Interobserver Agreement for the Assessment of Handicap in Stroke Patients

J.C. van Swieten, MD, P.J. Koudstaal, MD, M.C. Visser, H.J.A. Schouten, PhD, and J. van Gijn, MD

Interobserver agreement for the assessment of handicap in stroke patients was investigated in a group of 10 senior neurologists and 24 residents from two centers. One hundred patients were separately interviewed by two physicians in different combinations. The degree of handicap was recorded by each observer on the modified Rankin scale, which has six grades (0-5). The agreement rates were corrected for chance (κ statistics). Both physicians agreed on the degree of handicap in 65 patients; they differed by one grade in 32 patients and by two grades in 3 patients. κ for all pairwise observations was 0.56; the value for weighted κ (with quadratic disagreement weights) was 0.91. Our results confirm the value of the modified Rankin scale in the assessment of handicap in stroke patients; nevertheless, further improvements are possible. (Stroke 1988;19:604-607)

A reliable measure of deficits after a stroke is important in the analysis of a therapeutic trial. The purpose of the trial determines whether the measurement concerns a specific or more global function. In theory, the spectrum ranges from exact quantification of the force of one muscle to estimation of the quality of life.1,2 Between these extremes, four different levels of measurement can be distinguished in practice. First, some clinical trials have limited the analysis to muscle strength and tendon jerks.3 A slightly more complex level is represented by tests that have been developed for the assessment of partial functions, such as the use of the hemiplegic arm.4 Tests for estimating the severity of aphasia are another example. However, such tests give no information about the function of the patient as a whole. The third level, formed by several disability indexes, of which the Barthel Index is one of the best known,5 measures the activities of daily living (ADL). In several therapeutic trials such an assessment, with emphasis on motor functions, was used.6,7 Other indexes, such as the Activity Index devised in Sweden, also take into account disorders of language and cognition.8

Finally, on the fourth level are scales that measure independence rather than performance of specific tasks and in this way incorporate mental as well as physical adaptation to the neurologic deficits. The score on such a scale gives a better impression of whether patients can look after themselves in daily life than ADL scores, and represents handicap rather than disability.9 The Glasgow Outcome Scale and the scale introduced by Rankin are examples of this type of scale.10,11 The Glasgow Outcome Scale was devised for head injury, but its general terms made it also suited for cerebrovascular disease.12 The Rankin scale has been slightly modified by Warlow and associates for the UK-TIA study13 to accommodate language disorders and cognitive defects (Table 1). It is currently used in the European Carotid Surgery Trial and the Dutch TIA trial.14 This modified Rankin scale not only measures the overall independence of stroke patients and allows comparison between patients with different kinds of neurologic deficits, but it also adds one further dimension by referring to previous activities. This is important because patients may be independent yet dissatisfied by restriction of their former lifestyle. On the other hand, patients can be restricted in their activities by complaints (arthritis, intermittent claudication) existing long before their stroke.

Clinical assessments are liable to disagreement between different observers.15 Interobserver variation has been investigated for the performance of individual ADL items16 but not for an overall handicap scale. The aim of our study was to determine the extent of interobserver agreement for the grading of stroke patients with the modified Rankin scale.

Subjects and Methods

Our aim was to imitate the circumstances of a multicenter clinical trial as closely as possible by involving many physicians with different levels of clinical experience and by performing the study in two different hospitals (the University Hospital Utrecht and the University Hospital Dijkzigt, Rotterdam).

During the study period (March 1 to September 1, 1986) we tried to include all patients in whom cerebral infarction was diagnosed by the referring physician or by a resident in either department of neurology. One hundred patients were assessed, 50 in each center; 67 were men. Patients were included only if the neurologic deficit had lasted for >24 hours. Inpatients (86) were eligible within the first week after a brain infarct, outpatients (14) within 5 months after their stroke. In one center six senior neurologists and 14 residents participated in the study, in the other four senior neurologists and 10

From the University Departments of Neurology, Utrecht and Rotterdam, and the University Department of Medical Statistics, Maastricht, The Netherlands.

Address for correspondence: J.C. van Swieten, Department of Neurology, University Hospital Nicolaas Beetsstraat 24, 3511 HG Utrecht, The Netherlands.

Received August 7, 1987; accepted November 25, 1987.
TABLE 1. The Modified Rankin Scale

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No symptoms at all</td>
</tr>
<tr>
<td>1</td>
<td>No significant disability despite symptoms: able to carry out all usual duties and activities</td>
</tr>
<tr>
<td>2</td>
<td>Slight disability: unable to carry out all previous activities but able to look after own affairs without assistance</td>
</tr>
<tr>
<td>3</td>
<td>Moderate disability: requiring some help, but able to walk without assistance</td>
</tr>
<tr>
<td>4</td>
<td>Moderately severe disability: unable to walk without assistance, and unable to attend to own bodily needs without assistance</td>
</tr>
<tr>
<td>5</td>
<td>Severe disability: bedridden, incontinent, and requiring constant nursing care and attention</td>
</tr>
</tbody>
</table>

Original Rankin scale did not contain Grade 0, defined Grade 1 as "No significant disability: able to carry out all usual duties," and defined Grade 2 as "Slight disability: unable to carry out some of previous activities ... ."

To record the degree of handicap, the modified Rankin scale was used (Table 1); the terms were explained to the physicians in a training session. The assessment was carried out by questioning the patients on activities of daily living, including outdoor activities. Information about the patient's neurologic deficits on examination, including aphasia and intellectual deficits, was given beforehand. Results of computed tomography (CT scan) were also transmitted to the physicians. The nursing staff in the hospital or a relative could be interviewed about the degree of independence of the patient. All aspects of physical and mental performance and speech were combined in the choice of a single handicap grade. The two physicians graded the patient within 6 hours of each other to avoid disagreement caused by a change in the patient's condition.

The degree of agreement between the 100 pairs of observers was calculated with 𝜅 statistics. If all degrees of disagreement are of equal importance, the coefficient of agreement is expressed as 𝜅 = (𝑝o − 𝑝c)/(100 − 𝑝c) where 𝑝o is the percent agreement observed and 𝑝c is the percent agreement expected by chance. Weighted 𝜅 (𝜅w) is used when the degree of disagreement is taken into account. In our calculations four times as much weight was given to a difference of two grades as to a difference of one grade and nine times as much weight to a difference of three grades (quadratic disagreement weights 𝑣ij = (𝑖 − 𝑗)2 between Rankin grades 𝑖 and 𝑗). Perfect agreement was assigned 0 (diagonal in Table 2). 𝜅w is calculated as

\[
\kappa_w = 1 - \frac{\sum v_i p_w}{2 \sum v_i p_{w0}}
\]

where 𝑣i is the disagreement weight, 𝑝w is the observed percentage of a certain degree of disagreement between Rankin scores 𝑖 and 𝑗, 𝑝w is the corresponding chance percentage of disagreement. 𝜅w is 0 when there is only chance agreement and 1 when there is perfect agreement.

Results

The neurologic deficits consisted of only motor deficit in 58 patients, motor deficit with hemianopia or aphasia in 33 patients, and only hemianopia or aphasia in 9 patients. A CT scan of the brain was available on the day of assessment in 90 patients. Twenty-two patients had an infarct in the left and 26 patients in the right cerebral hemisphere, and in 42 patients abnormalities were not or not yet visible.

The 100 pairs of observers agreed about the degree of handicap in 65 of 100 patients (Table 2). In 32 patients the assessments differed by one grade, and in three patients the difference was two grades. The corresponding 𝜅 is 0.56, 𝜅w is 0.91. Agreement for the different grades of the modified Rankin scale was best for Grades 0 and 5, which might be expected because disagreement is possible in only one direction, and worst for the Grades 2, 3, and 4. The neurologic deficit in patients about whom the observers did not agree was motor deficit alone in 25 of 58 patients, motor deficit with hemianopia or aphasia in 7 of 33 patients, and hemianopia or aphasia alone in 3 of 9 patients. The agreement rates involving each kind of neurologic deficit are shown in Table 3. Observers disagreed about 2 of 15 outpatients and about 33 of 85 inpatients. 𝜅 was 0.82 for outpatients and 0.51 for inpatients. This was a significant difference even though 𝜅w was hardly different, 0.91 and 0.89, respectively.

The results from the two centers were not significantly different; for the University Hospital Rotterdam 𝜅 was 0.50, 𝜅w 0.90. For the University Hospital Utrecht, these values were 0.62 and 0.91, respectively.

Discussion

If a handicap scale is used for assessing outcome in a therapeutic trial of patients with cerebrovascular disease, the results of the trial might be influenced by variation in grading between physicians. In our study the interobserver agreement for the modified Rankin scale was 0.91.

TABLE 2. Agreement Between 100 Pairs of Observers for Degree of Disability Expressed Using Modified Rankin Scale

<table>
<thead>
<tr>
<th>Observer 1</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>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>25</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td></td>
<td></td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>24</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>10</td>
<td>21</td>
<td>16</td>
<td>22</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

TABLE 3. 𝜅 According to Kind of Neurologic Deficit

<table>
<thead>
<tr>
<th>Kind of Neurologic Deficit</th>
<th>Total number</th>
<th>Number with disagreement</th>
<th>𝜅</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motor deficit only</td>
<td>58</td>
<td>25</td>
<td>0.47</td>
</tr>
<tr>
<td>Hemianopia or aphasia only</td>
<td>9</td>
<td>3</td>
<td>0.59</td>
</tr>
<tr>
<td>Motor deficit with hemianopia or aphasia</td>
<td>33</td>
<td>7</td>
<td>0.63</td>
</tr>
</tbody>
</table>
scale was satisfactory. The observers agreed on the extent of handicap in 65 of 100 patients. This corresponds with a true agreement rate, after correction for chance, of 0.56, which is substantial. The observers differed by one grade in 32 patients and by two grades in three patients. The latter cases concerned a difference in the assessment of the ability to perform previous activities in one patient and of the level of independence in the other two patients. If a difference of two grades is given four times as much weight as a difference of one grade, the weighted true agreement is excellent (0.91). These results are more convincing if the great number of observers, and particularly that of less experienced residents, is taken into account.

Because this is the first interobserver study of an overall handicap scale in stroke patients, our results cannot be compared with those of earlier studies. It is also difficult to find authoritative criteria on what represents a satisfactory $\kappa$. Some have assumed that when $\kappa$ is $>0.80$, the agreement can be considered excellent, that $\kappa$ between 0.40 and 0.80 represents moderate to substantial agreement, $\kappa$ between 0.20 and 0.40 fair agreement, and $\kappa$ of $<0.20$ slight or poor agreement. The same type of statistics is used for grading examination papers that consist of multiple choice questions, and then the limit is often set at 0.6. On the other hand, many common clinical signs and symptoms fail to attain this limit when subjected to an interobserver study. Interobserver agreement for symptoms and signs in stroke patients showed $\kappa$ between 0.40 and 0.70. For the assessment of overall outcome following severe head injury with a five-category scale $\kappa$ was 0.77. Interobserver agreement for individual ADL items (such as dressing, feeding, and walking) proved to be good, but this assessment did not take into account the degree of dependence.

To further improve the agreement rate, it is necessary to unravel possible causes of variation. First, Grades 3 and 4 are defined in a way that assumes a constant relation between the ability to walk and the ability to lead an independent life. This assumption is not always correct and may lead to ambiguities. As overall handicap is clearly the main theme of the scale, walking should perhaps not be an explicit criterion. Such a modification is presently used in the Oxfordshire Community Stroke Project. Furthermore, it may be difficult to assess restrictions of lifestyle in hospital inpatients, as was done for the most part in this study, and perhaps the handicap after stroke should not be graded until 6 months after the stroke. Second, uniformity might be improved if the observers would use a checklist of activities of daily living as a guide in questioning the patient, as was found for the case of the diagnosis of transient ischemic attacks (TIAs). Third, discrepancies between observers are most striking for Grades 2, 3, and 4. This corresponds with the low $\kappa$ for the intermediate level (Grade 3) of the Hunt-Hess scale in the grading of patients with subarachnoid hemorrhage. Reducing the modified Rankin scale to a four- or even a three-point scale would probably improve the interobserver variation. The Barthel Index has this advantage, with only two or three points for each item, although this scale is not very sensitive toward the upper end of the handicap range. On the other hand, modest but clinically relevant differences between patients can no longer be detected if the scale is contracted too much. Therefore, the modified Rankin scale is probably an acceptable compromise. Fourth, variation might theoretically arise from a difficulty in combining the impacts of different neurologic deficits such as hemiparesis, hemianopia, or aphasia on the overall handicap of the patient. Nevertheless, the results were contrary to this hypothesis because patients with only motor deficit turned out to be the most difficult to assess. These patients were probably over-represented among the middle parts of the scale, in which disagreement can go both ways. Finally, variation can perhaps also be reduced if the scale is thoroughly discussed with all participating physicians before the start of the study and if the observers practice the use of the scale, but such training is hardly realistic in the context of a multicenter trial.

In conclusion, although we found a satisfactory interobserver agreement for the grading of stroke patients with the modified Rankin handicap scale, further improvement may be possible in two ways. The first is devising a simple pro forma with questions that are most useful in detecting restrictions of the patients' lifestyle. The second is the removal of walking from the scale, leaving overall handicap as the leading theme. It is important to include all causes of handicap in patients with TIA or minor stroke because they may suffer from other complications such as angina, myocardial infarction, intermittent claudication, or retinal infarction. Even nonvascular events may be side effects of the preventive treatment that is under study and ought to be included in the assessment.

Acknowledgments

We thank all physicians involved in the assessments. We received helpful comments from Prof. C.P. Warlow (Edinburgh), Dr. D.T. Wade (Oxford), Dr. J.M. Bamford (Leeds), and Prof. H. van Crevel (Amsterdam).

References


**Key Words**: cerebrovascular disorders • disability evaluation • interobserver agreement
Interobserver agreement for the assessment of handicap in stroke patients.
J C van Swieten, P J Koudstaal, M C Visser, H J Schouten and J van Gijn

Stroke. 1988;19:604-607
doi: 10.1161/01.STR.19.5.604

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/19/5/604