![University of Glasgow logo]

http://eprints.gla.ac.uk/18438/

Deposited on: 18 January 2012

Reliability of the Modified Rankin Scale : A Systematic Review
Terence J. Quinn, Jesse Dawson, Matthew R. Walters and Kennedy R. Lees

The online version of this article, along with updated information and services, is
located on the World Wide Web at:

http://stroke.ahajournals.org/content/40/10/3393

Reliability of the Modified Rankin Scale
A Systematic Review

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

Background and Purpose—A perceived weakness of the modified Rankin Scale is potential for interobserver variability. We undertook a systematic review of modified Rankin Scale reliability studies.

Methods—Two researchers independently reviewed the literature. Crossdisciplinary electronic databases were interrogated using the following key words: Stroke*; Cerebrovasc*; Modified Rankin*; Rankin Scale*; Oxford Handicap*; Observer variation*. Data were extracted according to prespecified criteria with decisions on inclusion by consensus.

Results—From 3461 titles, 10 studies (587 patients) were included. Reliability of modified Rankin Scale varied from weighted $\kappa=0.95$ to $\kappa=0.25$. Overall reliability of mRS was $\kappa=0.46$; weighted $\kappa=0.90$ (traditional modified Rankin Scale) and $\kappa=0.62$; weighted $\kappa=0.87$ (structured interview).

Conclusion—There remains uncertainty regarding modified Rankin Scale reliability. Interobserver studies closest in design to large-scale clinical trials demonstrate potentially significant interobserver variability. (*Stroke*. 2009;40:3393-3395.)

Key Words: clinical trials ■ clinometrics ■ modified Rankin Scales ■ outcome assessment ■ systematic review

The modified Rankin Scale (mRS) is the most prevalent functional outcome measure in contemporary stroke research.[1] A weakness of mRS grading is the potential for interobserver variability. Variability implies end point misclassification and can weaken statistical power.[1] Various attempts to quantify mRS reliability have been reported.

Clinical trial use of the mRS is global[2] and often used by research nurses and professions allied to medicine.[2] A contemporary, systematic review of the international literature, including allied healthcare journals, was performed.

Methodology

Two clinical researchers independently reviewed the literature. Throughout the process, we adhered to Meta-analysis of Observational Studies in Epidemiology (MOOSE) guidelines for meta-analysis.[3]

Participants

Study populations included human stroke survivors only. We used no restrictions for mRS assessor.

Study Methodology

All studies purporting to measure mRS reliability through patient interview (inter-/intraobserver variability) were reviewed.

Bias and Trial Quality

As a minimum data set to allow assessment of trial quality, we collected details on: patient selection; blinding of mRS interviewers (to others scores); mRS methodology; number of patients/observers; details of withdrawals; and dropouts.

Outcomes

No restriction on the basis of mRS assessment methodology was applied. Studies using mRS derivatives (Rankin Scale and Oxford Handicap Scale) were reviewed for references but not included.

Search Strategy

A comprehensive battery of crossdiscipline electronic databases were interrogated: AMED; British Nursing Index; CINAHL; Embase; Health and Psychosocial Instruments; Internurse.com; Medline; and PsychINFO (all inception to December 2008). Key words were formulated using MeSH headings and designed to be as inclusive as possible (Figure).

To identify studies not yet in print, proceedings of scientific meetings were hand-searched: International Stroke Conference; European Stroke Conference; and World Stroke Congress (January 2006 to November 2008). Bibliographies of retrieved articles were searched for further references and the process repeated until no new articles were found.

We retrieved full text of articles that either reviewer suspected may be relevant. Data were extracted according to prespecified criteria. Decisions on inclusion were by consensus. Where potentially relevant data were not published, electronic contact with original authors was attempted. For studies not published in English, professional translation services were used.

Statistics

To allow for comparison and where data permitted, we described results using κ; quadratic weighted κ, and percentage agreement. Based on previous work suggesting a beneficial effect of a structured interview approach,[4] we performed separate analysis comparing "structured" and "traditional" mRS. A one-group descriptive study using average absolute difference with a fixed effects model was performed using MIX software Version 1.7 (www.mix-for-meta-analysis.info).

Search terms

stroke* OR cerebrovasc*   observer variation*
AND   AND
modified Rankin* OR Rankin* OR Oxford Handicap*

Figure. Profile of the systematic review.

Table 1. Reliability of Traditional mRS

Study	κ*	Weighted κ*	Agreement, %†
van Swieten, 1988[8]	0.56 (0.45–0.68)	0.91 (0.71–1.00)	65%
Wolfe, 1991	N/A	0.87 (0.84–0.97)	80%
Berger, 1999[9]	0.56 (0.41–0.71)	0.88 (0.58–1.00)	N/A
Wilson, 2002[10]	0.44 (0.29–0.62)	0.78 (0.53–1.00)	57%
Newcommon, 2003[11]	0.72 (0.55–0.89)	N/A	N/A
Wilson, 2005	0.25 (0.16–0.35)	0.71 (0.53–0.88)	43%
de Caneda, 2006[12]	0.45 (0.31–0.60)	0.45 (0.58–0.90)	N/A
Gur, 2007[13]	N/A	0.95 (0.89–1.00)	N/A
Meyer, 2008[14]	N/A	0.90 (0.59–1.00)	N/A
Quinn, 2009	0.64 (0.48–0.79)	0.91 (0.65–1.00)	72%
Totals	0.46 (0.41–0.51)	0.90 (0.86–0.94)	71%

*95% CIs in parentheses.
†Agreement between observers.
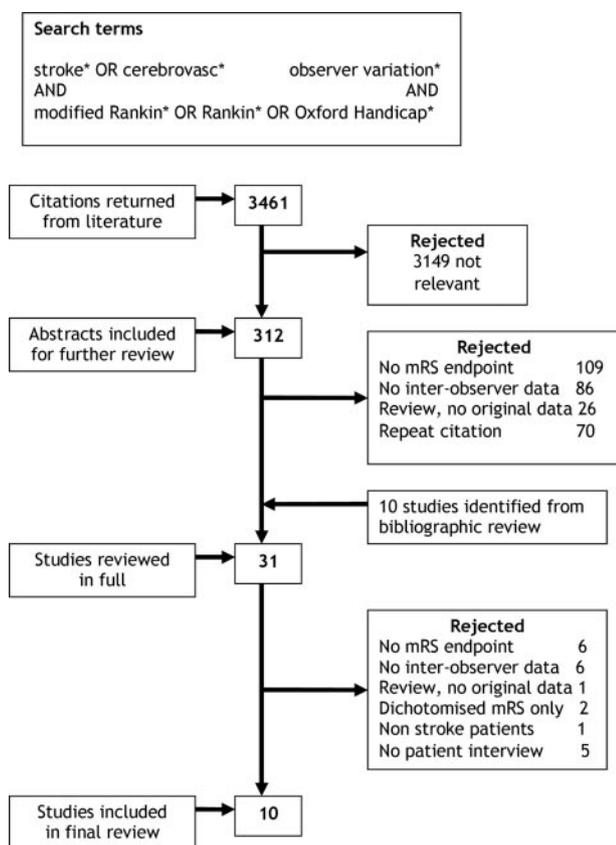NA indicates not available.

Results

The review profile is detailed in the Figure. Ten studies involving 587 patients were included in the final analysis (reference list available as supplemental data; available at http://stroke.ahajournals.org). Median number of included patients was 47; median number of researchers performing mRS was 2.

Interobserver variability of mRS varied from "near perfect" (weighted κ=0.95) to "poor" (κ=0.25). Overall, reliability was "moderate" for the 2 approaches to mRS (Table 1 and 2). Three studies[4–6] (162 patients) measured intraobserver variability of mRS. Overall reliability was very good with combined weighted κ=0.94 and percentage agreement 84%.

In the included studies, diverse methodologies were used to administer and study mRS (Supplemental Table, available online at http://stroke.ahajournals.org). No study met our "minimum" criteria to allow assessment of quality: no description of patient selection (n=5); no data on blinding (n=5); inadequate description of mRS methodology (n=2); and no description of location/timing of mRS (n=2). As a result, we included all relevant studies regardless of methodological quality.

Discussion

We have demonstrated that overall reliability of mRS is moderate, but there remains potential for improvement. The effect of structured interview on reliability remains unproven; apparent benefits seen with combined κ analysis

were lost when "weighted" κs were applied. We should be cautious in interpreting these data; only 4 studies purported to examine the structured approach and almost two thirds of the data came from studies performed by the authors of the original structured interview. The nonparametric nature of κ does not allow for comparative meta-analysis and so the safest conclusion is that structuring mRS may partly improve mRS reliability, but effects have not been consistent.

It is interesting that studies with larger numbers of patients and observers reported poorer reliability. The importance of interobserver reliability in clinical trials becomes apparent when the number of potential end point assessors is considered. In the recent Stroke-Acute Ischemic NXY Treatment (SAINT) trials, >1000 assessors from 25 countries were trained in outcome assessment.[7] We have shown that numbers of assessors included in mRS reliability studies are considerably smaller. The ideal methodology to assess mRS reliability would involve observers of differing backgrounds and from differing international centers. Only one study approaches this "ideal" and it reports a concerningly low reliability for standard mRS.[4] We should note that all included studies measured reliability across mRS 0 to 5. For clinical trials, a grade of mRS 6 (death) is often added; addition of this objective end point may improve overall reliability.

Table 2. Reliability of mRS Using a Structured Interview Approach

Study	κ*	Weighted κ*	Agreement, %†
Wilson, 2002[10]	0.70 (0.56–00.85)	0.93 (0.67–1.00)	78%
Newcommon, 2003[11]	0.34 (0.17–0.55)	N/A	50%
Wilson, 2005	0.74 (0.64–0.84)	0.91 (0.73–1.00)	81%
Quinn, 2009	0.50 (0.34–0.68)	0.74 (0.455–1.00)	63%
Totals	0.62 (0.56–0.69)	0.87 (0.75–1.00)	73%

*95% CIs in parentheses.
†Agreement between observers.
NA indicates not available.

Intraobserver studies suggested excellent reliability. However, 2 studies measured mRS at distinct periods in the patient's recovery and thus are prone to recall bias and potential for functional improvement between assessments.[4,5] Although theoretically interesting, intraobserver variability may be of less relevance to clinical trials, in which primary outcome assessment is performed once only.

There was heterogeneity between studies in several important aspects of methodology (Supplemental Table). As an example, only one study[6] made use of the recognized mRS training resource.[7] This study reported no beneficial effect of structured interview, perhaps suggesting that structured approach is unnecessary if assessors are adequately trained. Collation of studies with differences in methodology potentially weakens our meta-analysis but is perhaps necessary; recent review of mRS reported substantial heterogeneity in its application.[1] Quality of studies varied and for all those included, certain data were incomplete. Again this weakens our analysis because we were unable to exclude potentially biased studies.

Heterogeneity was further evident in the statistical methods used in the studies: κ; weighted κ; intraclass correlation coefficient; and percentage agreement were all used. All are appropriate and there is no accepted optimal test. However, the resultant data were not readily interchangeable and offered limited potential for meta-analysis without access to original individual patient data. Our own use of statistics demands discussion. No universally accepted analysis method for multiple κ statistics has been described. Recognizing this limitation, we used a group analysis technique that made the fewest assumptions of the underlying data.

Accepting these limitations, our study does have certain strengths. Literature searching was comprehensive and systematic, considering reports from non-English and "nonmedical" sources and excluding studies that would not help describe mRS reliability in a clinical trial setting.

There remains uncertainty regarding mRS reliability. Available studies are likely underpowered and have design flaws that limit generalization. Studies closest in design to large-scale trials demonstrate potentially significant variability. Although we await definitive data on mRS reliability, we must acknowledge a degree of interobserver variability is inherent in mRS.

Acknowledgments

Disclosures

References

1. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. *Int J Stroke*. 2009;4:200–206.
2. 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.
3. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB for the Meta-analysis of Observational Studies in Epidemiology (MOOSE) Group. Meta-analysis of Observational Studies in Epidemiology: a proposal for reporting. *JAMA*. 2000;283:2008–2012.
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. Wolfe CDA, Taub NA, Woodrow EJ, Burney PGJ. Assessment of scales of disability and handicap for stroke patients. *Stroke*. 1991;22:1242–1244.
6. Quinn TJ, Dawson J, Walters MR, Lees KR. Exploring the reliability of the modified Rankin Scale. *Stroke*. 2009;40:762–766.
7. 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.
8. 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.
9. Berger K, Weltermann B, Kolominsky-Rabas P, Meves S, Heuschmann P, Böhner J, Neundörfer B, Hense HW, Büttner T. The reliability of stroke scales: The German version of NIHSS, ESS and Rankin scales. *Fortschritte der Neurologie-Psychiatrie*. 1999;67:81–93.
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. 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.
12. de Caneda, MAG, Fernandes JG, de Almeida AG, Mugnol FE. Reliability of neurological assessment scales in patients with stroke. *Arq Neuro-Psiquiatr*. 2006;64:690–697.
13. Gur AY, Lampl Y, Gross B, Royter V, Shopin L, Bornstein NM. A new scale for assessing patients with vertebrobasilar stroke-the Israeli Vertebrobasilar Stroke Scale (IVBSS): inter-rater reliability and concurrent validity. *Clinical Neurology & Neurosurgery*. 2007;109:317–322.
14. Meyer BC, Raman R, Chacon MR, Jensen M, Werner JD. Reliability of site-independent telemedicine when assessed by telemedicine-naive stroke practitioners. *Journal of Stroke & Cerebrovascular Diseases*. 2008;17:181–186.