derived from multiattribute measures such as the EQ-5D (30), Short-Form 6D (SF-6D; 4), Quality of Well Being scale (QWB; 18), and Health Utilities Index (HUI; 9;10;12;16). A standard design for studies investigating test–retest reliability is to administer the instrument to a group of respondents who are expected not to experience changes in health status and then to readminister the instrument shortly thereafter (24;37). Often the interval between administrations is selected to be long enough that respondents are likely to have forgotten their previous responses but short enough that health status is unlikely to have changed. Alternatively, using the generalizability theory, one can calculate the intraclass correlation coefficient (ICC) between scores at the two administrations by dividing the between-subject variation by the total variation (37). Finally, one can use longitudinal studies to obtain estimates of test–retest reliability by assessing agreement among scores for stable patients. A challenge in this approach is to classify subjects as stable versus nonstable (25).

METHODS

Study Design for Main Study

This investigation of test–retest reliability is part of a larger prospective cohort study examining recovery after hip fracture. Inclusion criteria included age 65 or older, ability to speak English, availability of a friend or family member who could act as a proxy respondent, and residence in the Capital Health region (Edmonton and surrounding area) of Alberta, Canada. Exclusion criteria included a pathological fracture other than one caused by osteoporosis, Paget disease, readmission for a previous fracture, and previous fracture within the past 5 years. Patients were recruited from October 2000 until December 2001. Questionnaires were administered to patients who had a Folstein Mini-Mental State Examination (MMSE) scores of 18 or higher (11;39). Data for patients with MMSE scores >18 were collected from proxy respondents. For this analysis, we report only the results from patients with MMSE scores greater than or equal to 18 at baseline.

The baseline assessment was performed in person by a trained professional interviewer within 3 to 5 days after surgery. Follow-up interviews were conducted by telephone at 1, 3, and 6 months after fracture. Follow-up interviews included the same battery of instruments used at baseline plus a nine-point global change (over the past 1 month) question (extremely worse; a lot worse; somewhat worse; a little worse; no change; a little better; somewhat better; a lot better; extremely better).

Data on clinical and demographic characteristics of patients were collected. Information on comorbidities was obtained in interviews with patients using a list derived from chronic conditions listed in the Charlson Comorbidity Index (5) and the Statistics Canada National Population Health Survey instrument.

Although many patients improved substantially during the first month and continued to improve over the next 2 months, descriptive results for the entire cohort indicated little or no change in the overall health of the cohort between the month 3 and month 6 assessments. We, therefore, based our examination of test–retest reliability on a comparison of the 3- and 6-month scores.

Measures

HUI2. The HUI2 is a multiattribute utility measure that includes a health-status descriptive system and a multiplicative multiattribute utility function that provides overall utility scores for HUI2 health states on the conventional dead $= 0.00$ to perfect health $= 1.00$ scale (9;12). The HUI2 covers seven attributes (dimensions) of health status: sensation (vision, hearing, and speech), mobility, emotion, cognition, self-care, pain, and fertility. (The initial application of HUI2 involved survivors of cancer in childhood for whom low fertility and infertility are issues; the fertility dimension was omitted from the study reported here.) Each attribute has four or five levels, ranging from highly impaired, levels 4 or 5 (for instance, level 5, unable to control or use arms and legs, for mobility), to normal, level 1. The HUI2 focuses on capacity rather than performance. The multiplicative HUI2 scoring function is based on preference scores obtained from a random sample of parents in the general population in Hamilton, Ontario, Canada (40).

HUI3. The attributes included in the HUI3 are vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain (10;12;16). There are five or six levels per attribute in the HUI3. The multiplicative scoring function for the HUI3 is based on preference scores obtained from a random sample of respondents 16 years of age and older in Hamilton, Ontario, Canada (10). Scores range from -0.36 (the all-worst HUI3 state) to 0.00 for dead to 1.00 for perfect health.

Folstein MMSE

The MMSE is a screening instrument for cognitive status (11). Eleven questions cover orientation to time, orientation to place, registration of three words, attention and calculation, recall, language, and visual construction. We defined the severity of cognitive impairment by using three cutoff levels (32;39): no cognitive impairment (24 to 30 points), mild impairment (18 to 23 points), and severe impairment (0 to 17 points; patients in this group at baseline were excluded from the analyses presented here).

Functional Independence Measure™

The Functional Independence Measure™ (FIM™) is a performance-based measure of disability based on the amount of assistance required to perform basic activities of daily living (13). The FIM™ includes eighteen items covering self-care, sphincter control, transfers, locomotion, communication, social adjustment/cooperation, and cognition/problem solving. Scores range from 18 to 126, with a higher score representing greater independence. Several studies provide evidence on the use of the telephone version of the FIM™ and/or its use in patients with hip fracture or elderly patients (13;14;26–29;34;36).

Wallace et al. (41) have suggested that the minimal clinically important difference for FIM™ scores is 11. Conservatively, we chose a slightly more stringent criterion to define stable patients—those experiencing a change of 10 points or fewer.

Criteria for Classifying Patients as Stable

In one approach to assessing test–retest reliability, the usefulness of the assessment of test–retest reliability relies on identifying a "known group" of stable patients. Following the example of Deyo et al. (7), we classified patients as stable if they fulfilled two criteria: (i) a less than clinically important change in the overall FIM™ score (10 or fewer); and (ii) a response on the global change questions of no change, a little worse, somewhat worse, a little better, or somewhat better (± two categories). Given that recall of previous health status is less than ideal (25), we did rely solely on results from the global change question. In a secondary analysis, a more stringent criterion was used: no change, a little worse, or a little better (± one category). In addition, using the generalizability theory (37), we estimated test–retest reliability for all available patients by assessing agreement between HUI scores at 3 and 6 months.

Statistical Analyses

Agreement among measures was assessed using the kappa statistic (19). Kappa values of <0.00 are interpreted as indicating poor agreement, values of 0.00–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–0.1.00 as almost perfect (19).

We assessed test–retest reliability using an ICC (7;33;35) derived from a mixed model two-way analysis of variance in which assessment was the fixed factor and patients were modeled as random. (Estimates were also done using a two-way random effects model in which both time and patient were random.) For the analyses based on the generalizability theory, the ICC was calculated as between-subject variability divided by total variability. Analyses were performed with SPSS Version 12 (SPSS, Inc., Chicago, IL). Using criteria proposed by Juniper et al. (17), ICCs >0.80 were classified as excellent agreement, 0.61 through 0.80 as good agreement, 0.41 through 0.60 as moderate agreement, and ≤ 0.40 as poor to fair. Ethics approval was obtained before data collection and medical chart review from the Health Research Ethics Review Board of the University of Alberta and Capital Health Region.

RESULTS

In total, 383 patients were enrolled in the study. Of these, 265 patients had baseline MMSE scores greater than or equal to 18 and, thus, were potentially available for analysis. Complete data based on patient self-assessment at months 3 and 6 are available for 195 patients. Using the ± two categories criterion from the global change question, 167 (85.6 percent) of these patients were classified as stable and 28 (14.4 percent) as not stable. Using the criterion of a change in overall FIM™ score ≤ 10, 160 (82.1 percent) patients were classified as stable and 35 (17.9 percent) as not stable. The percentage agreement between the two criteria was 77; the kappa statistic (unweighted) was 0.54, indicating moderate agreement (19).

Combining the two criteria, 141 (72.3 percent) patients were classified as stable by both criteria, 9 (4.6 percent) were classified as not stable by both, 19 (9.7 percent) were classified as stable by the FIM™ but not by the global change question, and 26 (13.3 percent) were classified as stable by the global change question but not by the FIM™. The 141 patients classified as stable by both criteria were used to assess test–retest reliability (Tables 1 and 2).

In the secondary analysis using the ± one category criterion for the global change question and change in FIM™ score ≤ 10, 124 (63.6 percent) patients were classified as stable by both criteria; the percentage agreement between the two criteria was 69 and the kappa was 0.37, indicating fair agreement (19).

For the analyses based on the generalizability theory, complete HUI2 scores were available for the 3- and 6-month assessments for 136 patients. For the HUI3, complete data were available for 137 patients.

Several patients skipped one or more items on the HUI questionnaire; frequently skipped questions included those on vision and hearing. As a result, complete data are available at both the 3- and 6-month assessments for only 104 patients for HUI2 (37 missing) and 105 for HUI3 (36 missing). At the 3-month assessment, patients for whom HUI data were incomplete may have been less healthy than those for whom data were complete. This tendency is not apparent at the 6-month assessment. Nonetheless, the mean HUI scores for the patients used in the analysis of test–retest reliability should not be regarded as representative of results for the cohort.

Mean change scores between the 3- and 6-month assessments indicate little change over the period (Table 3). The ICCs for the overall HUI2 and HUI3 scores were 0.71 and

https://doi.org/10.1017/S0266462305050518 Published online by Cambridge University Press

Table 1. Demographic Characteristics of Stable Patients at the 3-Month Assessment

	n	Mean	Median	Minimum	Maximum	Standard deviation
Age (yr)	141	79.7	79.5	65.2	95.2	7.47
Gender, % female	141	72				
Chronic conditions (n)	139	5	5	0		2.48
MMSE	141	20	21	11		2.66

MMSE, Folstein Mini-Mental State Examination scores range from 0 to 30, with higher scores indicating higher cognitive function. Scores less than 18 indicate severe cognitive impairment; scores 18–23 indicate mild cognitive impairment; scores 24 or higher indicate no cognitive impairment (32;39).

Table 2. HUI and FIM™ Scores for Stable Patients at the 3- and 6-Month Assessments

	Mean	Median	Minimum	Maximum	Standard deviation
3-Month assessment					
HUI2 overall, n = 104	0.65	0.67	0.19	0.97	.17
HUI3 overall, n = 105	0.56	0.61	−0.07	1.00	.26
FIM™ overall, n = 105	108	112	61	126	14.29
6-Month assessment					
HUI2 overall, n = 104	0.67	0.67	0.19	0.97	.19
HUI3 overall, n = 105	0.57	0.61	−0.12	1.00	.27
FIM™ overall, n = 105	109	111	58	126	14.92

HUI2, Utilities Index Mark 2; HUI3, Utilities Index Mark 3; FIM™, Functional Independence Measure.

Table 3. Mean Change Scores for Stable Patients between the 3- and 6-Month Assessments

	HUI2 overall	HUI3 overall	FIM™ overall
Mean	0.02	0.01	−0.01
Median	0.00	0.00	0.00
Minimum	−0.48	−0.51	−10.00
Maximum	0.39	0.52	10.00
Standard deviation	0.14	0.20	4.40

HUI2, Utilities Index Mark 2; HUI3, Utilities Index Mark 3; FIM™, Functional Independence Measure.

0.72 (Table 4). Results are virtually identical for the ± two or ± one category criteria from the global change question. Results are also virtually identical when using a two-way random effects model (data not shown). Results based on the generalizability theory are similar but somewhat higher, 0.76 and 0.77. When patients are divided into two groups, not cognitively impaired versus cognitively impaired at 3 and 6 months, the ICCs based on the generalizability theory are 0.79 for HUI2 and 0.77 for HUI3 for the not cognitively impaired compared with 0.65 and 0.67 for the cognitively impaired.

DISCUSSION

The results provide evidence that the test–retest reliability of HUI2 and HUI3 falls into the acceptable level of 0.70 or higher generally recommended as required for group-level comparisons (15;22;31). The ICCs are below the 0.90 level generally recommended for individual-level use of scores. The ICCs for agreement for the cohort are higher, 0.76 and 0.77, and again acceptable for group-level comparisons.

Our reliability estimates are consistent with those from previous studies of the HUI. In an assessment of test–retest reliability conducted as part of a pretest of the Statistics Canada National Population Health Survey using a provisional scoring system for HUI3, Boyle et al. (1) report an ICC of 0.77. Suarez-Almazor et al. (38) reported 3-month and 6-month test–retest reliability ICCs of 0.78 and 0.80 for HUI2 in a cohort of patients with low back pain. In the same study, ICCs for EQ-5D index scores were 0.76 and 0.50. (ICCs for EQ-5D visual analogue scale scores are reported in the studies by Dorman et al. and Macran (8;21).) Our results are also similar to those of Luo et al. (20) for patients with rheumatic disease, who reported test–retest ICCs for EQ-5D and HUI3 of 0.64 and 0.75.

Brazier et al. (2) reported a Spearman correlation of 0.67 between test and retest EQ-5D index scores in a study of elderly female patients in the United Kingdom. They also reported additional evidence for EQ-5D index scores of 0.83 in patients with chronic obstructive pulmonary disease and 0.55 in patients with rheumatoid arthritis. Coons et al. (6) found an ICC of 0.78 in a 2-week test–retest study using EQ-5D index scores. One-day test–restest reliability ICCs for the QWB scale range from 0.78 to 0.99, with most values exceeding 0.90 (3).

https://doi.org/10.1017/S0266462305050518 Published online by Cambridge University Press

Table 4. Intra-Class Correlation Coefficients for Test–Retest Reliability for HUI2 and HUI3

Measure	ICC	95% Confidence interval
Criteria: Change in FIM^{TM} $\leq$10 and $\pm$ two categories on global change question		
HUI2 overall score	0.71	0.60–0.79; n = 104
HUI3 overall score	0.72	0.62–0.80; n = 105
Criteria: Change in FIM^{TM} $\leq$10 and $\pm$ one categories on global change question		
HUI2 overall score	0.71	0.60–0.80; n = 94
HUI3 overall score	0.73	0.62–0.81; n = 94
Criteria: Agreement between 3- and 6-month scores for all patients with complete data		
HUI2 overall score	0.76	0.67–0.82; n = 136
HUI3 overall score	0.77	0.69–0.83; n = 137

HUI2, Utilities Index Mark 2; HUI3, Utilities Index Mark 3; ICC, intra-class correlation; FIM^{TM}, Functional Independence Measure.

It is also important to compare the ICCs for the preference-based multiattribute measures with ICCs for generic health-profile measures. In their review, Coons et al. (6) reported ICCs of 0.60 to 0.81 for various domain scores derived from SF-36, 0.87 to 0.97 for the Sickness Impact Profile, and 0.67 to 0.97 for the Dartmouth COOP Charts. McHorney and Tarlov (23) report ICCs of 0.77 to 0.85 for the Nottingham Health Profile, 0.42 to 0.88 for the Dartmouth COOP Charts, 0.30 to 0.78 for the Duke Health Profile, and 0.60 to 0.81 for the SF-36.

Several study limitations should be noted. First, our results are based on self-assessments of health status and systematically exclude cognitively impaired and very ill patients. Second, because many respondents skipped questions on vision or hearing or other dimensions of health status, there were missing data for the HUI. Clearly, the mean HUI scores are not representative of the entire cohort. The reliability estimates may reflect the experience of somewhat healthier respondents and may not be fully generalizable.

The results reported here for the test–retest reliability of the HUI2 and HUI3 are consistent with other results for the HUI, results for other multiattribute preference measures, and results for generic health profile measures. Test–retest reliability appears to be acceptable for group-level comparisons. Additional empirical evidence on test–retest reliability for multiattribute utility scores in other patient groups would be welcome.

CONTACT INFORMATION

C. Allyson Jones, PT, PhD, Assistant Professor (Allsyson.Jones@ualberta.ca) Department of Physical Therapy, Faculty of Rehabiliations Medicine, 2-50 Corbett Hall, University of Alberta, Edmonton, Alberta T6G 2G4, Canada
David Feeny, PhD (david.feeny@ualberta.ca), Professor of Economics, Public Health Sciences, and Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta T6G 2H4, Canada; Institute of Health Economics, 10405 Jasper Avenue, Suite 1200, Edmonton, Alberta T5J 3N4, Canada
Ken Eng, MA, Research Associate (keng@ihe.ca), Institute of Health Economics, 10405 Jasper Avenue, Suite 1200, Edmonton, Alberta T5J 3N4, Canada

REFERENCES

1. Boyle MH, Furlong W, Feeny D, Torrance G, Hatcher J. Reliability of the Health Utilities Index - Mark III used in the 1991 Cycle 6 General Social Survey Health Questionnaire. *Qual Life Res*. 1995;4:249-257.
2. Brazier J, Walters SJ, Nicholl JP, Kohler B. Using the SF-36 and Euroqol on an elderly population. *Qual Life Res*. 1996;5:195-204.
3. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess*. 1999;3:i-iv, 1-164.
4. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health status from the SF-36. *J Health Econ*. 2002;21:271-292.
5. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chron Dis*. 1987;40:373-383.
6. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. *Pharmacoeconomics*. 2000;17:13-35.
7. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Control Clin Trials*. 1991;12:142S-158S.
8. Dorman P, Slattery J, Farrell B. Qualitative comparison of the reliability of health status assessments with the EuroQol and the SF-36 questionnaires after stroke. *Stroke*. 1998;29:63-68.
9. Feeny D, Furlong W, Barr RD, et al. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *J Clin Oncol*. 1992;10:923-928.
10. Feeny D, Furlong W, Torrance GW, et al. Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Med Care*. 2002;40:113-128.
11. Folstein MF, Folstein SE, McHugh PR. Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatric Res*. 1975;12:189-198.
12. Furlong WJ, Feeny DH, Torrance GW, Barr RD. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Ann Med*. 2001;33:375-384.

13. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hen MM. Functional assessment scales: A study of persons with multiple sclerosis. *Arch Phys Med Rehabil*. 1990;71:870-875.
14. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles for the functional independence measure. *Am J Phys Med Rehabil*. 1993;72:84-89.
15. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res*. 1993;2:441-449.
16. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index (HUI®): Concepts, measurement properties and applications. *Health Qual Life Outcomes*. 2003;1:54.
17. Juniper EF, Guyatt GH, Jaeschke R. How to develop and validate a new health-related quality of life instrument. In: Spilker S, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996: 49-56.
18. Kaplan RM, Anderson JP. The general health policy model: An integrated approach. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1996: 309-322.
19. Landis RJ, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
20. Luo, N, Chew LH, Fong KY, et al. A comparison of the EuroQol-5D and the Health Utilities Index Mark 3 in patients with rheumatic disease. *J Rheumatol*. 2003;30:2268-2274.
21. Macran S. Test-retest performance of EQ-5D. In: Brooks R, Rabin R, de Charro F, eds. *The measurement and valuation of health status using EQ-5D: A European perspective. Evidence from the EuroQol BIOMED Research Programme*. Dordrecht: Kluwer Academic Publishers; 2003: 43-54.
22. McDowell I, Newell C. *Measuring health: A guide to rating scales and questionnaires*. 2nd ed. New York: Oxford University Press, 1996.
23. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Qual Life Res*. 1995;4:293-307.
24. Medical Outcomes Trust, Scientific Advisory Committee. Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res*. 2002;11:193-205.
25. Norman G. Hi! How are you? Response shift, implicit theories and differing epistemologies. *Qual Life Res*. 2003;12: 239-249.
26. Ottenbacher KJ. Mann WC, Granger CV, et al. Inter-rater agreement and stability of functional assessment in the community-based elderly. *Arch Phys Med Rehabil*. 1994;75:1297-1301.
27. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: A quantitative review. *Arch Phys Med Rehabil*. 1996;77:1226-1232.
28. Petrella N, Overend T, Chesworth B. FIM after hip fracture: Is telephone administration valid and sensitive to change? *Am J Phys Med Rehabil*. 2002;81:639-644.
29. Pollak N, Rheault W, Stoecker JL. Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Arch Phys Med Rehabil*. 1996;77:1056-1061.
30. Rabin R, de Charro F. EQ-5D: A measure of health status from the EuroQol group. *Ann Med*. 2001;33:337-343.
31. Revicki D, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Qual Life Res*. 2000;9:887-900.
32. Roccaforte WH, Burke WJ, Bayer BL, Wengel SP. Validation of a telephone version of the Mini-Mental State Examination. *J Am Geriatr Soc*. 1992;40:697-702.
33. Schuck P. Assessing reproducibility for internal data in health-related quality of life questionnaires: Which coefficient should be used? *Qual Life Res*. 2004;13:571-586.
34. Segal, ME, Gillard M, Schall R. Telephone and in-person proxy agreement between stroke patients and caregivers for the functional independence measure. *Am J Phys Med Rehabil*. 1996;75:208-212.
35. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
36. Smith PM, Illig SB, Fiedler RC, Hamilton BB, Ottenbacher KJ. Intermodal agreement of follow-up telephone functional assessment using the functional independence measure in patients with stroke. *Arch Phys Med Rehabil*. 1996;77:9431-435.
37. Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. 2nd ed. Oxford: Oxford University Press; 1995.
38. Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic, and preference-based instruments. *Rheumatology*. 2000;39:783-790.
39. Tombaugh TN, McIntyre NJ. The Mini-Mental State Examination: A comprehensive review. *J Am Geriatr Soc*. 1992;40:922-935.
40. Torrance GW, Feeny DH, Furlong WJ, et al. Multi-attribute preference functions for a comprehensive health status classification system: Health Utilities Index Mark 2. *Med Care*. 1996;34:702-722.
41. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the functional independence measure in stroke. The impact of using different methods for measuring responsiveness. *J Clin Epidemiol*. 2002;55:922-928.