Stroke
JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association | American Stroke Association

Improving the Reliability of Stroke Disability Grading in Clinical Trials and Clinical Practice: The Rankin Focused Assessment (RFA)

Jeffrey L. Saver, Bogdan Filip, Scott Hamilton, Anna Yanes, Sharon Craig, Michelle Cho, Robin Conwit and Sidney Starkman

Stroke. 2010;41:992-995; originally published online April 1, 2010;
doi: 10.1161/STROKEAHA.109.571364
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628

Improving the Reliability of Stroke Disability Grading in Clinical Trials and Clinical Practice

The Rankin Focused Assessment (RFA)

Jeffrey L. Saver, MD; Bogdan Filip, MD; Scott Hamilton, PhD; Anna Yanes, RN;
Sharon Craig, RN; Michelle Cho, BS; Robin Conwit, MD; Sidney Starkman, MD;
for the FAST-MAG Investigators and Coordinators

Background and Purpose**—**The modified Rankin Scale rates global disability after stroke and is the most comprehensive and widely used primary outcome measure in acute stroke trials. However, substantial interobserver variability in modified Rankin Scale scoring has been reported. This study sought to develop and validate a short, practicable structured assessment that would enhance interrater reliability.

Methods**—**The Rankin Focused Assessment was developed by selecting and refining elements from prior instruments. The Rankin Focused Assessment takes 3 to 5 minutes to apply and provides clear, operationalized criteria to distinguish the 7 assignable global disability levels. The Rankin Focused Assessment was prospectively validated 3 months poststroke among 50 consecutive patients enrolled in the Phase 3 National Institutes of Health Field Administration of Stroke Therapy–Magnesium (FAST-MAG) Trial.

Results**—**Among the 50 patients, mean age was 71.5 years (range, 43 to 93 years), 48% were female, and stroke subtype was hemorrhagic in 24%. At Day 90, 43 patients were alive and 7 had died. The modified Rankin Scale median was 2.0 and mean was 2.8. When pairs of 14 raters assessed all enrolled patients, the percent agreement was 94%, the weighted κ was 0.99 (95% CI, 0.99 to 1.0), and the unweighted κ was 0.93 (95% CI, 0.85 to 1.00). Among the 43 surviving patients, the percent agreement was 93%, the weighted κ was 0.99 (0.98 to 1.0), and the unweighted κ was 0.91 (0.82 to 1.00).

Conclusions**—**The Rankin Focused Assessment yields high interrater reliability in the grading of final global disability among consecutive patients with stroke participating in a randomized clinical trial. The Rankin Focused Assessment is brief and practical for use in multicenter clinical trials and quality improvement activities. **(*Stroke*. 2010;41:992-995.)**

Key Words: cerebral infarction ■ clinical trial ■ disability ■ outcomes ■ scales

The modified Rankin Scale (mRS) is the most comprehensive and most widely used primary outcome measure in contemporary acute stroke trials.[1–3] The mRS is an ordinal, hierarchical scale that assigns patients among 7 global disability levels ranging from 0 (no symptoms) to 5 (severe disability) and 6 (death). Formal clinometric investigations have demonstrated that the mRS has good responsiveness and excellent construct and convergent validity. However, substantial interobserver variability in mRS scoring has been reported.[4–6] Interrater variability introduces noise into trial outcome assessments and reduces the power of clinical trials to detect treatment effects.

A variety of approaches to minimize interrater variation of the mRS have been described or proposed, including (1) use of a formal structured interview[7]; (2) training and certification programs using written and video case vignettes[8]; and (3) central panel adjudication of local site-recorded video assessments.[6] However, the instruments and approaches developed to date have not consistently been shown to reduce interrater variability.

The purpose of this study was to develop and validate a systematic, structured assessment tool to guide raters in assigning mRS grades.

Methods

Assessment Tool Construction

The Rankin Focused Assessment (RFA) was developed by a working group consisting of physicians with extensive stroke clinical trial experience (J.L.S., S.S.), the head nurse coordinator (A.Y.), and the study monitor (S.C.) of the National Institutes of Health (NIH) Field Administration of Stroke Therapy–Magnesium (FAST-MAG) Phase 3 clinical trial with additional input from the 14 nurse-coordinators

Received October 26, 2009; final revision received December 13, 2009; accepted December 17, 2009.

From the Stroke Center and Department of Neurology (J.L.S., B.F., A.Y., S.C., M.C., S.S.), University of California, Los Angeles, Los Angeles, Calif; Stanford University, Palo Alto, Calif (S.H.); and the National Institute for Neurological Disorders and Stroke (R.C.), National Institutes of Health, Bethesda, Md.

Correspondence to Jeffrey L. Saver, MD, UCLA Stroke Center, 710 Westwood Plaza, Los Angeles, CA 90095. E-mail jsaver@ucla.edu

Stroke is available at http://stroke.ahajournals.org DOI: 10.1161/STROKEAHA.109.571364

Table 1. Crosstabulation of Paired, RFA-Guided mRS Ratings of 50 Consecutive Patients

	0	1	2	3	4	5	6
0	8	1					
1		8					
2			9	1			
3				2			
4					3	1	
5						10	
6							7

Table 2. Interrater Agreement for All Dichotomizations of the mRS

Rankin Cut Point	Observed Agreement	κ
0 versus 1–6	98.0%	0.93
0–1 versus 2–6	100%	1.00
0–2 versus 3–6	98.0%	0.96
0–3 versus 4–6	100%	1.00
0–4 versus 5–6	98.0%	0.96
0–5 versus 6	100%	1.00

performing outcome assessments at 47 participating hospitals in the trial. This mixed group of expert and novice trial staff selectively extracted, revised, and combined elements of prior instruments and generated new elements to construct the assessment tool. Important sources for tool construction were the mRS itself, the Structured Interview developed by Wilson and colleagues,[7,9] the videotapes and teaching booklet developed by Lees and colleagues,[8] and the working group's daily experience in implementing the mRS in an ongoing trial. The assessment tool was piloted and iteratively refined in small groups of patients. The final tool was then prospectively tested in 50 consecutive trial patients.

The RFA consists of a 4-page form accompanied by a 5-page instruction sheet (Supplemental Data; available at http://stroke.ahajournals.org). When performed after brief review of medical records and an NIH Stroke Scale examination, the RFA is typically completed in 3 to 5 minutes. The assessment specifies clear, operationalized criteria to distinguish among the 7 assignable global disability levels. To determine which criteria a patient meets, the assessment permits and encourages the rater to gather data from all available useful sources, including interviews with the patient and caregivers, medical records, rehabilitation therapist notes, and the rater's own examination of the patient. In addition to checkmarked items, the assessment tool includes text boxes in which the rater specifies the particular, concrete functional difficulties identified that led to an item being checked, facilitating review of the accuracy of a particular rating and ongoing training of novice by more expert raters. Separate versions of the RFA have been developed to assess a patient's current poststroke functional status and their historical prestroke functional status. In this study, the RFA to determine the patient's current, poststroke mRS score was evaluated.

Prospective Validation

The prospective validation study was performed in 50 consecutively enrolled patients undergoing 90-day mRS assessment in the NIH FAST-MAG Trial. At the 90-day visit, 2 different nurse-coordinators performed the mRS in succession with neither present in the room during the other's evaluation and the second coordinator blinded to the first's rating. Coordinator pairs were selected from a pool of 14 active coordinators (13 nurse-coordinators, 1 nonnurse-coordinator) in the trial. One coordinator was the individual assigned to perform the primary 90-day outcome evaluation by study operating procedures. The second coordinator was selected based on geographic and scheduling availability. No single rater performed >11 patient ratings.

Statistical Analysis

The primary outcome measure was the weighted κ coefficient reflecting agreement over the entire range of the mRS above chance among the rater pairs. Following standard convention κ scores of 0.0 to 0.2 would be considered poor, 0.21 to 0.40 fair, 0.41 to 0.6 moderate, 0.61 to 0.8 good, and 0.81 to 1.0 excellent. In addition, to permit comparison with the range of reliability measures reported in prior studies of the mRS, we also calculated the unweighted κ over the entire range of the mRS considering all ratings, the κ for dichotomizations of the mRS, and the crude rate of agreement of raters (percent agreement) unadjusted for chance concurrence. These reliability scores were calculated for the paired ratings obtained among all 50 consecutively enrolled patients and among all survivors from this group at Day 90.

Results

Among the 50 patients with stroke, average age was 71.5 years (range, 43 to 93 years), 48% were female, and final diagnosis was ischemic stroke in 66%, hemorrhagic stroke in 26%, and transient ischemic attack in 8%. Neurological deficit at the time of enrollment in the trial in the prehospital setting was a median Los Angeles Motor Scale score of 4 (range, 1 to 5), whereas the first NIH Stroke Scale obtained after hospital arrival (and after exposure to prehospital study drug) was median 10.5 (range, 0 to 40).

At Day 90, 43 patients were alive and 7 had died. Across all 50 patients, the median NIH Stroke Scale was 3 (interquartile range, 0 to 10; range, 0 to 42). In the 43 alive patients, the Mini Mental Status Examination score was median 29 (interquartile range, 23 to 30).

Frequencies of mRS scores among all 100 ratings were: mRS 0 in 17%, mRS 1 in 17%, mRS 2 in 19%, mRS 3 in 5%, mRS 4 in 7%, mRS 5 in 21%, and mRS 6 in 14%. The mRS median was 2.0 and mean 2.8. The crosstabulation of pair ratings in shown in Table 1. Raters' scores concurred fully in 47 of the 50 patients, and in the remaining 3 patients, scores differed by 1 level. Consequently, among all enrolled patients, for assigning patients among all possible mRS scores, the percent agreement was 94%, the weighted κ was 0.99 (95% CI. 0.99 to 1.00), and the unweighted κ was 0.93 (95% CI, 0.85 to 1.00). Among the 43 surviving patients, the percent agreement was 93%, the weighted κ was 0.99 (0.98 to 1.0), and the unweighted κ was 0.91 (0.82 to 1.00).

The κ scores for the 6 possible dichotomizations of the mRS are shown in Table 2 and ranged from 0.93 to 1.00.

Discussion

In this investigation, raters using the RFA achieved excellent interrater reliability in assigning final outcome mRS disability ratings to patients 3 months after an index stroke. The interrater reliability in assigning mRS grades achieved with use of the RFA was substantially better than in most prior studies. In a recent meta-analysis of 10 prior studies of the interrater reliability of the mRS, the combined achieved unweighted κ was moderate at 0.46[10] compared with the unweighted κ of 0.93 (95% CI, 0.85 to 1.00) observed in this study.

When he initially presented the scale over 50 years ago, Rankin provided only brief, broad descriptions for the categories of the mRS without clear operational criteria distinguishing 1 level from the next.[11] Consequently, the original scale leaves substantial leeway open to raters to develop idiosyncratic criteria or to apply the scale in an impressionistic manner.[7] Because of the only moderate reliability of unstructured methods for assigning mRS grades, more formalized approaches have been previously developed by other groups. However, these have achieved only inconsistent or modest improvements in reliability and have additional potential drawbacks. The RFA was designed to incorporate elements of, and lessons learned from, these prior algorithms.

The Structured Interview (SI) for the mRS,[7,9] developed by Wilson and colleagues, was a pioneering instrument that first introduced a systematic approach to assigning mRS levels. However, the SI is somewhat complex to implement, which has limited its deployment in actual clinical trials. In addition, the SI improves the reliability of the mRS only moderately (κ in meta-analysis of 0.62).[10] The training DVD digital system developed by Quinn and colleagues is modeled on the successful NIH Stoke Scale training and certification system and has been widely adopted in clinical trials.[8] However, this system has also been found to only moderately improve mRS reliability.[10]

The RFA differs from these and other prior instruments in several distinct ways. Like the SI, the RFA has raters elicit information regarding specific functional items in 5 sections: (1) constant care; (2) basic activities of daily living; (3) instrumental activities of daily living; (4) limitations in participation in usual social roles; and (5) the presence of common stroke symptoms. However, the RFA encourages the rater to gather information on patient functional performance from all available sources, including patient self-report, caregiver observations, physical therapist notes, physician and nursing records, and the rater's own examination and interaction with the patient. In contrast, the SI is written in a manner that encourages elicitation of information from a single informant, a potentially problematic approach because individuals often have incomplete or biased perceptions of performance. Patients with anosognosia may underestimate and patients with the catastrophic reaction may overestimate their deficits. Individual family members may only see patients in limited settings and not have a fully rounded picture of performance.

The RFA rates the patient based on current actual capacity and performance. In contrast, the SI asks the rater to factor out prestroke disability when assign a rating, forcing the rater to speculate on what the patient's capacity and performance would have been if they had no other complicating conditions. Rating only specifically stroke-related dysfunction has advantages and disadvantages. A theoretical advantage is that scores reflect a treatment's effect on the target condition uncontaminated by pre-existing deficits. However, in practice, identifying what a patient's functional status would have been had they not had any pre-existing conditions requires conjecture by raters likely to decrease interrater reliability. Also, it makes the assessment instrument more complex, requiring documentation of which deficits are due to prior disability and which are due to stroke-only disability. Moreover, this approach differs from that taken for other standard outcome measures in stroke clinical trials. The Barthel Index, the NIH Stroke Scale, and mortality status are all scored based on all-cause sources, not just those speculated to be due to stroke alone. Additionally, restricting disability rating to stroke-specific items is problematic in clinical trials because interventions can alter functional outcome through nonstroke mechanisms, for example, an adverse effect producing disabling congestive heart failure.

Like the SI, the RFA provides a detailed algorithm for scoring that concretely operationalizes criteria for distinguishing 1 mRS level from another. This approach is appropriate for instruments designed to enhance diagnostic judgments. The digital training vignette system, in contrast, does not as clearly provide operationalized criteria for assigning rankings. The experiential training and certification process works well for promoting interrater reliability in performing tasks that are intrinsically highly operationalized such as physical examination techniques (like the NIH Stroke Scale) but may be less useful for tasks requiring rendering of complex diagnostic judgments that have not been clarified by formal performance algorithms. Also, in developing the RFA, several items from the SI that raters found ambiguous or difficult were reworded or eliminated. For example, the SI item asking if patients need assistance to look after household expenses was reframed in terms of patient capacity rather than actual recent activity, because many fully capable elders do not perform this activity even when completely healthy, relying on their spouse.

The RFA is logistically much simpler to use in multicenter clinical trials and in local quality improvement projects than performance of mRS ratings by a central core laboratory either by direct teleconference interviews or by central rating of videotapes of locally performed interviews. Also, remote adjudication panel rating approaches have not yet been prospectively validated.

An additional strength of the present study is that it was conducted using researchers working in an actual clinical trial interviewing real stroke survivors. The study accordingly has ecological validity for this common application of the mRS. Also, the coordinator raters in the study had a wide range of prior trial experience, from limited to extensive. Accordingly, the findings are likely to be generalizable to a wide range of raters.

We analyzed RFA performance using several statistical indices, weighted κ, unweighted κ, and percent agreement. Each provides useful insight. The percent agreement measure perhaps most accords with lay and clinician understanding of interrater agreement but does not take into account chance concurrences. The unweighted κ reflects performance above chance but penalizes near misses to the same degree as wide disagreements. This metric is perhaps most appropriate for dichotomized applications of the mRS. The weighted κ penalizes near misses to a lesser degree than wide disagreements and is most relevant when the mRS is analyzed over several levels.[12–14]

We included patients with fatal outcome in our main analysis. Because contemporary clinical trials routinely use the 7-level mRS version that includes a fatal outcome level, it is important to include these patients to obtain interrater reliability estimates that accurately indicate how the RFA will perform in clinical trials. Some prior mRS reliability studies have not included these patients, which will tend to lower the estimate of interrater reliability, because agreement on the hard end point of death is straightforward. To permit comparison with these studies, we also reported data on RFA performance in stroke survivors only, and it showed excellent reliability even when confined to this group.

This study has limitations. The patient sample was moderate in size. Physician-investigators did not participate as raters, nonnurse-coordinators participated only to a limited degree, and all assessors were from a single trial group and a single country. Further reliability testing of the RFA in more diverse assessor groups would be beneficial. The study did not compare in the same patients mRS scores obtained with the RFA and scores obtained with any of the current common methods of scoring. Although there is no single widely accepted scoring method that can serve as a pre-existing gold standard, such comparisons with past practices would be of interest.

The RFA is brief and practical for use in multicenter clinical trials and routine practice. In this study, the RFA yielded high interrater reliability in grading the final global disability of patients participating in a randomized acute stroke treatment clinical trial. Further testing and validation of the RFA in larger studies with a more diverse group of assessors is desirable.

Acknowledgments

We are grateful to the FAST-MAG study coordinators who performed RFA assessments: Star Feudi, Miranda Gordon, Theresa Haley, Suzanne Henthorn, Keta Hodgson, Amy Ingram, Kanika McCall, Chisom Onwunyi, Vicky Padilla, Robert Phoenix, Brenda Pierce, Randy Sanoff, Heavenly Swendsen, and Barbara Tone. A full listing of the FAST-MAG Investigators is available at www.fastmag.info.

Source of Funding

This study was supported by National Institutes of Health–National Institute of Neurological Disorders and Stroke Award U01 NS 44364.

Disclosures

None.

References

1. Duncan PW, Jorgensen HS, Wade DT. Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. *Stroke*. 2000;31:1429–1438.
2. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin Scale: implications for stroke clinical trials: a literature review and synthesis. *Stroke*. 2007;38:1091–1096.
3. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. *Int J Stroke*. 2009;4:200–205.
4. New PW, Buchbinder R. Critical appraisal and review of the Rankin Scale and its derivatives. *Neuroepidemiology*. 2006;26:4–15.
5. Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in modified Rankin scoring across a large cohort of international observers. *Stroke*. 2008;39:2975–2979.
6. Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified Rankin Scale. *Stroke*. 2009;40:762–766.
7. Wilson JTL, Hareendran A, Grant M, Baird T, Schulz UGR, Muir KW, Bone I. Improving the assessment of outcomes in stroke: use of a structured interview to assign grades on the modified Rankin Scale. *Stroke*. 2002;33:2243–2246.
8. Quinn TJ, Lees KR, Hardemark HG, Dawson J, Walters MR. Initial experience of a digital training resource for modified Rankin Scale assessment in clinical trials. *Stroke*. 2007;38:2257–2261.
9. 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.
10. Quinn T, Dawson J, Walters M, Lees K. Reliability of the modified Rankin Scale: a systematic review. *Stroke.* 2009;40:3393–3395.
11. Rankin J. Cerebral vascular accidents in patients over the age of 60, II: prognosis. *Scott Med J*. 1957;2:200–215.
12. Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *Int J Stroke*. 2008;3:78–84.
13. Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. *Neurology*. 2009;72:1310–1315.
14. Saver JL, Gornbein J, Grotta J, Liebeskind D, Lutsep H, Schwamm L, Scott P, Starkman S. Number needed to treat to benefit and to harm for intravenous tissue plasminogen activator therapy in the 3- to 4.5-hour window: joint outcome table analysis of the ECASS 3 trial. *Stroke*. 2009;40:2433–2437.