Reliability of the Modified Rankin Scale Across Multiple Raters
Benefits of a Structured Interview

J.T. Lindsay Wilson, PhD; Asha Hareendran, PhD; Anne Hendry, FRCP; Jan Potter, FRCP; Ian Bone, MD; Keith W. Muir, MD

Background and Purpose—The modified Rankin Scale (mRS) is widely used to assess global outcome after stroke. The aim of the study was to examine rater variability in assessing functional outcomes using the conventional mRS, and to investigate whether use of a structured interview (mRS-SI) reduced this variability.

Methods—Inter-rater agreement was studied among raters from 3 stroke centers. Fifteen raters were recruited who were experienced in stroke care but came from a variety of professional backgrounds. Patients at least 6 months after stroke were first assessed using conventional mRS definitions. After completion of initial mRS assessments, raters underwent training in the use of a structured interview, and patients were re-assessed. In a separate component of the study, intrarater variability was studied using 2 raters who performed repeat assessments using the mRS and the mRS-SI. The design of the latter part of the study also allowed investigation of possible improvement in rater agreement caused by repetition of the assessments. Agreement was measured using the κ statistic (unweighted and weighted using quadratic weights).

Results—Inter-rater reliability: Pairs of raters assessed a total of 113 patients on the mRS and mRS-SI. For the mRS, overall agreement between raters was 43% (κ=0.25, κw=0.71), and for the structured interview overall agreement was 81% (κ=0.74, κw=0.91). Agreement between raters was significantly greater on the mRS-SI than the mRS (P<0.001). Intrarater reliability: Repeatability of both the mRS and mRS-SI was excellent (κ=0.81, κw=0.94).

Conclusions—Although individual raters are consistent in their use of the mRS, inter-rater variability is nonetheless substantial. Rater variability on the mRS is thus particularly problematic for studies involving multiple raters. There was no evidence that improvement in inter-rater agreement occurred simply with repetition of the assessment. Use of a structured interview improves agreement between raters in the assessment of global outcome after stroke. (Stroke. 2005; 36:777-781.)

Key Words: clinical trials ■ disability evaluation ■ outcome

The modified Rankin Scale (mRS) is the most popular assessment of global outcome in stroke2-3 and is increasingly being adopted as a primary endpoint in clinical trials in acute stroke, either on its own or as a component of the global endpoint using the common odds ratio methodology.4 The scale describes 6 grades of disability after a stroke (grade 5 denotes severe disability, bedridden; and grade 0 denotes no symptoms at all), and a particular strength of the mRS is its ability to capture the full spectrum of limitations in activity and participation after stroke. Although the inter-rater reliability of the mRS meets criteria for a satisfactory clinical assessment, there may still be room for improvement.2-5 The assignment of grades is open to variability in the interpretation of the descriptions by raters, and the kinds of information elicited from informants. Such differences are likely to be greater when multiple raters are involved in a study and when raters come from different professional backgrounds. Variability is of particular importance because misclassification on the endpoint reduces the power of a clinical trial and reduces the apparent size of any treatment benefit.6 Although loss of power can be compensated by an increase in the total number of cases recruited, differences between groups will remain diminished by rater variation, and consequently it may be difficult to demonstrate convincing clinical benefits of treatment.

Previous evaluations of inter-rater reliability of the mRS1,2,5,7,8 have used raters of similar background, usually from a single center,1,5,7,8 and may therefore have understi...
mated potential inter-rater variability. One way of investigating the importance of differences between raters is by comparing the conventional open-ended use of the mRS with a structured interview (SI). The SI for the mRS\(^1\) was developed to standardize the manner of administration of the scale while still allowing room for clinical judgment. It was developed based on input from patients and clinicians regarding aspects of disability after a stroke. We previously examined the inter-rater reliability of the mRS and SI (the mRS-SI), and concluded that using SI improved reliability.\(^1\)

The current study investigated rater variability in outcome assessment using the mRS. To simulate more closely the conditions of a multicenter study, we investigated inter-rater reliability when using numerous raters with different professional backgrounds. We also investigated the repeatability of assessments and studied the extent to which repetition effects might be contributing to improvement in agreement between raters.

**Subjects and Methods**

**Multicenter Inter-rater Agreement**

Fifteen raters from staff at 3 local stroke centers were recruited to represent a variety of professions involved in stroke care. They included 7 nurses, 4 physiotherapists, 2 neurologists, an occupational therapist, and a stroke physician. The conventional description of the mRS was given to raters.\(^2\) Staff who had not used the scale before performed assessments on 8 to 12 patients before commencing the study. During the practice assessments, raters also used the Barthel Index.\(^3\)

For the study, each patient was interviewed independently by 2 raters. After all assessments on the mRS had been completed, raters were trained in the use of the SI and the method of obtaining a Rankin grade.\(^4\) Training consisted of an explanation by the investigator, followed by a video showing an example interview that the raters scored, and 2 example transcripts that the raters scored. Each patient was then assessed again by the same 2 raters using the SI.

**Intrarater Agreement**

Repeatability of the mRS and the mRS-SI was investigated for 2 raters. The raters were nurses with experience of working in neurology but were without a specific background in stroke care and were not involved in the inter-rater study. A description of the mRS was given to the raters, and they each performed 12 assessments of stroke patients with the scale before commencing the study. During the practice interviews, raters also used the Barthel Index.

To investigate repeatability (test-retest reliability) of the mRS, each rater independently assessed patients on 2 occasions using the mRS (at an interval of 7 to 14 days). After all assessments had been completed, the raters were trained in the use of the SI and the procedure for grading patients on the mRS. They then rated another set of patients on 2 occasions (at an interval of 7 to 14 days).

Possible effects of repeat assessment were investigated by comparing inter-rater agreement on first interviews with inter-rater agreement on second interviews for each of the 2 parts of the intrarater agreement study. If repeat assessment improved consistency between raters, then this would be expressed as greater agreement on the second interviews than first interviews.

**Statistical Analysis**

Agreement between raters is analyzed using the \(\kappa\) statistic. This statistic corrects for agreement by chance, and a weighted value (\(\kappa_w\)) takes into consideration the size of disagreements. To allow comparison with previous studies, we give both the unweighted and weighted values of \(\kappa\), together with the 95% confidence intervals (CIs) of the values. The weighted value of \(\kappa\) (using quadratic weights) is equivalent to the intraclass correlation coefficient used for continuous measures.\(^10\) Conventional interpretation of the degree of agreement indicated by \(\kappa\) is: 0 to 0.20, “poor;” 0.21 to 0.40, “fair;” 0.41 to 0.60, “moderate;” 0.61 to 0.80, “good;” and 0.81 to 1.00, “very good.”\(^11\)

To compare methods of rating, we used appropriate nonparametric tests.

**Results**

**Multicenter Inter-rater Agreement**

Individual raters interviewed between 8 and 28 patients. A total of 117 stroke patients were assessed with the mRS, 113 (48 male) of whom returned for assessment using the SI. The mean age of patients assessed on both mRS and mRS-SI was 70 years (range, 30 to 92 years), and the mean interval after stroke on first assessment was 13.3 months (range, 6 to 24 months).

Ratings on the mRS and on the SI are shown in Tables 1 and 2. On the mRS, there was exact agreement between raters in only 43% of cases, and it was most common to disagree by one category (50% of cases). Table 1 shows that disagreement was greater at some category boundaries than others. For example, there was agreement for 88.5% of cases regarding whether patients had Rankin score 0 to 1 or less, whereas there was only agreement for 71% of cases regarding whether they had Rankin score 0 to 2 or less.

The unweighted kappa (\(\kappa\)) value for inter-rater agreement on the conventional mRS was 0.25 (95% CI, 0.16 to 0.35), and the weighted kappa value (\(\kappa_w\)) was 0.71 (95% CI, 0.53 to 0.83).

**TABLE 1. Agreement Between Raters on the mRS**

<table>
<thead>
<tr>
<th>Rater 2</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>15</td>
<td>18</td>
<td>1</td>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>12</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>17</td>
<td>30</td>
<td>19</td>
<td>5</td>
<td></td>
<td>113</td>
</tr>
</tbody>
</table>

Pairs of interviewers, arbitrarily designated in the Table as rater 1 or rater 2, classified each patient.

**TABLE 2. Agreement Between Raters on the Structured Interview**

<table>
<thead>
<tr>
<th>Rater 2</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td>13</td>
<td>6</td>
<td>1</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>19</td>
<td>1</td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>36</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>17</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>15</td>
<td>29</td>
<td>42</td>
<td>18</td>
<td>7</td>
<td>113</td>
</tr>
</tbody>
</table>

Patients previously rated on the mRS (Table 1) returned for assessment using a structured interview by the same pairs of independent interviewers.
0.88). On the SI, there was agreement in 81% of cases: $\kappa = 0.74$ (95% CI, 0.64 to 0.84) and $\kappa_w = 0.91$ (95% CI, 0.73 to 1.0). To compare the studies, we analyzed category differences between raters. Grades assigned by rater 1 were subtracted from grades assigned by rater 2, and the absolute values were computed. Comparison of the extent of disagreement between raters when using the mRS and mRS-SI showed that there was less disagreement using the SI (Wilcoxon $z = -5.35; P = 0.001$).

A comparison of the ratings made on the mRS in the first part of the study with the ratings made for the same patients using the mRS-SI is shown in Table 3. There was exact agreement in 48% of cases: $\kappa = 0.32$ (95% CI, 0.26 to 0.39) and $\kappa_w = 0.72$ (95% CI, 0.60 to 0.85), indicating good agreement between gradings. However, there was a shift in the gradings toward categories indicative of greater disability when using the SI. The overall difference between the distributions of Rankin grades was significant (Wilcoxon $z = -5.1; P < 0.001$).

### Intrarater Agreement

Fifty patients were recruited for the initial assessment, 48 (27 male) of whom returned for a second assessment. The median interval between first and second assessments was 7 days (range, 4 to 13 days). The mean age of patients assessed twice was 68.6 years (range, 41 to 86 years), and the mean interval poststroke was 14.9 months (range, 4 to 13 days). The mean age of patients assessed twice for the mRS-SI was 68.4 years (range, 32 to 86 years), and the mean interval after stroke was 14.2 months (range, 6 to 24 months). Twenty-five patients were interviewed by rater 1 first, and 23 by rater 2 first. On the mRS-SI, there was agreement between first and second assessment on 88% of cases for rater 1, and 98% for rater 2; $\kappa = 0.84$ (95% CI, 0.68 to 0.99) for rater 1 and $\kappa = 0.97$ (95% CI, 0.81 to 1.0) for rater 2, and $\kappa_w = 0.96$ (95% CI, 0.67 to 1.0) for rater 1 and $\kappa_w = 0.99$ (95% CI, 0.71 to 1.0) for rater 2.

### Impact of Repeat Assessment

It is conceivable that the differences in agreement on the mRS and mRS-SI observed in the study with multiple raters are caused by repetition of the assessment; that is, repetition of the assessment may improve agreement between raters. To examine whether this was likely to be the case, we looked at changes in agreement when the assessment was repeated. Calculation of inter-rater agreement on the mRS showed that exact agreement between raters was 71% on first interview and 65% on the second interview; this difference was not significant on a Wilcoxon test ($z = -1.0, P = 0.32$). Agreement between raters on the mRS-SI was 58% on first interview and 56% on the second interview; this was not significantly different (Wilcoxon test, $z = -0.38, P = 0.71$).

### Discussion

The results of this study with multiple raters showed that there was exact agreement between observers on the mRS in less than half of cases assessed, reflected in an unweighted $\kappa$ value of only 0.25. Previous studies of the reliability of the mRS have reported rather higher levels of inter-rater reliability.1,2,5,8 The current findings thus suggest that use of multiple raters from different professions for the conventional mRS is likely to increase variability. On the SI, there was significant improvement in agreement between raters. The results thus confirm that the reliability of assessment on the mRS can be improved by use of SI.

Information variance (raters having different levels/sources of information) and criterion variance (raters using different criteria) have been identified as the 2 most important sources of unreliability of questionnaires used to classify patients.12 The SI provides a systematic framework for the collection of information, suggesting questions to be asked for each level. Although it allows rephrasing and follow-up of questions and the collection of both objective and subjective information, the framework promotes better consistency of the level of information. There is also a guide for interpreting responses to inform the criteria to assign a patient to a Rankin grade.

Most disagreements on the mRS were by only one category, and the scale thus achieves a satisfactory level of inter-rater reliability assessed by the weighted $\kappa$ statistic. Quadratic weighting emphasizes the size of disagreements (differences are squared), and this is reasonable when the $\kappa$ statistic is used as a measure of the reliability of the mRS as a 6-point scale used to describe overall disability in a sample. However, in clinical trials the mRS is almost always dichotomized and consideration is confined to intergroup differences above and below a specific boundary (for example, mRS 0 to 2 versus mRS 3 to 5). In this case, weighting is irrelevant, and what matters is the extent of agreement at the particular cut-point chosen. The current findings indicate that inter-rater agreement at some boundaries is relatively poor.

#### TABLE 3. Agreement Between Ratings on the mRS and the Structured Interview

<table>
<thead>
<tr>
<th>mRS</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>19</td>
<td>16</td>
<td>3</td>
<td></td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>22</td>
<td>34</td>
<td>4</td>
<td></td>
<td></td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>39</td>
<td>15</td>
<td>1</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7</td>
<td>21</td>
<td>5</td>
<td>34</td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>36</td>
<td>51</td>
<td>83</td>
<td>41</td>
<td>12</td>
<td>226</td>
</tr>
</tbody>
</table>

By guest on July 14, 2017 http://stroke.ahajournals.org/ Downloaded from
The intrarater reliability study shows that there is excellent repeatability for both the conventional mRS and the SI. The level of agreement for the mRS is similar to that reported by Wolfe et al for repeatability of this assessment. The results thus confirm that individual raters are consistent over time in their use of the conventional mRS and the SI. The results also indicate that improvement in agreement between raters does not occur simply with repeat testing. Inter-rater agreement on second testing was very similar to agreement on second testing for both methods of assessing disability. Similarly, the SI gives consistent results when repeated, with no evidence of improvement in agreement simply through repeat assessment. Effects of repetition are thus not an explanation of the improvement in agreement between raters observed in the first study. An unexpected observation in the intrarater reliability study is that using the SI did not lead to greater agreement between raters. This part of the study was not designed to allow comparison of inter-rater reliability, and it may simply be that the patients recruited for the second part of the study involved more problematic cases (for example, cases close to borderlines between categories, cases with pre-existing disability, or cases with inconsistent reports).

The results show that whereas individual raters are consistent in their use of the mRS, inter-rater variability can nonetheless be substantial. An implication of the disagreement between raters found in the present study is that different raters focus on different aspects of disability. If so, then there is no single consensus on the application of the mRS. It is likely for example that some raters focus on physical disability and use some simple rules of thumb to assess patients, whereas other raters take more account of social limitations, residual symptoms, and level of functioning before the stroke. As a consequence, some raters showed a large shift in ratings when using the SI, whereas others showed little or no change.

This study simulates the scenario in most multicenter trials, in which outcome assessment may be performed by individuals from a variety of professional backgrounds. Whereas internal or local use of the conventional mRS by a small cohort of individuals with similar training or background may provide an adequate measure of disability, conventional use of the mRS in multicenter clinical trials is likely to reduce the power of the study as a result of increased variability. The effect of misclassifications on the power of clinical trials is discussed by Choi et al in a hypothetical example using dichotomized outcomes, a 10% misclassification rate reduced an overall difference in favor of treatment from 7.5% to 6%, whereas a 20% misclassification reduced the difference to 4.5%. In the current study, the disagreement rate for dichotomized outcomes on the conventional mRS reached 29% when ratings were split between Rankin 2 and 3. Our study indicates that any desire to continue using the conventional mRS for reasons of comparability with previous data is likely to be greatly outweighed by its poor inter-rater reliability. The conventional mRS has proved quite contentious when used alone as the primary endpoint, with discrepant results being obtained in some trials based on arbitrary choices of dichotomization cut-point (eg, PROACT 2 E CASS II). Use of the SI produced a shift in the overall distribution of ratings in comparison to the conventional mRS. It is likely that the distribution obtained with the SI results from taking a more systematic and comprehensive approach to assessment of disability and handicap after stroke. We have previously found evidence of the validity of the mRS-SI with respect to other measures of impairment, activities of daily living, and quality of life. The shift obtained in the distribution of Rankin grades emphasizes the importance of further work on the construct validity of the SI.

The potential for improvement in the power of clinical trials and consequent reduction in sample size required as a result of reduced variability is an important aspect to consider while designing trials. The findings suggest a number of practical ways in which the effects of observer variability in the assessment of outcome can be reduced in multicenter studies. The choice of where to dichotomize outcomes is often based on arguments concerning the clinical meaning of the split; however, other things being equal, agreement on the conventional mRS seems better at the Rankin 1 to 2 boundary than the Rankin 2 to 3 boundary. The number of raters involved in the study should be kept to the minimum that is practical (for example, no more than 1 rater per center). Agreement on the mRS is improved if raters use standardized questions and guidelines, and training is given. In most clinical trials, raters perform a number of assessments (including, for example, the National Institutes of Health or Scandinavian Stroke Scales, and the Barthel Index) before finally assigning a global outcome rating on the mRS. Although this will standardize at least some of the information that is collected by raters and should reduce inter-rater variability, there is also a need for consensus guidelines for resolving issues that arise when assessing outcome on the mRS. Training for raters in the interpretation of this information may improve variability of outcome assessment, but the value of unstructured training is uncertain. Finally, the use of SI after appropriate training appears to produce consistent and substantial improvements in reliability that should be considered the optimal approach at the present time.

Acknowledgments

We are grateful to participants in the study from Wishaw General Hospital, the Southern General Hospital, Glasgow, and the Victoria Infirmary, Glasgow. The study was supported by a research grant from Pfizer Global Pharmaceuticals, UK to the University of Stirling.

References


Reliability of the Modified Rankin Scale Across Multiple Raters: Benefits of a Structured Interview
J. T. Lindsay Wilson, Asha Hareendran, Anne Hendry, Jan Potter, Ian Bone and Keith W. Muir

Stroke. 2005;36:777-781; originally published online February 17, 2005;
doi: 10.1161/01.STR.0000157596.13234.95
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

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/36/4/777