Stroke
JOURNAL OF THE AMERICAN HEART ASSOCIATION
American Stroke Association
A Division of American Heart Association

Variability in Modified Rankin Scoring Across a Large Cohort of International Observers

Terence J. Quinn, Jesse Dawson, Matthew R. Walters and Kennedy R. Lees

Stroke 2008, 39:2975-2979: originally published online August 7, 2008
doi: 10.1161/STROKEAHA.108.515262
Stroke is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0039-2499. Online ISSN: 1524-4628

Variability in Modified Rankin Scoring Across a Large Cohort of International Observers

Terence J. Quinn, MRCP; Jesse Dawson, MRCP; Matthew R. Walters, MD; Kennedy R. Lees, MD

Background and Purpose—The modified Rankin scale (mRS) is the most commonly used outcome measure in stroke trials. However, substantial interobserver variability in mRS scoring has been reported. These studies likely underestimate the variability present in multicenter clinical trials, because exploratory work has only been undertaken in single centers by a few observers, all of similar training. We examined mRS variability across a large cohort of international observers using data from a video training resource.

Methods—The mRS training package includes a series of "real-life" patient interviews for grading. Training data were collected centrally and analyzed for variability using kappa statistics. We examined variability against a standard of "correct" mRS grades; examined variability by country; and for UK assessors, examined variability by center and by professional background of the observer.

Results—To date, 2942 assessments from 30 countries have been submitted. Overall reliability for mRS grading has been moderate to good with substantial heterogeneity across countries. Native English language has had little effect on reliability. Within the United Kingdom, there was no significant variation by profession.

Conclusion—Our results confirm interobserver variability in mRS assessment. The heterogeneity across countries is intriguing because it appears not to be related solely to language. These data highlight the need for novel strategies to improve reliability. **(*Stroke*. 2008;39:2975-2979.)**

Key Words: clinical trials ■ modified Rankin Score ■ outcome assessment ■ stroke

The modified Rankin Scale (mRS) is the most commonly used measure of poststroke disability and is increasingly used as a primary outcome in stroke trials.[1] mRS is an ordinal hierarchical scale that describes grades of disability from 0="no symptoms" to 5="severe disability; bedridden, incontinent and requiring constant nursing care and attention"; with some authors adding an extra grade of mRS 6 to denote fatal outcome.[2] Although a strength of the mRS is the broad range of activity encompassed at each grade, distinctions between grades are poorly defined and until recently, mRS users had little guidance on rating.

This lack of guidance likely contributes to the significant interobserver variability present in traditional mRS grading.[3] Several groups have tested the reliability of mRS with varying results. Early single-center study of the clinometric properties of the scale reported moderate reliability,[2] but subsequent studies using raters from multiple centers within a city reported a concerning poor reliability for standard mRS assessment.[4] These trials provide useful data; however, by limiting themselves to a few experienced raters from a single center or city, they likely underestimate the true variability of mRS assessment manifest in a large-scale, international trial.

A better understanding of the reliability of mRS assessment is of more than academic interest. Variability in mRS grading has the potential to increase trial end point misclassification, which in turn will have deleterious effects on trial power and quality.[5]

To improve quality of multicenter trials, a standardized approach to outcome measurement is required. One potential method of improving consistency is to offer training and examination in end point classification. Case-based training in application of the National Institutes of Health Stroke Scale is well established, and successful completion of the National Institutes of Health Stroke Scale training is a prerequisite for several stroke trials.[6] We have developed a comprehensive training package for mRS assessment consisting of written educational materials, video-based tutorials, and mRS cases for grading. A full description of the package is available elsewhere.[3] In brief, investigators study the training resources and then attempt an assessment exercise that comprises a series of real-time mRS interviews for grading. Examination attempts are externally graded and success leads to certification in mRS training.

Our mRS training resource has been available for almost 5 years and has been widely used. Thousands of assessors from various countries and backgrounds have attempted the certification examination. These data have been collated centrally and offer a powerful resource for analysis of mRS variability

Received January 20, 2008; final revision received March 17, 2008; accepted April 10, 2008.

From the Division of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK.

Correspondence to Terence J. Quinn, MRCP, Gardiner Institute of Cardiovascular and Medical Sciences, University of Glasgow, Western Infirmary, Church Street, Glasgow G11 6NT, UK. E-mail Tjq1t@clinmed.gla.ac.uk

Stroke **is available at http://stroke.ahajournals.org** **DOI: 10.1161/STROKEAHA.108.515262**

and its determinants. Using these training data, we describe variability in mRS assessment across a large cohort of international researchers. To explore reasons for mRS variability, we further described mRS reliability by country, language, and background speciality of the assessor.

Methods

mRS Training Data

The video-based mRS certification exercise comprises 5 nonscripted interviews with stroke survivors. There is a further recertification examination comprising 4 cases. Interviews were originally recorded in English. Fully translated training packages, with native speakers overdubbing the interview, have been made available for Finnish, French, German, Italian, Portuguese, Spanish, and Swedish researchers; a subtitled Chinese version has also been produced.

Certification data are collated and scored centrally. Investigators can submit responses individually or through a local study coordinator. Standardized paper or electronic score sheets are used and mRS grades are then transcribed directly onto an electronic spreadsheet. Score sheets that are incorrectly completed or poorly legible are returned to the assessor for resubmission. An assessor who fails to achieve certification is invited to review the training materials and resubmit on the assessment exercise; no guidance is given on errors in original grading. There is no limit on the number of attempts an assessor can make. Initial certification is valid for 1 year after which time raters are invited to take a further assessment.

"Correct" mRS grades for each video interview were derived by 2 experts in mRS grading and were informed by the results of pilot data. mRS certification is graded using a "pass"/"fail" system. Embedded formulae within a dedicated database automatically calculate the investigator's final grade. A complete description of the process used to score mRS grading is available elsewhere.[3]

Certification data were anonymized before analysis of reliability; data on center, country, and profession were maintained but any identifying information was removed. The program was provided to trials on condition that we would have access to scoring data for quality control and research, and investigators' consent is expected to be covered by trial agreements. Like with any registry, participants can ask for details to be removed. Throughout the period of data collection and after publication of the mRS training system methodology,[3] we have received no such request.

Statistical Analyses

To assess for significant differences in certification performance between countries, professions, and centers, we used χ^2 analyses comparing proportions achieving the following certification results: fail; pass 3/5 correct; pass 4/5 correct; and pass 5/5 correct. Because numbers in the group "pass 3/5" were small, to prevent statistical error, they were combined with the "fail" group.

Reliability is a measure of consistency in multi-item scales. Quantifying the reproducibility of repeat scoring between graders gives interobserver variability. For our analyses, variability was described using kappa statistics (k), where $k=1$ defines perfect agreement between assessors, whereas $k=0$ defines no agreement other than that expected by chance. Standard definitions of poor ($k=0$ to 0.20), fair ($k=0.21$ to 0.40), moderate ($k=0.41$ to 0.60), good ($k=0.61$ to 0.80), and very good ($k=0.81$ to 1.00) agreement were used.[7]

Agreement was described across the cohort of observers and against the "standard" of predefined correct answers. Using an equivalent approach, we also described variability at each potential mRS grading. We performed 2 principal analyses: first describing reliability for assessors' initial attempt at the mRS exercise and a second analysis limited to those assessors who successfully completed the exercise. Further subanalysis was performed to describe agreement by country and by native English and nonnative English language countries. Native English speakers were arbitrarily defined as any assessors from centers in Australia, Canada, New Zealand, South Africa, the United Kingdom, and the United States.

Variability by background profession was also described. Background profession was classified as neurology or geriatrics/care of the elderly; specialist stroke physicians not specifically trained in neurology or geriatric medicine were classified as "general medicine." Nonphysician, clinical research assistants were classified as "research nurses." Background was extracted from submitted assessment documentation. Where background was not available, these data were collected from the host institution. To eliminate the effect of language and country, we limited the background analysis to UK-based assessors. To assess for potential bias in grading, mean mRS and 95% CI were calculated for each mRS grading. An equivalent analysis describing variability by institution was also performed for UK assessors.

All available certification attempts have been included in the analyses. Kappa statistics were described using attribute agreement functions. Statistics were performed using Minitab software (version 13.1; Minitab Inc).

Results

Certification assessments have been collated since March 2003; for this analysis, we finished data collection on January 2008. The total number of assessments to date is 2942 (2636 first certifications; 306 recertifications). Thus, data on 14 404 mRS assessments were available. Certification cases spanned a range of potential mRS scores: for mRS Grade 2, we included one video case for assessment; mRS 3=4 cases; mRS 4=3 cases; and mRS 5=one case.

Country of origin was available for 2349 certification attempts. Assessors were based in a variety of international centers (n=30 countries). The majority of assessors (1958 [75%]) achieved certification at the first attempt; 20 assessors required a third attempt at certification; and 4 assessors required 4 or more attempts.

Proportions of countries achieving certification grades of fail, pass 3/5 correct, pass 4/5 correct, and pass 5/5 correct are presented graphically for all countries submitting more than 50 assessments. Presented data represent performance according to assessor's first attempt (Figure A) and performance was limited to those assessors who achieved certification (Figure B). There was a significant difference in the performance of countries for both analyses ($P<0.0001$).

Interobserver variability and variability against the "correct" grade are presented for all countries submitting more than 50 certification assessments along with total variability across the cohort (Table 1). Other countries included in the overall analysis comprise a mix of Asian, European, Canadian, and African centers. Variability by country ranged from fair to very good. Variability for the entire cohort was good with no difference between English-speaking and non-English-speaking countries. For the complete cohort of assessors, variability at each level of mRS grading is presented (Table 2).

UK assessors comprised a mixed group of disciplines. Although reliability in mRS grading varied with professional background (Table 3), there was no statistically significant difference in certification assessment results ($P=0.321$). All groups tended to underscore disability. Heterogeneity in reliability was also present across the various UK centers. For those centers with greater than 5 raters completing the examination, (anonymous) results were: Center 1 (n=22) k: 0.59; Center 2 (n=12) k: 0.70; Center 3 (n=11) k: 0.74; Center 4 (n=11) k: 0.43; Center 5 (n=6) k: 0.67; and Center

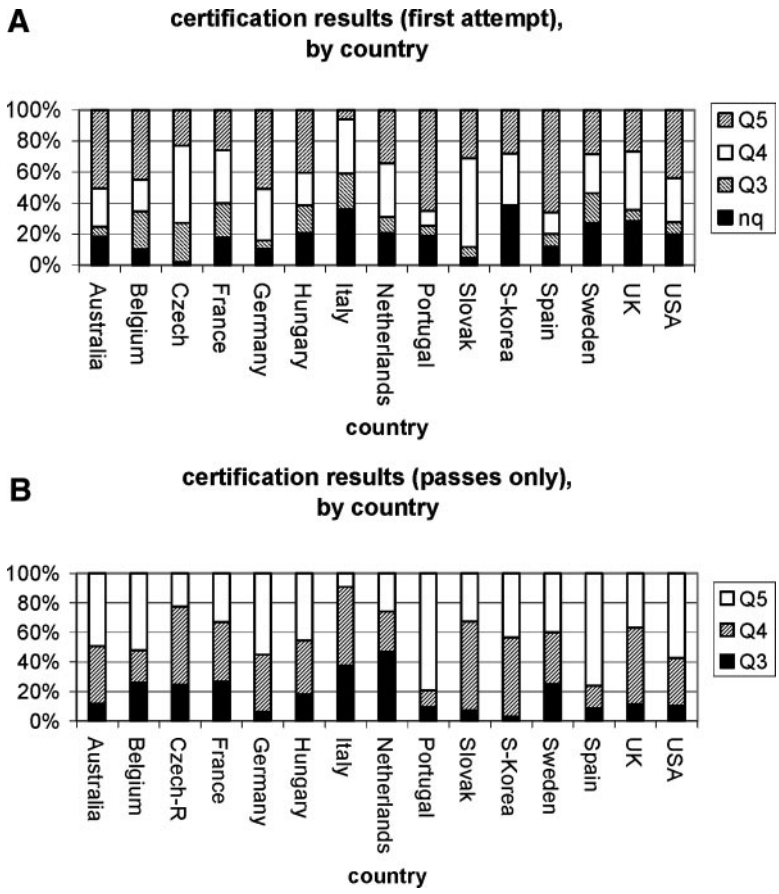

Figure. A, Performance on the mRS certification exercise at the first attempt. Data are for all countries submitting more than 50 assessments. nq indicates not qualified; Q3, qualified 3 of 5 answers correct; Q4, qualified 4 of 5 answers correct; Q5, qualified 5 of 5 answers correct. B, Performance on the mRS certification exercise limited to researchers who passed the certification examination. Data are for all countries submitting more than 50 assessments. Q3 indicates qualified 3 of 5 answers correct; Q4, qualified 4 of 5 answers correct; Q5, qualified 5 of 5 answers correct.

6 (n=6) *k*: 0.80. For this analysis, there was a statistically significant difference in performance across the UK centers (P=0.001).

Discussion

We have demonstrated considerable interobserver variability in mRS grading across a large cohort of international investigators. Variability was apparent for all included countries; however, within the cohort, there was substantial heterogeneity with a number of countries achieving only fair to moderate reliability. Our data confirm the potential for end point misclassification in a clinical stroke trial and suggest some reasons for this variability in scoring.

The statistical techniques used to describe variability demand some discussion. There is no accepted standard test for measurement of reliability. Use of kappa statistics has been criticized, because the basic assumptions underlying the calculations rely on observer independence.[8] We recognize that complete rater independence can never be guaranteed in a trial setting but chose to use kappa statistics in this study because clinicians are familiar with the test and previous studies of mRS reliability have used a similar approach. Traditional kappa statistics do not allow for comparative analysis. To assess whether the differences seen between countries were significant, we compared proportions achieving a "fail" or one of the "pass" grades using accepted techniques. Having thus established a significant difference in mRS grading between countries, we then described the interobserver variability in terms of kappa statistics.

We present 2 analyses of interobserver variability that describe differing aspects of mRS reliability. Analysis of assessors' first attempt at mRS gives an approximation of reliability across all potential stroke investigators. The second analysis, limited to assessors who successfully completed the training exercise, provides a better representation of the variability that would be seen in a contemporary stroke trial (the majority of trials that make use of the mRS training resource demand successful certification before the investigator can assess trial patients). Although this second measure is consistently better than the first, there is still substantial variability with heterogeneity across countries. By defining "correct" mRS grades for each patient interview, we were able to compare variability against a predefined standard. For certain countries, there was a marked discrepancy between interobserver variability and variability against the standard; this suggests that cohorts of raters were consistent in their grading but that this grading was wrong.

Our data compare favorably with previous estimates of mRS interobserver variability, in which *k* has been estimated at levels between fair (k=0.25)[4] and moderate (k=0.56).[2] This improvement in reliability will in part represent the beneficial effect of our mRS training resource[3] and perhaps increasing familiarity with mRS as a method of functional outcomes assessment.[1] However, there is no room for complacency; even those assessors who successfully completed the assessment demonstrated variability. Thus, although our results are encouraging, there is still some way to go before mRS reliability is optimized.

Table 1. modified Rankin Scale Variability by Country of Assessor for All Countries Submitting More Than 50 Certification Attempts*

	(1) First Attempt	(2) Passes Only
Australia	*k*=**0.60**	*k*=0.79
n=111 attempt 1/n=110 pass	(0.77 standard)	(0.86 standard)
Belgium	0.64	0.72
n=49/46	(0.73)	(0.78)
Czech Republic	0.70	0.74
n=49/48	(0.68)	(0.70)
France	**0.60**	0.83
n=72/67	(0.64)	**(0.52)**
Germany	0.66	0.77
n=162/159	(0.78)	(0.84)
Hungary	**0.60**	0.75
n=57/54	(0.70)	(0.79)
Italy	**0.55**	0.62
n=147/130	**(0.34)**	**(0.38)**
The Netherlands	**0.53**	0.75
n=58/49	**(0.50)**	(0.76)
Portugal	0.66	0.84
n=63/53	(0.80)	(0.91)
Slovakia	0.72	0.75
n=42/40	(0.75)	(0.77)
South Korea	**0.52**	0.74
n=55/30	(0.67)	(0.81)
Spain	0.73	0.83
n=314/299	(0.84)	(0.90)
Sweden	**0.55**	0.71
n=56/41	(0.65)	(0.74)
UK	**0.59**	0.74
n=109/95	(0.69)	(0.77)
US	0.61	0.77
n=172/162	(0.73)	(0.84)
Native English	0.66	0.69
n=580 total/389 pass	(0.77)	(0.77)
Nonnative English	0.67	0.71
n=1769/1251	(0.76)	(0.78)
All	0.67	0.71
n=2942/2151	(0.76)	(0.78)

*Variability (*k*) presented as interobserver variability and variability against a standard of correct grade. Analysis (1) observers; first attempt at assessment exercise; (2) limited to those assessors who achieved an overall "pass." Reliabilities scored as moderate or poorer are highlighted in bold text.

Table 2. Columns (1) and (2): Variability (*k*) at Each Grade of modified Rankin Scale; Columns (3) and (4): Mean mRS at Each 'Correct' Grade for All Attempts at mRS Restricted to Those Who Achieved Certification*

mRS	(1) Interobserver Variability	(2) Variability Against Standard	(3) Mean mRS (SD) All Assessors	(4) Mean mRS (SD) 'Passes' Only
0	...	...		
1	0.19	...		
2	0.48	0.56	1.7 (0.46)	1.8 (0.47)
3	0.74	0.79	2.8 (0.51)	2.8 (0.40)
4	0.95	0.97	3.9 (0.43)	3.9 (0.31)
5	...	...		

*Original assessment exercise did not include patients from full range of disabilities.

Across the complete cohort of assessors, interobserver variability was greatest for Grades 1 and 2. This finding is of particular significance for clinical trial end point analysis in which mRS outcomes are often dichotomized at grades of 1, 2, or 3. Thus, in our cohort, variability is potentially greatest for those mRS grades most likely to influence the final trial result. Increased variability at these middle grades is well described. It may be attributable to better definition of the highest and lowest categories or to the potential for misclassification in one direction only at extremes of mRS.[9]

The substantial variation in reliability observed between countries is intriguing. Our data suggest that this is not purely a function of language as countries with native English performed similarly to other countries. We acknowledge the global nature of contemporary medical practice where a given institution may have a number of international staff and that most centers recruiting for international trials will be staffed by teams familiar with English. It is possible that sociocultural factors related to perceptions of disability and handicap may influence patterns of grading. For this reason, in collaboration with our unit, international centers are producing new assessment cases in the native language and featuring local stroke survivors. We recognize that our data will include countries that may have a number of centers relatively inexperienced in stroke trials/mRS assessment. Increasing use and familiarity with the scale may help to remove some of this variability. Our available data do not allow for assessment of training effects; we encourage stroke trialists to continue to use available and forthcoming recertification materials because data from these resources will allow for future analysis of training effect and hopefully should demonstrate greater improvements in reliability.

Our analyses of UK assessors suggest that heterogeneity in measured reliability is not only accounted for by nationality. Accepting the smaller numbers of certification attempts available, there was again heterogeneity between centers and between professions, although only the analysis across differing centers revealed a significant difference in performance. Interobserver variability between differing professional backgrounds has been demonstrated for other neurological outcome scales including the Barthel Index[10] and the Unified Parkinson's Disease Rating Scale motor examination.[11]

Some aspects of our methodology demand discussion. Although describing variability through performance on a video-based assessment allows for standardization across a large number of raters, there is the potential to overestimate reliability using this method. Variability recorded in a series of traditional face-to-face mRS gradings would likely be substantially higher, because assessors can use various ap-

Table 3. Variability (*k*) in mRS Scoring and Mean Submitted Grade for UK Assessors by Background Profession

	Interobserver Variability	Variability Against Standard	Mean Grade (95% CI) mRS 2	Mean Grade (95% CI) mRS 3	Mean Grade (95% CI) mRS 4
General medicine	0.66	0.77	1.67	2.87	3.95
(n=13)			(1.35–1.98)	(2.72–3.02)	(3.87–4.04)
Geriatrics	0.54	0.68	1.91	2.91	3.89
(n=23)			(1.73–2.10)	(2.73–3.10)	(3.79–3.98)
Neurology	0.56	0.65	2.00	3.03	4.00
(n=16)			(1.80–2.19)	(2.78–3.28)	(4.00–4.00)
Research nurse	0.65	0.72	1.79	2.77	4.00
(n=58)			(1.67–1.91)	(2.66–2.88)	(3.97–4.02)

proaches to patient interviewing. In creating the mRS assessment, we had to strike a balance between including a broad range of cases at varying levels of disability and having an assessment that was short enough to be acceptable to a large population of researchers. Our final choice of included cases deliberately focuses attention on those mRS grades known to demonstrate greatest variability, ie, mRS 2 to 4. For future training packages, we intend to include more cases from extremes of the mRS spectrum.

The strengths of our analysis are its size and international scope. Major acute stroke trials may involve as many as 400 hospitals at international sites, each with 2 to 5 raters.

Previous descriptions of the clinometric properties of mRS have been limited to few observers often from single centers. Certification in mRS grading was required for a number of recently completed and ongoing multicenter stroke trials, and investigators from many of the leading stroke research centers are included in this analysis. The initial users of our novel resource may well have overrepresented enthusiasts and specialists in the field who already have considerable experience of mRS. For this reason, we have waited until the mRS teaching package was well established before attempting any analysis of training data.

Having demonstrated this interobserver variability in mRS, it is incumbent on stroke trialists to take steps to improve reliability in outcomes assessment. The DVD training package was designed for this purpose and has been a success. However, even those raters who achieve certification still demonstrate a degree of interobserver variability, and we suspect that training alone will not eradicate variability. Other methodologies to improve mRS assessment are available or are being developed. Use of a structured interview may significantly improve reliability,[4] although results from other groups using a structured approach have been less favorable.[12] End point adjudication using 2 or more remote reviewers is now commonplace in multicenter trials; we are piloting a group-based mRS assessment methodology that may further improve reliability.

We have demonstrated interobserver variability in a large representative cohort of international researchers. Although country and background profession seem to influence this variability, the significant differences between similar local UK centers suggest that reasons for variability are more complex than simple sociogeographic differences. Because variability between assessors may never be fully explained, we should continue to pilot novel methodologies to minimize interobserver variability in clinical trials.

Acknowledgments

We are grateful to all the investigators who have attempted the mRS certification assessment. We acknowledge also the excellent work of Mrs Pamela Mackenzie in collecting and inputting raw training data.

Disclosures

K.R.L. has assisted with a number of trials that used the mRS training resource: international principal investigator of the SAINT I trial and chair of the steering committee for the CHANT and SAINT I and II trials. He has received fees, expenses, and institutional grants relating to these and other trials from GlaxoSmithKline, AstraZeneca, and several other pharmaceutical companies that have been or are developing treatments for stroke. K.R.L., M.R.W., J.D., and T.J.Q. have applied for academic grant support to continue work on developing stroke outcome assessments using mRS and have published data in support of mRS as the optimal end point for acute stroke trials.

References

1. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome assessment in contemporary stroke trials [Abstract]. *Stroke*. 2008;39:692.
2. van Swieten Koudstaal PJ, Visser MC, Schouten HJA, Gijn JV. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604–607.
3. Quinn TJ, Dawson J, Lees KR, Hardemark HG, Walters MR. Initial experience of a digital training resource for modified Rankin Scale assessment in clinical trials. *Stroke*. 2007;38:2257–2261.
4. Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke*. 2005;36:777–781.
5. Choi SC, Clifton GL, Marmarou A. Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. *J Neurotrauma*. 2002;19:17–22.
6. Lyden P, Brott T, Tilley B, Welch KMA, Mascha EJ, Levine S, Haley EC, Grotta J, Marler J; NINDS TPA Stroke Study Group. Improved reliability of the NIH Stroke Scale using video training. *Stroke*. 1994;25:2220–2226.
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
8. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. *J Clin Epidemiol*. 1988;41:59–68.
9. Banks JL, Marotta AC. Outcomes validity and reliability of the modified Rankin Scale: implications for stroke clinical trials. *Stroke*. 2007;38:1091–1096.
10. Richards SH, Peters TJ, Coast J, Gunnell DJ, Darlow MA, Pounsford J. Inter-rater reliability of the Barthel ADL index: how does a researcher compare to a nurse? *Clin Rehabil*. 2000;14:72–78.
11. Post B, Merkus MP, de Bie RM, de Haan RJ, Speelman JD. Unified Parkinson's Disease Rating Scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov Disord*. 2005;20:1577–1584.
12. Newcommon NJ, Green TL, Haley E, Cooke T, Hill MD. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin scale. *Stroke*. 2003;34:377–378.