Assessment of Scales of Disability and Handicap for Stroke Patients

C.D.A. Wolfe, MD; N.A. Taub, MSc; E.J. Woodrow, BA; and P.G.J. Burney, FFPHM

Background and Purpose: The purpose of the study is to compare the reliability of the Barthel activities of daily living score, which assesses disability, with the Rankin scale, which assesses handicap, and to determine their mutual agreement.

Methods: Fifty patients with stroke of varying severity were identified by a community-based stroke register and interviewed by two of three research nurses on two occasions that were 2–3 weeks apart.

Results: There was no evidence of a systematic difference between the first and second measurements. Repeatability was assessed using a κ statistic with quadratic disagreement weights (κw) to take account of extreme differences. This measure was very good for both Barthel (κw = 0.98) and Rankin (κw = 0.95) scales. There was also excellent agreement between raters for the Barthel scale (κw ≥ 0.88), but some indication of disagreement (κw = 0.75) between raters for the Rankin scale. Analysis of variance confirmed these findings. A conversion from the Barthel to the Rankin scale can be derived by assigning the most common Rankin score for the subjects with a given Barthel score, producing a κw of 0.91 for agreement.

Conclusions: The Barthel scale is a more reliable and less subjective scale for assessing disability, from which a Rankin handicap score can then be derived to enable those managing stroke patients to assess aspects of handicap as well as disability. (Stroke 1991;22:1242-1244)

It is necessary for all those involved in the management of stroke patients to assess disability and handicap using valid, reliable scales in a standardized manner. A disability is defined by the World Health Organization as any restriction or lack of ability to perform an activity in a manner or within a range considered normal for human beings and reflects the consequences of impairment in terms of functional performance and activity by the individual. A handicap is a disadvantage, resulting from an impairment or a disability, that limits or prevents the fulfillment of a role that is normal for that individual and reflects interaction with and adaptation to the individual's surroundings. Stroke results in a wide range of disabilities, making comparisons between patients difficult, but the functional consequences can be assessed using activity of daily living (ADL) scores. These are made up of separate items relating to self-care, mobility, and continence. The component scores are often combined into a global score either with no weighting or with weighting such that some individual components attach more importance than others. These scores have been reported to have a direct relation with the degree of independence of the patient. The Barthel scale is numerical and scores 10 functions on a scale 0 (fully dependent) to 20 (independent). Use of such ADL scores by all professionals allows the standardization of treatment and follow-up records, which would be of benefit to the planning and audit of patient care and research. In a comparison of three well-documented ADL scores, Gresham and colleagues found that the Barthel score possessed certain advantages, including completeness, sensitivity to change, suitability for statistical manipulation, and greater familiarity due to more widespread use.

There are also measures of independence, rather than of the ability to perform specific tasks, which incorporate a patient's deficits and give a better impression than the ADL scores of whether patients can look after themselves in daily life. Such scores, including the Rankin, have been reported as portraying handicap rather than disability. The Rankin scale was modified for use in the Oxfordshire Community Stroke Study to accommodate language disorders and cognitive defects, but its relation to ADL, its reliability, and its repeatability have not been formally assessed. The modified Rankin scale is scored 0–5: 0 (no symptoms), 1 or 2 (functionally...
We are currently engaged in a population-based stroke registry project. One of the aims of the project is to describe the change in patients' disability and handicap over time. Previous stroke registries have used either the Barthel or modified Rankin scales, but the relation between them has not been assessed. The aims of this study are, therefore, to compare the reliability of these two scores and to determine their agreement.

### Subjects and Methods

Fifty patients diagnosed clinically as having suffered a stroke at least 3 months previously were visited at home or in hospital. The World Health Organization definition of a stroke was used to identify patients for the study. Patients were registered in a community-based stroke registry and were selected to represent a range of mild to severe strokes of all ages. The three observers were registered general nurses working as research assistants for the stroke registry. Before the study they had 6 months' experience using the Barthel and Rankin scales on first-time stroke patients. Assessments were made by questioning the patient directly and also, where necessary, by questioning the nurses caring for them. None of the patients had received any rehabilitation in the interval between assessments. The Wilcoxon test for matched pairs was used to test whether there was a tendency for subjects' scores to rise or fall between the two assessments.

The degree of repeatability between the 50 pairs of observations was calculated using the χ statistic. If all degrees of disagreement are of equal importance, the coefficient of agreement is \( \chi^2 = (p_e - p_o)/((100 - p_o)) \), where \( p_e \) is the percent agreement observed and \( p_o \) is the percent agreement expected by chance. \( \chi \) is an index of the agreement over and above that expected by chance alone; for most purposes, \( \chi \) values >0.75 may be taken to represent excellent agreement. Fourteen patients were assessed twice by the same observer within a 2-week time interval, to obtain replicate measurements. This time interval was chosen to minimize the effect of disease progression on the one hand and observer recall on the other. A weighted χ statistic with quadratic disagreement weights (\( \chi^w \)) was used to take account of extreme differences and has been shown to be approximately equivalent to the intraclass correlation coefficient for measures on a continuous scale.

The differences between raters' mean scores were estimated by analysis of variance using Generalized Linear Interactive Modeling (GLIM). Initially the difference between each subject's first and second scores was plotted against their mean, to check that these were approximately normally distributed and were independent of the severity of the subject's condition. The χ statistic for all subjects and for each pair of raters is shown in Table 1. The percentage exact agreement between raters is also displayed in Table 1.

### Results

There was no evidence of a systematic difference between the first and second measurement of either score (Wilcoxon test, \( p > 0.1 \)). The Barthel score ranged from 0-4 (16%), 5-9 (21%), 10-14 (27%), 15-19 (23%), and 20 (13%). The Rankin score took the values 0 (3%), 1 (3%), 2 (16%), 3 (21%), 4 (29%), and 5 (28%). The χ statistic for all subjects and for each pair of raters is shown in Table 1. The percentage exact agreement between raters is also displayed in Table 1.

The high values of \( \chi^w \) for the subjects measured twice by the same rater, \( \chi^w = 0.98 \) for Barthel and \( \chi^w = 0.95 \) for Rankin, indicate excellent repeatability on each scale. There is excellent agreement between raters on the Barthel scale (\( \chi^w = 0.88 \) for each pair of raters). There is some indication of disagreement between raters on the Rankin scale, with raters 1 and 3 agreeing closely (\( \chi^w = 0.96 \)) but differing from rater 2.

Analysis of variance using GLIM demonstrated that there was no evidence of any systematic difference between the raters' Barthel scores (\( F_{2,48} = 0.25, p > 0.5 \)) but there was strong evidence of a systematic difference between the raters' Rankin scores (\( F_{2,48} = 6.02, p = 0.005 \)). The analysis of variance indicated that raters 1 and 3 estimated the Rankin score 0.42 and 0.33 points (estimated with standard error of 0.13) higher than rater 2.

An empirical comparison of Barthel scores with Rankin scores can be derived by assigning the most common Rankin score for these subjects with a given Barthel score. There was no Barthel score that corresponded with a Rankin score of 0 or 1. Even the subjects with a Barthel score of 20 were most likely to have a Rankin score of 2 (8 of 13). Therefore Rankin scores 0, 1, and 2 were grouped together. The

### Table 1. Weighted χ (\( \chi^w \)) for Repeatability and Rater Agreement

<table>
<thead>
<tr>
<th>Raters 1 and 2</th>
<th>n</th>
<th>( \chi^w )</th>
<th>95% CI</th>
<th>% exact agreement</th>
<th>Rankin scale</th>
<th>Raters 1 and 3</th>
<th>Raters 2 and 3</th>
<th>Replicates</th>
<th>All subjects</th>
<th>n</th>
<th>( \chi^w )</th>
<th>95% CI</th>
<th>% exact agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters 1 and 2</td>
<td>11</td>
<td>0.88</td>
<td>0.75-1.00</td>
<td>27</td>
<td>0.75</td>
<td>0.47-1.00</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raters 1 and 3</td>
<td>15</td>
<td>0.94</td>
<td>0.90-0.98</td>
<td>27</td>
<td>0.96</td>
<td>0.90-1.00</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raters 2 and 3</td>
<td>10</td>
<td>0.98</td>
<td>0.96-1.00</td>
<td>50</td>
<td>0.85</td>
<td>0.71-0.98</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates</td>
<td>14</td>
<td>0.98</td>
<td>0.95-1.00</td>
<td>50</td>
<td>0.95</td>
<td>0.88-1.00</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All subjects</td>
<td>50</td>
<td>0.96</td>
<td>0.94-0.98</td>
<td>38</td>
<td>0.90</td>
<td>0.84-0.97</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI, confidence interval.
Table 2. Conversion of Barthel to Rankin Scales on 100 Observations

<table>
<thead>
<tr>
<th>Barthel scale score</th>
<th>Rankin scale score</th>
<th>K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>19, 20</td>
<td>0, 1, 2</td>
<td>0.72</td>
<td>17</td>
</tr>
<tr>
<td>14–18</td>
<td>3</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>8–13</td>
<td>4</td>
<td>21</td>
<td>29</td>
</tr>
<tr>
<td>0–7</td>
<td>5</td>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

conversion is illustrated in Table 2. The unweighted ω was 0.72 (95% confidence interval 0.61–0.82) and ω = 0.91 (0.87–0.95).

Discussion

The King’s Fund consensus document advocates a multidisciplinary approach to stroke care that would require reliable, standardized measures of disability and handicap to be used by all professionals; this study has assessed the use by research nurses of two commonly used scales. There was no evidence that the scores changed systematically between the two measurements, making a 2–3 week interval between patient assessments legitimate in this study. The results also indicated a high degree of repeatability, indicating that the measurements are not subject to random fluctuation with time.

The ω statistics for agreement for interobserver variation were excellent and in agreement with previous studies. If such scales are to be used by members of a team, the Barthel interobserver variation is excellent, but there was some disagreement between rater 2 and the others for the Rankin score such that rater 2 scores approximately 0.4 less than raters 1 and 3 on the Rankin scale. Although this is not as good as the Barthel agreement, this magnitude of difference is unlikely to be clinically relevant. The percentage exact agreement for the Barthel scale was lower than for the Rankin scale (38 versus 80) because the former is a 20-point scale and the latter only a 6-point scale. As van Swieten has discussed, the Rankin score is subjectively assessed, instead of being scored on stated criteria, and misclassification is therefore possible. Rater 2 had a tendency to classify patients lower because of incorporation of asymptomatic comorbidities into the score. A suggestion of reducing the Rankin scale to 3 or 4 points to improve interobserver variation would be supported by our observations. The use of an ADL checklist for scoring the Rankin scale would overcome, in part, the subjective nature of the score.

Although both scales have been used in studies, the nature of the relation between them has not been addressed. Previous comparisons of ADL scores have shown them to have unweighted agreements of ω = 0.77 and ω = 0.42. To allow comparison, groups 0, 1, and 2 on the Rankin scale were combined, making it a 4-point scale; the unweighted ω was 0.72 and ω was 0.91, which indicate very substantial agreement. The close association between disability and handicap in this group of patients allows the conversion from one score to the other. In view of the better interobserver and intraobserver agreement for the Barthel scale, it is suggested that a Barthel score be assessed and a Rankin score derived to assess handicap. This would be in agreement with research groups in the United Kingdom who have tended to use the Barthel score as the most complete, sensitive, and widely used scale.

Acknowledgments

We thank Mrs. Richardson for her fieldwork, Ms. Law for statistical advice, and Ana Childs for manuscript preparation.

References


KEY WORDS • stroke assessment • stroke outcome
Assessment of scales of disability and handicap for stroke patients.
C D Wolfe, N A Taub, E J Woodrow and P G Burney

Stroke. 1991;22:1242-1244
doI: 10.1161/01.STR.22.10.1242

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/22/10/1242