IMPROVING THE RELIABILITY OF STROKE DISABILITY GRADING IN CLINICAL TRIALS AND CLINICAL PRACTICE: THE RANKIN FOCUSED ASSESSMENT (RFA)

Jeffrey L. Saver, MD[1], **Bogdan Filip, MD**[1], **Scott Hamilton, PhD**[2], **Anna Yanes, RN**[1], **Sharon Craig, RN**[1], **Michelle Cho, BS**[1], **Robin Conwit, MD**[3], and **Sidney Starkman, MD**[1] for the FAST-MAG Investigators and Coordinators

[1] Stroke Center and Department of Neurology, University of California, Los Angeles

[2] Stanford University

[3] National Institute for Neurological Disorders and Stroke, National Institutes of Health

Abstract

Background—The modified Rankin scale (mRS) rates global disability after stroke and is the most comprehensive and widely employed primary outcome measure in acute stroke trials. However, substantial inter observer variability in mRS scoring has been reported. This study sought to develop and validate a short, practicable structured assessment that would enhance inter-rater reliability.

Methods—The Rankin Focused Assessment (RFA) was developed by selecting and refining elements from prior instruments. The RFA takes 3–5 minutes to apply and provides clear, operationalized criteria to distinguish the 7 assignable global disability levels. The RFA was prospectively validated 3 months poststroke among 50 consecutive patients enrolled in the phase 3 NIH FAST-MAG Trial.

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

Conclusions—The Rankin Focused Assessment yields high inter-rater reliability in the grading of final global disability among consecutive stroke patients participating in a randomized clinical trial. The RFA is brief and practical for use in multicenter clinical trials and quality improvement activities.

Keywords

Cerebral infarction; Clinical Trial; Disability; Scales; Outcomes

The modified Rankin scale (mRS) is the most comprehensive and most widely employed primary outcome measure in contemporary acute stroke trials.[1, 2, 3] The mRS is an ordinal, hierarchical scale that assigns patients among seven global disability levels, ranging from 0

Address for Correspondence: Jeffrey L. Saver, MD, UCLA Stroke Center, 710 Westwood Plaza, Los Angeles, CA 90095, Tel 310-794-6379/Fax 310-267-2063, jsaver@ucla.edu.

(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] Inter-rater variability introduces noise into trial outcome assessments and reduces the power of clinical trials to detect treatment effects.

A variety of approaches to minimize inter-rater variation of the modified Rankin Scale have been described or proposed, including 1) use of a formal structuredinterview,[7] 2) training and certification programs using written and videocasevignettes,[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 inter-rater 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 (JLS, SS), the head nurse coordinator (AY), and the study monitor (SC) of the NIH Field Administration of Stroke Therapy – Magnesium (FAST-MAG) phase 3 clinical trial, with additional input from the 14 nurse-coordinators 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 Rankin Focused Assessment consists of a 4 page form, accompanied by a 5 page Instruction Sheet(Supplemental Online Materials). When performed after brief review of medical records and an NIH Stroke Scale examination, the RFA is typically completed in 3–5 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 checkmark items, the Assessment tool includes text boxes in which the rater specifies the particular, concrete functional difficulties identified that led 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, two 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 non-nurse 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 more than 11 patient ratings.

Statistical Analysis

The primary outcome measure was the weighted kappa coefficient, reflecting agreement over the entire range of the mRS above chance among the rater pairs. Following standard convention kappa scores of 0.0–0.2 would be considered poor, 0.21–0.40 fair, 0.41–0.6 moderate, 0.61–0.8 good, and 0.81–1.0 excellent. In addition, to permit comparison with the range of reliability measures reported in prior studies of the mRS, we also calculated the unweighted kappa over the entire range of the mRS considering all ratings, the kappa 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 stroke patients, average age was 71.5 (range 43–93) 48% were female, and final diagnosis was ischemic stroke in 66%, hemorrhagic stroke in 26%, and transient ischemic attack in 8%. Neurologic deficit at the time of enrollment in the trial in the prehospital setting was a median Los Angeles Motor Scale score of 4 (range 1–5), while the first NIH Stroke Scale obtained after hospital arrival (and after exposure to prehospital study drug) was median 10.5 (range 0–40).

At day 90, 43 patients were alive and 7 had died. Across all 50 patients, the median NIHSS was 3, IQR 0–10, range 0–42. In the 43 alive patients, the Mini Mental Status Examination score was median 29, IQR 23–30.

Frequencies of mRS scores among all 100 ratings were: mRS 0 – 17, mRS 1–17, mRS 2 – 19, mRS 3 – 5, mRS 4 – 7, mRS 5–21, mRS 6 – 14. The mRS median was 2.0 and mean 2.8. The cross-tabulation 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 kappa was 0.99 (95% CI 0.99–1.00), and the unweighted kappa was 0.93 (95% CI 0.85–1.00). Among the 43 surviving patients, the percent agreement was 93%, the weighted kappa was 0.99 (0.98–1.0), and the unweighted kappa was 0.91 (0.82–1.00).

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

Discussion

In this investigation, raters using the Rankin Focused Assessment achieved excellent inter-rater reliability in assigning final outcome modified Rankin Scale disability ratings to patients three months after an index stroke. The inter-rater 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 inter-rater reliability of the mRS, the combined achieved unweighted kappa was moderate at 0.46,[10] compared with the unweighted kappa of 0.93 (95% CI 0.85–1.00) 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 one 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 (kappa 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 five 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) 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 as 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 inter-rater reliability. Also, it makes the assessment instrument more complex, requiring documentation of which deficits are due to prior disability and which 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 since interventions can alter functional outcome through nonstroke mechanisms, e.g. 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 one 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 inter-rater reliability in performing tasks that are intrinsically highly operationalized, such as physical exam techniques (like the NIHSS), 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 upon their spouse.

The RFA is logistically much simpler to employ 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 ecologic 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 kappa, unweighted kappa, and percent agreement. Each provides useful insight. The percent agreement measure perhaps most accords with lay and clinician understanding of inter-rater agreement, but does not take into account chance concurrences. The unweighted kappa 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 kappa 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. Since contemporary clinical trials routinely employ the 7 level mRS version that includes a fatal outcome level, it is important to include these patients to obtain inter-rater 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 inter-rater reliability, since agreement on the hard endpoint 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, non-nurse 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 Rankin Focused Assessment is brief and practical for use in multicenter clinical trials and routine practice. In this study, the RFA yielded high inter-rater 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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. [PubMed: 10835468]

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. [PubMed: 17272767]

3. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. Int J Stroke 2009;4:200–205. [PubMed: 19659822]

4. New PW, Buchbinder R. Critical appraisal and review of the rankin scale and its derivatives. Neuroepidemiology 2006;26:4–15. [PubMed: 16272826]

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. [PubMed: 18688002]

6. Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified rankin scale. Stroke 2009;40:762–766. [PubMed: 19131664]

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. [PubMed: 12215594]

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. [PubMed: 17600236]

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. [PubMed: 15718510]

10. Quinn T, Dawson J, Walters M, Lees K. Reliability of the modified rankin stroke scale – a systematic review. Stroke. In Press.

11. Rankin J. Cerebral vascular accidents inpatients over the age of 60, ii: Prognosis. Scott Med J 1957;2:200–215. [PubMed: 13432835]

12. Calculation of sample size for stroke trials assessing functional outcome: Comparison of binary and ordinal approaches. Int J Stroke 2008;3:78–84. [PubMed: 18705999]

13. Saver JL, Gornbein J. Treatment effects for which shift or binary analyses are advantageous in acute stroke trials. Neurology 2009;72:1310–1315. [PubMed: 19092107]

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. [PubMed: 19498197]

Table 1

Cross-tabulation 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

Inter-Rater Agreement for All Dichotomizations of the Modified Rankin Scale

Rankin Cutpoint	Observed Agreement	Kappa
0 vs 1–6	98.0%	0.93
0–1 vs 2–6	100%	1.00
0–2 vs 3–6	98.0%	0.96
0–3 vs 4–6	100%	1.00
0–4 vs 5–6	98.0%	0.96
0–5 vs 6	100%	1.00