Improving the Assessment of Outcomes in Stroke
Use of a Structured Interview to Assign Grades on the Modified Rankin Scale

J.T. Lindsay Wilson, PhD; Asha Hareendran, PhD; Marie Grant, BSc; Tracey Baird, MD;
Ursula G.R. Schulz, MD; Keith W. Muir, MD; Ian Bone, MD

Background and Purpose—The modified Rankin Scale is widely used to assess changes in activity and lifestyle after stroke, but it has been criticized for its subjectivity. The purpose of the present study was to compare conventional assessment on the modified Rankin Scale with assessment through a structured interview.

Methods—Sixty-three patients with stroke 6 to 24 months previously were interviewed and graded independently on the modified Rankin Scale by 2 observers. These observers then underwent training in use of a structured interview for the scale that covered 5 areas of everyday function. Eight weeks after the first assessment, the same observers reassessed 58 of these patients using the structured interview.

Results—Interrater reliability was measured with the κ statistic (weighted with quadratic weights). For the scale applied conventionally, overall agreement between the 2 raters was 57% (κ = 0.78); 1 rater assigned significantly lower grades than the other (P = 0.048). On the structured interview, the overall agreement between raters was 78% (κ = 0.93), and there was no overall difference between raters in grades assigned (P = 0.17). Rankin grades from the conventional assessment and the structured interview were highly correlated, but there was significantly less disagreement between raters when the structured interview was used (P = 0.004).

Conclusions—Variability and bias between raters in assigning patients to Rankin grades may be reduced by use of a structured interview. Use of a structured interview for the scale could potentially improve the quality of results from clinical studies in stroke. (Stroke. 2002;33:2243-2246.)

Key Words: clinical trials □ disability evaluation □ outcome □ outcome assessment

The Rankin Scale\(^1\) has wide acceptance as a measure of functional outcome after stroke and has become one of the most popular end points for clinical trials.\(^2\) The version of the scale most commonly used in trials is the modified Rankin Scale,\(^3\) which is a simple 6-point assessment that includes reference to both limitations in activity and changes in lifestyle.

The reliability of the modified Rankin Scale has been investigated and found to be satisfactory.\(^3\) However, comparison with the Barthel Index\(^4\) indicates lower levels of interrater agreement for the modified Rankin Scale\(^5\) and suggests that some raters may systematically assign higher or lower Rankin grades than others. The descriptions given for the categories of the modified Rankin Scale are broad and open to subjective interpretation. Walking is the only explicit criterion for assessment mentioned, and even for this criterion, it is not specified whether someone requiring an aid should be considered able to walk. It is therefore left open to raters to develop idiosyncratic criteria or to apply the scale in an impressionistic manner. Discrepancies are particularly striking for Rankin grades 2, 3, and 4, and it has been suggested that interviewers should use a checklist of activities of daily living (ADL) to produce greater uniformity in the application of the scale.\(^5\) Wolfe and colleagues\(^5\) have advocated using Barthel scores to generate ratings on the Rankin Scale and shown that this improves reliability. However, this approach can be applied only to the lower outcome categories of the modified Rankin Scale that relate to the basic ADL assessed by the Barthel Index.

The Glasgow Outcome Scale\(^6\) is similar in concept to the Rankin Scale, and the problem of impressionistic use of the Glasgow Outcome Scale has been addressed through the use of a structured interview.\(^7\) The purpose of the present study was to develop a structured interview for the modified Rankin Scale and to compare this form of assessment with the conventional application of the scale. We wanted to investigate whether using a structured interview could increase agreement between raters on the modified Rankin Scale. We...
TABLE 1. The Modified Rankin Scale and Corresponding Sections of the Structured Interview

<table>
<thead>
<tr>
<th>Modified Rankin Scale</th>
<th>Structured Interview for the Modified Rankin Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>5=Severe disability; bedridden, incontinent, and requiring constant nursing care and attention.</td>
<td>5=Severe disability; someone needs to be available at all times; care may be provided by either a trained or an untrained caregiver. Question: Does the person require constant care?</td>
</tr>
<tr>
<td>4=Moderately severe disability; unable to walk without assistance, and unable to attend to own bodily needs without assistance.</td>
<td>4=Moderately severe disability; need for assistance with some basic ADL, but not requiring constant care. Question: Is assistance essential for eating, using the toilet, daily hygiene, or walking?</td>
</tr>
<tr>
<td>3=Moderate disability; requiring some help, but able to walk without assistance.</td>
<td>3=Moderate disability; need for assistance with some instrumental ADL but not basic ADL. Question: Is assistance essential for preparing a simple meal, doing household chores, looking after money, shopping, or traveling locally?</td>
</tr>
<tr>
<td>2=Slight disability; unable to carry out all previous activities but able to look after own affairs without assistance.</td>
<td>2=Slight disability; limitations in participation in usual social roles, but independent for ADL. Questions: Has there been a change in the person’s ability to work or look after others if these were roles before stroke? Has there been a change in the person’s ability to participate in previous social and leisure activities? Has the person had problems with relationships or become isolated?</td>
</tr>
<tr>
<td>1=No significant disability despite symptoms; able to carry out all usual duties and activities.</td>
<td>1=No significant disability; symptoms present but not other limitations. Question: Does the person have difficulty reading or writing, difficulty speaking or finding the right word, problems with balance or coordination, visual problems, numbness (face, arms, legs, hands, feet), loss of movement (face, arms, legs, hands, feet), difficulty with swallowing, or other symptom resulting from stroke?</td>
</tr>
<tr>
<td>0=No symptoms at all.</td>
<td>0=No symptoms at all; no limitations and no symptoms.</td>
</tr>
</tbody>
</table>

Carried out 2 interrater reliability studies with the same patients 8 weeks apart, the first with the conventional scale and the second with the structured interview. To reduce the possibility of change between the first and second assessments, only patients who had suffered stroke ≥6 months previously were included in the study.

Methods

Structured Interview

The structured interview (Table 1) differs from the conventional guided interview for the modified Rankin Scale by defining specific questions to grade each category. The structured interview developed for the study consists of 5 sections: (1) constant care, (2) basic ADL, (3) instrumental ADL, (4) limitations in participation in usual social roles, and (5) checklist for the presence of common stroke symptoms. Items for inclusion in the interview were selected after review of outcome assessments used in stroke and focus groups held with stroke patients. An initial draft of the interview was piloted before the final version for the study was produced. Section 2 of the interview is based on items from the Barthel Index and followed the definitions provided by Collin et al.8 Sections 3 and 4 are adapted from the structured interview for the extended Glasgow Outcome Scale.7 Unlike a questionnaire, a structured interview allows reformulation of questions to suit the particular circumstances. In each section, restrictions in activities before stroke are recorded, and raters are instructed to discount preexisting limitations in the final rating. Raters were encouraged to interview a relative when possible and to base their assessments on ability to do the task rather than performance. Further details concerning the principles involved in administering a structured interview are given elsewhere.7 The structured interview took ≈15 minutes to administer.

Participants

The study was confined to patients surviving stroke by ≥6 months. Study inclusion criteria were as follows: age ≥18 years; diagnosis of stroke 6 to 24 months previously; living at home, living in an institution, and/or attending outpatient clinics; and ability to respond appropriately to interview in English. Excluded from the study were patients with terminal cancer, seizure disorder, dementia, substance or alcohol abuse, and major organ failure (unstable cardiopulmonary function, impaired hepatic or renal function resulting in episodic alterations in functional ADL); those unable and/or unlikely to comprehend and follow the study protocol; and patients not contacted on the advice of their general practitioners. Informed consent was obtained for each study participant.

Procedure

Both raters were neurologists in training; rater 1 (T.B.) was a specialist registrar in neurology with 4 years of experience; and rater 2 (U.G.R.S.) was a senior house officer with 2 years of experience. Before beginning the study, the raters practiced applying the modified Rankin Scale in a stroke population. Patients were assessed on 2 occasions 8 weeks apart. On the first occasion, the 2 raters interviewed each patient independently and assigned a rating on the modified Rankin Scale. The raters were instructed not to confer about ratings of individual patients. Rankin grades were assigned immediately after the initial interview. After all patients had been assessed, the raters were trained to use the structured interview to assign Rankin grades. The patients were then recalled, and each patient was independently assessed with the structured interview.

Statistical Analysis

Strength of agreement between raters is described with the κ statistic that corrects for agreement by chance. When there are >2 points on an assessment, it is appropriate to use a weighted value (κw) to take into account the size of disagreements. To facilitate comparison with previous studies, we used quadratic weights for this analysis. Quadratic weights penalize extreme disagreements particularly heavily (differences are squared), and it has been shown that when weighted this way, κw is comparable to the intraclass correlation coefficient used for continuous measures.9 Brennan and Silman10 suggest the following interpretation of the κ statistic (weighted appropriately) for the agreement between clinical measures: 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. The 95% CIs for κw values are given. Ratings were also compared through the use of appropriate nonparametric tests.

Results

Sixty-three patients were initially recruited into the study and took part in the first assessment in which the modified Rankin Scale was applied in a conventional manner; 58 patients returned (92%) for a second assessment 8 weeks after the first with the Rankin structured interview. Reasons for loss to
follow-up were death (n=1), serious illness (n=2), alcohol abuse (n=1), and did not attend (n=1).

The 58 patients (31 men) who took part in both assessments were between 37 and 90 years of age (mean, 68.3 years; SD, 10.95 years). The first assessment took place 6 to 24 months after stroke (mean, 17.1 months; SD, 5.2 months).

Ratings from the first interview, in which the modified Rankin Scale was applied in the conventional manner, are given in Table 2. The overall agreement between raters was 57%; the unweighted $\kappa$ statistic was 0.44, and $\kappa_w$ was 0.78 (95% CI, 0.53 to 1.0). In 8 cases, rater 1 rated patients less favorably than rater 2, and in 17 cases, rater 1 rated patients more favorably. Comparison of grades given by raters indicated a significant overall difference between observers (Wilcoxon $Z=-1.8$, $P=0.072$ for rater 1; $Z=-0.69$, $P=0.491$ for rater 2). The 2 assessments were also highly correlated (Spearman’s correlation, 0.82; $P<0.001$ for rater 1; Spearman’s correlation, 0.90; $P<0.001$ for rater 2).

**Table 2. Modified Rankin Scale: Cross-Tabulation of Rankin Grades Assigned by 2 Raters Who Independently Interviewed 58 Patients**

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>5</td>
<td>5</td>
<td>3</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td>7</td>
<td>10</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>11</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td>58</td>
</tr>
</tbody>
</table>

Ratings from the structured interview for the modified Rankin Scale are given in Table 3. The overall agreement between raters was 78%; $\kappa$ was 0.70, and $\kappa_w$ was 0.93 (95% CI, 0.67 to 1.0). In 4 cases, patients were rated less favorably by rater 1, and in 9 cases, they were rated more favorably. There was no significant difference in the overall rankings assigned (Wilcoxon $Z=-1.4$, $P=0.17$).

To compare the studies, we analyzed disagreements between raters. There were 25 disagreements between raters in the first study and 13 in the second study. Rankin grades for rater 2 were subtracted from Rankin grades for rater 1, and the absolute differences with and without the structured interview were compared. The analysis showed that the extent of disagreement was less when the structured interview was used (Wilcoxon $Z=-2.85$, $P=0.004$).

The overall distributions of ratings for each rater (given in the “total” columns in Tables 2 and 3) do not differ substantially between the 2 assessments. Although there were differences in the individual ratings between the 2 assessments (22 for rater 1, 19 for rater 2), only 1 difference was by $>1$ category. To test whether there was a significant shift in overall scoring for each rater, we compared the Rankin scores given without the structured interview with those obtained with the structured interview using the Wilcoxon test. For both observers, there was no overall difference in the ratings assigned on the 2 assessments (Wilcoxon $Z=-1.8$, $P=0.072$ for rater 1; $Z=-0.69$, $P=0.491$ for rater 2). The 2 assessments were also highly correlated (Spearman’s correlation, 0.82; $P<0.001$ for rater 1; Spearman’s correlation, 0.90; $P<0.001$ for rater 2).

**Table 3. Structured Interview: Cross-Tabulation of Rankin Grades Assigned by 2 Raters for 58 Patients**

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>21</td>
<td>10</td>
<td>9</td>
<td>17</td>
<td>58</td>
</tr>
</tbody>
</table>

**Discussion**

The results for the conventionally applied modified Rankin Scale indicate good interrater reliability but indicate that significant bias may be present. The structured interview for the modified Rankin Scale had very good interrater reliability, the extent of disagreement between raters was less, and significant bias was not present. Comparison of ratings on the structured interview and the conventional Rankin showed that they were highly correlated, indicating that the structured interview has satisfactory criterion validity when measured against the conventional Rankin as a standard.

The findings of the present study are consistent with previous studies of the reliability of the modified Rankin Scale. The interrater reliability of conventional assessment with the modified Rankin Scale is satisfactory but nonetheless open to improvement. Direct comparison of the present findings with previous reports is complicated because the distributions of gradings differ. The recruitment criteria for the present study tended to eliminate the most mildly disabled and the most severely disabled groups. To allow comparison with the present study, we reanalyzed the study of Van Swieten et al, confining analysis to 67 patients in Rankin categories 0 to 4. This analysis yielded an overall agreement of 61%, a $\kappa$ of 0.49 and a $\kappa_w$ of 0.80 (quadratic weights). Wolfe et al reported values of $\kappa_w$ ranging from 0.75 to 0.96 for interobserver agreement on the Modified Rankin, and Bamford et al gave a value of 0.72 for the version of the Rankin Scale used in the Oxfordshire Community Stroke Project. The reliability of the conventional Rankin Scale in the present study is thus very similar to previous reports. In agreement with the study of Wolfe et al, the present study also demonstrates that significant bias may be present even when the $\kappa$ value is satisfactory. Wolfe et al reported systematic differences in the overall rankings produced by their 3 raters.

Limitations of the present study are that only 2 raters were used, and both came from similar professional backgrounds. In large clinical trials, multiple observers contribute data, and 2 observers do not represent this situation. Using multiple observers with different backgrounds may lead to greater divergence in the application of the conventional Modified Rankin (ie, lower reliability), and there could consequently be a larger effect of introducing a standardized procedure. In the present study, it was not possible to counterbalance order of assessment with the conventional and structured interview,
because exposure to the structured interview will inevitably affect the style of approach adopted in the conventional assessment. Once exposed to the structured interview, raters may simply continue to ask the same questions when asked to assess patients on the conventional scale. It is possible that some of the reduction in variability is due to a practice effect for raters or another time difference. However, interrater reliability for the conventional assessment obtained in our study was similar to previous reports; and interrater reliability with the structured interview was similar to that found for the use of a structured interview for the Glasgow Outcome Scale in head-injured patients. Our results are therefore comparable with previous studies and suggest that major differences resulting from practice are unlikely. Further study could define the relevance of any time effects and investigate interrater reliability when the assessment is used by multiple raters from different professional backgrounds. After the study, the raters were debriefed in detail, and reasons for disagreements were identified. We have subsequently developed a set of detailed guidelines for the interview and a video for use in training raters. Training raters is particularly likely to be of importance in multicenter clinical trials, and raters should have the opportunity to observe an interview being conducted and record responses themselves. It may also be desirable to have an accreditation process to ensure that observers are applying the scale appropriately.

Van Swieten et al noted that further improvement in the Rankin Scale should be possible, and our results confirm that it is possible to reduce variability in ratings. The present findings probably have most relevance to the conduct of multicenter clinical trials. Use of the modified Rankin Scale as a functional assessment in clinical trials is supported by a recent analysis, and it is likely to remain a popular end point. Demonstrating convincing treatment effects in acute stroke trials has proved to be a challenge, and it has been suggested that use of less-than-optimal methods to measure outcome may be responsible for problems of inconsistent findings. Choi and colleagues have demonstrated that misclassification not only reduces the power of a clinical trial but also reduces the size of the observed treatment effect on dichotomous outcomes. The present investigation shows that use of a structured interview for the modified Rankin Scale may help to reduce variation between raters and improve the quality of results in clinical studies.

Acknowledgments
This project was supported by a research grant from Pfizer UK to the University of Stirling. We would like to thank Peter J. Snyder, PhD, Robert Bagdorf, MD, and Michael Krams, MD, for their comments and support of the project.

References
Improving the Assessment of Outcomes in Stroke: Use of a Structured Interview to Assign Grades on the Modified Rankin Scale

J.T. Lindsay Wilson, Asha Hareendran, Marie Grant, Tracey Baird, Ursula G.R. Schulz, Keith W. Muir and Ian Bone

Stroke. 2002;33:2243-2246
doi: 10.1161/01.STR.0000027437.22450.BD

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/33/9/2243