http://eprints.gla.ac.uk/17550/

Deposited on: 18 January 2012

Exploring the Reliability of the Modified Rankin Scale

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 prevalent outcome measure in stroke trials. Use of the mRS may be hampered by variability in grading. Previous estimates of the properties of the mRS have used diverse methodologies and may not apply to contemporary trial populations. We used a mock clinical trial design to explore inter- and intraobserver variability of the mRS.

Methods—Consenting patients with stroke attending for outpatient review had the mRS performed by 2 independent assessors with pairs of assessors selected from a team of 3 research nurses and 4 stroke physicians. Before formal assessment, interviewers estimated disability based only on initial patient observation. Each patient was then randomized to undergo the mRS using standard assessment or a prespecified structured interview. The second interviewer in the pair reassessed the patient using the same method blinded to the colleague's score. For each patient assessed, one rater was randomly assigned to video record their interview. After 3 months, this interviewer reviewed and regraded their original video assessment.

Results—Across 100 paired assessments, interobserver agreement was moderate ($k=0.57$). Intraobserver variability was good ($k=0.72$) but less than would be expected from previous literature. Forty-nine assessments were performed using the structured interview approach with no significant difference between structured and standard mRS. Researchers were unable to reliably predict mRS from initial limited patient assessment ($k=0.16$).

Conclusions—Despite availability of training and structured interview, there remains substantial interobserver variability in mRS grades awarded even by experienced researchers. Additional methods to improve mRS reliability are required. (*Stroke*. 2009;40:762-766.)

Key Words: clinical trials ■ clinometrics ■ modified Rankin Scale ■ outcome assessment
■ stroke ■ stroke treatment ■ video recording

I n stroke trials and in clinical work, robust measures of patient outcome are required. The modified Rankin Scale (mRS)[1] is the most prevalent functional outcome measure in contemporary stroke trials and has been used in several landmark studies.[2,3] The mRS quantifies disability using an ordinal hierarchical grading from zero (no symptoms) to 5 (severe disability) with some adding a category of mRS 6 (death).

Before any clinical use, assessment scales should be proven "fit for purpose." Clinometrics is the methodological discipline that describes clinical measurement quality. Outcome scales are traditionally assessed in terms of responsiveness, validity, and reliability.[4] The original Rankin Scale was not designed for clinical trial use and, like many other stroke assessment tools, the mRS became established as a study end point before any formal clinometric assessment.[5]

Recent studies have quantified responsiveness of the mRS[6] and proven excellent construct and convergent validity of the scale.[7,8] Estimates of mRS reliability have been less favorable with authors describing substantial interobserver variability.[9] This is a concern, because arguably for an instrument that will be used by many hundreds of raters in large-scale multicenter clinical trials, reliability is an important property of the scale.

Although all studies to date have confirmed poor reliability, there has been heterogeneity in the degree of variability described.[1,10] Differing methodologies used to assess the mRS may partly explain this. Because we are principally concerned with variability present in clinical trials, the most informative analysis of the mRS would be conducted using current researchers working in a clinical trial setting and interviewing real stroke survivors. Those few studies that have attempted this design used only limited numbers of assessors and/or patients.[1,10]

In recognition of this potential for variability, techniques to improve mRS assessment, including structured interview[10] and video training,[11] have been developed. Studies of the structured interview approach have yielded conflicting results,[10,12] although again limited numbers of patients and assessors compromise the generalizability of these reports.

Thus, there are several unanswered questions regarding reliability of mRS assessment. Using a mock clinical trial design, we set out to study the inter- and intraobserver variability
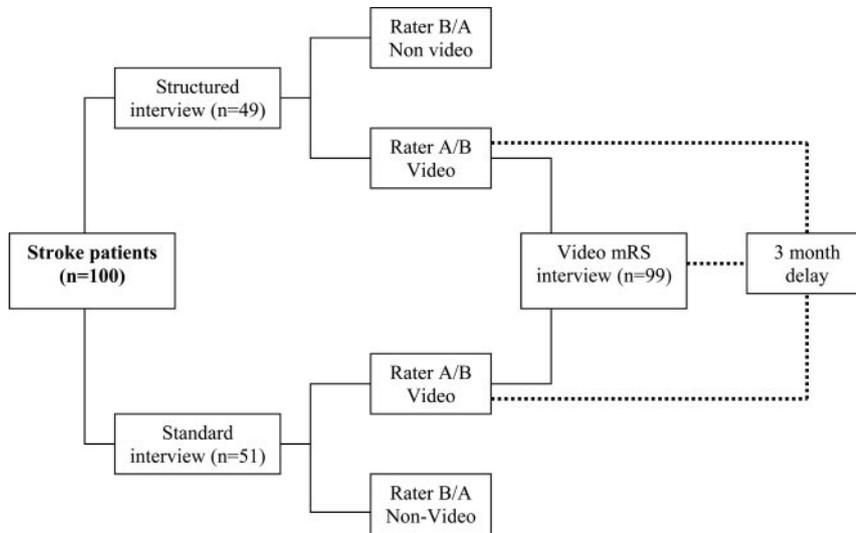
Figure. Schematic diagram of evaluation process for mRS reliability assessment. A–B, trained mRS assessors selected from a pool of 7 research nurses/stroke physicians. Order of video interview (first or second in pair) and designation to A or B were determined by randomization. At 3 months, video-recorded assessments were reviewed by the original interviewer and graded again.

when the mRS is applied by experienced and trained study personnel. We further describe the effect of a structured interview format on properties of the mRS. Finally, recognizing that initial clinical judgment often influences final scoring in assessment scales, we described the ability of researchers to estimate disability from limited review before formal mRS assessment.

Materials and Methods

Patients and Assessors

We approached sequential patients attending our university hospital cerebrovascular clinic for their routine poststroke assessment. Clinic patients have usually been inpatients in the local acute stroke unit and typically attend for review at 90 days poststroke; however, we did not set fixed time-related or geographic exclusion criteria. All patients were considered for inclusion. If cognitive impairment or language problems precluded satisfactory mRS interview, a proxy (family member or caregiver) was used. Informed consent was given by all participants or designated proxy before recruitment and reconfirmed after the assessment. The local ethics committee approved the study protocol.

To allow assessment of variability across a representative group, we involved 7 assessors: 4 stroke physicians and 3 research nurses. All assessors had been trained and certified in mRS assessment[10] and have considerable experience in mRS application in clinical trials.

We used a stratified assessment technique to test our related hypotheses (Figure). The selection of mRS assessors, interview methodology used, and selection for interview recording were all prespecified using an online randomization program (www.random. org/integers), and allocation was concealed from interviewers and patients using an opaque envelope system.

Statistical Analysis

Reliability, or agreement between observers, is traditionally described with kappa (k) statistics. The k statistic can be seen as a measure of agreement, above that expected by chance alone, with defined standard values of poor ($k=0.00$ 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.[13]

Kappa statistics were chosen for primary analysis because clinicians are familiar with the test and because previous studies of mRS reliability have used similar statistical techniques.

Formal comparisons between kappa statistics are problematic, particularly if numbers in each group are not comparable. Therefore, to allow for basic comparative analysis, we also calculated the number of interviews in which rater pairs agreed exactly on mRS

(expressed as percentage agreement) and compared values using χ^2 testing. Specific analyses performed for each hypothesis are detailed in the relevant subsections. All statistics were performed using Minitab software (Version 14.0; Minitab Inc).

Interobserver Variability for the Traditional Modified Rankin Scale

For each patient enrolled, 2 assessors allocated from our pool of 7 performed mRS grading. Interviews were performed using a standard mRS approach or a structured interview with choice of methodology randomly allocated. Thus, patients had 2 independent assessments in succession, each using the same interview methodology (structured or standard mRS) and blinded to colleagues' grading. We used the previously validated, questionnaire-style interview for the structured assessments as originally described by Wilson et al[9] with roughly half the assessments conducted using this structured interview approach.

This initial series of face-to-face paired interviews are further referred to as "traditional mRS." We measured interobserver agreement between the paired mRS assessments, first for all interviews and then with subanalysis to compare structured interview against standard mRS. We further evaluated duration of the interview for the structured and standard mRS interview using paired Student t testing.

Intraobserver Variability

One researcher from each interview pair was randomly selected for video recording. Following advice from the Media Services Department, University of Glasgow, audiovisual recording was captured using a portable digital camera (HDVR-HC1E 1080i digital HD video camera recorder; Sony) and stored on digital video disc using readily available image processing software (Windows Movie Maker; Microsoft).

At a later date, the interviewer who performed the original mRS assessment viewed this recording and rescored the mRS. We left a minimum 3-month delay between the interview and assessment of recording to reduce recall bias. Repeat scoring was performed independently and raters had no access to their previous scores. Assessment of intraobserver variability was made comparing all raters' original, traditional mRS score with their subsequent video review score.

Estimating the Modified Rankin Scale

To gauge the added value of formal interview, raters were asked to grade disability before beginning their formal mRS assessment. This meant assigning a preliminary mRS using only such information as would be available in the first few seconds of patient interaction, for example assessment of patient mobility in the consulting room or

Table 1. Group Reliability of Traditional mRS Assessment*

mRS	Average No. of Cases at This mRS	Variability (k) of Traditional mRS
0	16	0.70
1	16	0.43
2	33.5	0.51
3	16.5	0.57
4	15.5	0.66
5	2.5	0.79
Overall k	N/A	0.57
Structured interview k	N/A	0.5
Standard interview k	N/A	0.64

*To quantify spread of disability, the average no. of patients scoring at each mRS grade is presented; these numbers were derived from the total no. of scores at a particular grade divided by the total no. of assessors.

N/A indicates not applicable.

Table 2. Interobserver Variability in mRS Comparing Our Study With Other Published Data*

	mRS Variability (% Agreement)
Quinn—total mRS	k=0.57 (67%)
Standard mRS	k=0.64 (72%)
Structured mRS	k=0.50 (63%)
Newcommon—standard mRS	k=0.72 (N/A)
Structured mRS	k=0.38 (N/A)
Wilson '05—standard mRS	k=0.25 (43%)
Structured mRS	k=0.74 (81%)
Wilson '02—standard mRS	k=0.44 (57%)
Structured mRS	k=0.70 (78%)
Wolfe—standard mRS	kw=0.90 (80%)
Van Sweiten—standard mRS	k=0.56 (65%)

*Variability is described using kappa (k) and percentage exact agreement; there were no significant differences in percent agreement between studies.

N/A indicates not available.

interaction with nurses or caregivers and initial conversation. This score was recorded and sealed in an opaque envelope. Raters then conducted and scored the formal mRS assessment. Properties of the preliminary mRS score were described by comparing these estimates with the final mRS and by describing variability within the estimated scores.

Comparison With Previous Literature
To place our results in context, we extracted data from previous studies of mRS reliability and performed comparative analyses. We used 2 recently conducted systematic reviews[14,15] combined with our own literature search to select articles that reported on either inter- or intraobserver reliability of mRS. When possible, we extrapolated data on variability as measured by kappa statistics and percentage agreement rates. We compared results from our mRS study with those published using χ^2 analysis of proportions.

Results
Of 104 patients approached, 102 consented to mRS interview and video recording. Of these, 100 video recordings were of sufficient technical quality to allow repeat grading and were included in the final analysis. Patients reflected a heterogeneous group of stroke subtypes typical of 3-month survivors (total anterior circulatory stroke, 16; partial anterior circulatory stroke, 30; lacunar stroke, 43; posterior circulation stroke, 11; unclassified, 2). Mean age was 69.8 years (SD, 12.9); mean National Institutes of Health Stroke Scale score at baseline was 5.5 (SD, 5.2) and median time since the event was 12 weeks (interquartile range, 6 to 21). Five patients had problems with communication such that assessment involved a proxy.

Interobserver Variability for the Traditional Modified Rankin Scale
Variability in traditional mRS grading was moderate (k=0.57) for the group of 100 paired interviews with least variability at extremes of mRS (Table 1). Exact agreement in mRS was 67%; this was not significantly different from data from previously published studies (P=0.073; Table 2).

Of the traditional mRS assessments, 49 used a structured interview approach. There was no difference in spread of disability as graded on mRS between the 2 groups (P=0.699 on χ^2 testing). Use of the structured interview did not

decrease variability (P=0.295; Tables 1 and 2). Mean duration of mRS assessment was 4.9 minutes (SD, 2.4). There was a significant difference between duration of structured (5.6 minutes; SD, 2.5) and unstructured (4.2 minutes; SD, 2.1) interviews (P=0.003).

Intraobserver Variability for the Modified Rankin Scale
One patient withdrew consent for video assessment after recording, leaving 99 video assessments that could be reviewed and scored by the original mRS assessors. Intraobserver reliability was good for the group (k=0.72; 77% complete agreement; Table 3); this differs significantly from other published studies of mRS intraobserver variability (P<0.0001). Intraobserver variability for individual raters was calculated; percentage agreement was similar for most raters (Rater 1: 86%; Rater 2: 89%; Rater 3: 75%; Rater 4: 40%; Rater 5: 63%; Rater 6: 100%; Rater 7: 91%). Differences in numbers of patients assessed by each rater precludes formal statistical comparison.

Estimating the Modified Rankin Scale
A convenience sample of preliminary mRS interviews was included. Because estimation of mRS is dependent on confidence in basic mRS administration, we included only the latter 40 mRS interviews in this analysis to eliminate any potential training effect. Agreement between estimated and final mRS was 38% and reliability was poor (k=0.16). The

Table 3. Intraobserver Variability in mRS Comparing Our Study With Other Published Data*

	mRS Variability (% Agreement)
Quinn	k=0.72 (77%)
Wolfe	kw=0.95 (86%)
Wilson '05	k=0.83 (92%)†

*Variability is described using kappa (k) and percentage exact agreement.
†Average across structured and standard interview approaches.

mean estimated mRS was 1. (SD, 1.1); mean final mRS was 1.6 (SD, 0.9). Comparing estimated scores between the paired assessors, there was again poor agreement in 30% and significant variability ($k=0.38$).

Discussion

Using a mock clinical trial design, we assessed reliability of the mRS across a large number of patients. We have demonstrated substantial interobserver and intraobserver variability in mRS assessment. Furthermore, we have found that a structured interview approach does not significantly improve reliability and we confirmed that researchers are poor at estimating mRS if they do not conduct a formal interview.

Despite considerable experience in clinical use of the mRS, our team of clinicians and nurses show only moderate reliability in mRS grading. This interobserver variability is in keeping with previous published estimates.[1,10] With increasing use of the mRS as a trial end point[16] and ready availability of specific training resources,[11] some improvement in mRS reliability was expected. Diverse study methodologies preclude any more definitive comment on these differences; suffice to say that problems with reliability represent an ongoing limitation of standard mRS as a trial end point. In the absence of a pretraining "control," our findings do not allow us to comment on usefulness of the training resource or on any training effects associated with increasing experience of real-life mRS administration.

Variability was most apparent for mRS grades 1 to 4. This is of particular importance for clinical trial end point analysis, in which mRS outcomes are often dichotomized around these middle grades. Misclassification of end points increases the likelihood of Type II error and decreases statistical power. The potential impact of mRS variability on clinical trial results has yet to be modeled, but we must assume that poor reliability will influence final results. Real-life examples of trials compromised by variability in end point classification are well recognized[17] and may be particularly relevant in the field of acute stroke, in which some have argued that recent unexpected neutral trial outcomes have stemmed from underpowering in trial design.[18]

Quantification of intraobserver reliability for clinical scales is challenging and if methodology is poor, there is potential for bias. Measuring test–retest variability over a short time period will be biased by observer recall of previous grading; delaying the second grading can allow for patient improvement or disease progression. Previous published studies have not accounted for these sources of bias in their design and, as such, the negligible interobserver variability they report for mRS should be questioned.[8,9] Our use of videos provides a more rigorous assessment of intraobserver reliability and may explain the significantly higher variability demonstrated. Because trialists are unlikely to be performing serial mRS over short time periods, it could be argued that proving intraobserver variability of the mRS is of little clinical relevance. However, we describe our findings here as further evidence of the imperfections of the standard mRS as an end point assessment tool.

Use of a structured interview approach to mRS assessment did not reduce interobserver variability in our cohort. The authors of one questionnaire-style structured interview previously reported significant improvements in reliability[9]; however, other groups have failed to replicate these findings[12] and, at present, the structured interview is infrequently used by stroke trialists. Our results show that for experienced raters, fully trained in mRS administration, use of a structured approach may have little to add. The difference in interview duration between the traditional and structured approach with no improvements in reliability suggests that certain components of the structured interview may be redundant.

Our final analysis described efficacy of initial limited disability assessment as a predictor of final mRS grading. Such an approach is not without precedent. It is recognized that for many scales, raters may not perform a comprehensive assessment; rather, they will estimate final grading based on initial basic review and "clinical intuition."[19] For disability scales, including the mRS, full assessment has been reduced to a limited number of key questions while preserving clinometric properties.[20] The mRS is heavily weighted toward locomotor independence and so we hypothesized that distinction between higher and lower grades may be possible simply by observing the patient entering the clinic. We have shown that experienced raters are poor at predicting the final mRS from initial assessment and that a formal interview is still required to grade disability.

A particular strength of our study was the mock clinical trial design simulating those situations in which the mRS is likely to be used. We adopted an inclusive policy, studying a large representative cohort of stroke survivors. We deliberately selected a panel of assessors from different clinical backgrounds because previous work has suggested that profession and training may impact on reliability of outcomes assessment.[21] Limited numbers of patients and use of assessors from similar backgrounds have compromised previous studies of mRS reliability.[1,10] The use of video recording to assess intraobserver variability was successful with minimal expenditure in terms of money and training. Other centers have also demonstrated efficacy of remote video-based mRS assessment.[22] These results suggest feasibility of remote video-based mRS assessment as a further aid to improve reliability.

Although number of patients included in our analysis is greater than in many previous studies of mRS, we still had relatively few assessors and all were from the same department. Ideally, we would have involved multiple centers in our analysis. In this regard, our study is complemented by recent work describing moderate to good overall reliability on a 5-patient mRS assessment exercise across a large cohort of international trialists.[21]

We deliberately chose to test a number of related hypotheses using a predefined structured design, thus deriving substantial data from a single clinical encounter. However, we prespecified our several hypotheses to limit the risk of drawing false conclusions as a result of multiplicity. Our results do not negate the potential benefit of training and we would encourage trialists to continue to use specific mRS training resources. Future trials designed to improve mRS assessment are planned; pending these results, our current data encourage caution in use and interpretation of the

standard mRS. Because measures to date have not substantially improved mRS interobserver variability, a possible option for future trials is to limit the number of observers. Remote adjudication panel assessment of laboratory and imaging end points is commonplace in contemporary multicenter trials and perhaps should now become routine for assessment of functional outcomes.

In conclusion, we have shown that despite increasing familiarity with the mRS and availability of specific training packages, there remains substantial variability in the mRS that could compromise clinical trial results. Further measures to improve mRS reliability are urgently required.

Acknowledgments

Disclosures

References

1. van Swieten Koudstaal PJ, Visser MC, Schouten HJA, Gijn JV. Inter-observer agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604–607.
2. The European Cooperative Acute Stroke Study Group. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. *JAMA*. 1995;274:1017–1025.
3. Shuaib A, Lees KR, Lyden P, Grotta J, Davalos A, Davis SM, Diener HC, Ashwood T, Wasiewski WW, Emeribe U. SAINT II Trial Investigators. NXY-059 for the treatment of acute ischemic stroke. *N Engl J Med*. 2007;357:562–571.
4. de Vet HC, Terwee CB, Bouter LM. Current challenges in clinometrics. *J Clin Epidemiol*. 2003;56:1137–1141.
5. Quinn TJ, Dawson J, Walters M. Dr John Rankin; his life, legacy and the 50th anniversary of the Rankin Stroke Scale. *Scot Med J*. 2007;52:44–47.
6. Kwon S, Hartzema AG, Duncan PW, Min-Lai S. Disability measures in stroke: relationship among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale. *Stroke*. 2004;35: 918–923.
7. Quinn TJ, Dawson J, Lees JS, Chang TP, Walters MR, Lees KR, for the GAIN and VISTA Investigators. Time spent at home poststroke: 'home-time' a meaningful and robust outcome measure for stroke trials. *Stroke*. 2008;39:231–233.
8. Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ. Assessment of scales of disability and handicap for stroke patients. *Stroke*. 1991;22:1242–1244.
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. Wilson JT, Hareendran A, Grant M, Baird T, Schulz UG, 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.
11. Quinn TJ, Lees KR, Hardemark HG, Dawson J, Walters MR. Initial experiences of a digital training resource for modified Rankin scale assessment in clinical trials. *Stroke*. 2007;38:2257–2261.
12. Newcommon NJ, Green TL, Hayley 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.
13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
14. New PW, Buchbinder R. Critical appraisal and review of the Rankin Scale and its derivatives. *Neuroepidemiology*. 2006;26:4–15.
15. 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.
16. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials [Abstract]. *Stroke*. 2008;39:692.
17. Jaffar S, Leach A, Smith PG, Cutts F, Greenwood B. Effects of misclassification of causes of death on the power of a trial to assess the efficacy of a pneumococcal conjugate vaccine in the Gambia. *Int J Epidemiol*. 2003;32:430–436.
18. Fisher M, Albers GW, Donnan GA, Furlan AJ, Grotta JC, Kidwell CS, Sacco RL, Wechsler LR. Stroke Therapy Academic Industry Roundtable IV. Enhancing the development and approval of acute stroke therapies: Stroke Therapy Academic Industry roundtable. *Stroke*. 2005;36: 1808–1813.
19. Burleigh E, Reeves I, McAlpine C, Davie J. Can doctors predict patients' abbreviated mental test scores? *Age Ageing*. 2002;31:303–306.
20. Berge E, Fjaertoft H, Indredavik B, Sandset PM. Validity and reliability of simple questions in assessing short- and long-term outcome in Norwegian stroke patients. *Cerebrovasc Dis*. 2001;11:305–310.
21. Quinn TJ, Dawson J, Walters MR, Lees KR. Variability in modified Rankin scoring across a large cohort of observers [Abstract]. *Stroke*. 2008;39:692.
22. Shinohara Y, Minematsu K, Amano T, Ohashi Y. Modified Rankin Scale with expanded guidance scheme and interview questionnaire: inter-rater agreement and reproducibility of assessment. *Cerebrovasc Dis*. 2006;21: 271–278.