Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials
Prospective Reliability and Validity

Brett C. Meyer, MD; Thomas M. Hemmen, MD; Christy M. Jackson, MD; Patrick D. Lyden, MD

Background and Purpose—To prospectively evaluate the reliability and validity of this previously developed stroke scale in an independently collected cohort. The National Institutes of Health Stroke Scale (NIHSS) has been criticized for its complexity and variability. Prior formal clinimetric analyses were used to obtain a modified version of NIHSS (mNIHSS), which retrospectively demonstrated improved reliability and validity. We sought to prospectively measure the reliability and validity of the mNIHSS.

Methods—Forty-five patients with a history of stroke or intracerebral hemorrhage were evaluated at the University of California, San Diego, Stroke Center from September 2000 through March 2001. Each patient was tested by 2 NIHSS-certified neurologists using the NIHSS, mNIHSS, Barthel Index, and Modified Rankin scales.

Results—There were a large percentage of high $\kappa$ values using the mNIHSS. Only 10 (66.67%) of 15 NIHSS $\kappa$ scores showed excellent agreement, whereas 10 (90.91%) of 11 mNIHSS $\kappa$ scores showed excellent agreement. As predicted, the mNIHSS was more reliable than the NIHSS because of the exclusion of items with low $\kappa$ values. With the use of correlation coefficient analysis, the mNIHSS was as valid as the NIHSS.

Conclusions—This prospective study found high reliability and continued validity by using a previously developed mNIHSS. Items found to have low $\kappa$ values were consistent with the previously derived retrospective mNIHSS. The resulting mNIHSS scale has much higher $\kappa$ values. The mNIHSS showed improved agreement between examiners and was also easier to administer, having fewer and simpler items. Further prospective evaluation should assess whether the mNIHSS could be used in lieu of the NIHSS. (Stroke. 2002;33:1261-1266.)

Key Words: clinical trials ■ outcome ■ stroke ■ stroke assessment

The National Institutes of Health Stroke Scale (NIHSS) is a graded neurological examination that assesses speech, language cognition, inattention, visual field abnormalities, motor and sensory impairments, and ataxia. The scale was developed for use in acute-stroke therapy trials and has since been widely used as a standard part of the assessment in clinical trials. This scale, along with many others, has been evaluated in its clinical usefulness in the assessment of the stroke patient.

The ideal stroke scale should be valid, reliable, and easy to administer in multiple settings by a broad range of health care practitioners. The NIHSS satisfies many, but not all, of the criteria for an ideal stroke scale. The NIHSS is not time-consuming to administer, taking <8 minutes to perform. Overall interrater reliability has been shown in multicenter stroke trials. NIHSS reliability has been extended to nonneurologist physicians and nonphysician study coordinators in clinical trials, as well as in community-based studies. This reliability improves with personal and videotape training. Factor analysis demonstrated content validity of the NIHSS. Regarding outcomes, the NIHSS has very good sensitivity, specificity, and accuracy in predicting clinical results at 3 months. However, this scale does contain specific items with poor reliability and redundancy and has been criticized for its complexity and resulting variability.

An ideal stroke scale, one that is reliable, valid, and predictive of patient outcome, would be frequently used in the evaluation of stroke patients. Although the NIHSS is useful in clinical research trials, it may not be routinely used in other situations. When emergency department acute-stroke study data were evaluated, a CT scan was performed in almost all cases, and a neurology consult was obtained in the majority of cases, but only 1.2% of patients with acute stroke had evidence of an NIHSS being performed. Although the NIHSS is well suited for acute and general use, it still falls short of the ideal stroke scale requirements.

Formal clinimetric analyses were used to improve the NIHSS, resulting in a modified version of the NIHSS (mNIHSS). With fewer items and simpler grading scales, the mNIHSS was intended to be easier to administer. The reliability has been demonstrated with certification data used previously for reliability testing of the NIHSS. Items with poor $\kappa$ values were reduced from 20% to 14% with use of the

Received November 15, 2001; final revision received February 2, 2002; accepted February 19, 2002.
From the Department of Neurosciences, UCSD School of Medicine, San Diego, Calif.
Correspondence to Brett C. Meyer, MD, Department of Neurosciences, UCSD School of Medicine, Stroke Center (8466), 3rd Floor, OPC, Suite #3, 200 West Arbor Dr, San Diego, CA 92103-8466. E-mail bcmeyer@ucsd.edu
© 2002 American Heart Association, Inc.

Stroke is available at http://www.strokeaha.org

DOI: 10.1161/01.STR.0000015625.87603.A7
mNIHSS. Validity was tested with National Institute of Neurological Disorders and Stroke outcome result data, which has also been used previously to investigate the NIHSS. The mNIHSS demonstrated improved reliability and validity by use of factor and coefficient analysis. Power was greater with the mNIHSS, resulting in the potential for smaller sample sizes in clinical trials. The resulting simplified mNIHSS was felt to be a scale that is reliable, valid, and easy to administer in the clinical research setting. However, the mNIHSS has not been prospectively evaluated to confirm these findings. The present study prospectively evaluated the reliability and validity of the mNIHSS.

Subjects and Methods

The Modified NIHSS

The Figure displays the mNIHSS scoring sheet. Four questions were removed from the NIHSS to create the final mNIHSS. The first item, level of consciousness (LOC), was dropped because factor analysis demonstrated it to be redundant. The remaining "consciousness" items had higher $\kappa$ values and were retained. The ataxia, facial weakness, and dysarthria items exhibited poor reliability, with facial weakness and dysarthria items also being redundant. Therefore, these items were deleted. The sensory question was simplified to 2 choices because of poor reliability of the third choice item. The maximum possible score with the use of this simplified scale is 31, compared with 42 for the original scale. All examiners in the present study were certified in use of the NIHSS with the use of previously published methods.

Patients and Procedures

Patients seen at the University of California, San Diego, Stroke Center from September 2000 through March 2001 with a diagnosis of cerebrovascular accident or intracerebral hemorrhage were included in the present study. All patients exhibited deficits lasting $\geq 24$ hours. Hospitalized and clinic patients were included in the present study to test the reliability of stroke scales in a wide range of clinical settings. Patients with nonvascular structural lesions or seizures accounting for symptoms were excluded. A series of 45 patients was evaluated by a pool of 4 certified NIHSS examiners;
each patient was tested simultaneously by 2 of the 4 total examiners with the use of the NIHSS and mNIHSS. Scoring was performed in a blinded fashion, with neither examiner knowing the other examiner’s scoring results. The Barthel Index (BI)\(^16\) and Modified Rankin scale (MRS)\(^17\) were scored only for outpatients (n=27) to allow for comparison with assessments of activities of daily living and functional impairment, because it was not possible to determine accurate functional scales on patients with new neurological deficits in the acute hospital setting. The BI and MRS were obtained at the same time as the NIHSS and mNIHSS. Patient demographic data and risk factor information were also obtained.

**Reliability**

Nonparametric methods were used, because the evaluated stroke scale scores are ordinal level data.\(^5\)-\(^18\) Interobserver reliability was rated with weighted \(\kappa\) statistics for each item of the NIHSS and mNIHSS. The \(\kappa\) statistics measure agreement among observers over and above that expected by chance alone.\(^19\) In the present study, weighted \(\kappa\) values were used because this method assigns weights to disagreements based on the magnitude of disagreement and also accounts for agreements based on chance alone.\(^18\) This method is generally used for ordinal scales, such as the ones being evaluated here, in which changes in numerical values are not consistent across all questions. The weighted \(\kappa\) value is qualified as follows: \(\kappa<0.40\) defines poor agreement, \(\kappa\) between 0.40 and 0.75 defines moderate agreement, and \(\kappa>0.75\) defines excellent agreement.\(^18\) Reliability was tested by using data from all 4 potential observers. Specifically, interrater reliability was assessed by comparing stroke scale scores between 2 raters at a time. At times, however, each of the 4 raters were compared with each other for interrater reliability assessments. Each examiner was blinded to the other rater’s scoring.

**Validity**

Validity was examined with the Spearman rank coefficients to compare the stroke scales. In settings such as those in the present study, in which an independent gold-standard method of evaluation is not available, criterion validity cannot be assessed. Predictive/ outcome validity was evaluated by comparing the NIHSS and new mNIHSS with functional outcome and disability scales. Construct validity was assessed by comparing the accepted NIHSS with the newly proposed stroke scale (mNIHSS) directly.\(^20\) Validity measurements were performed on only those patients evaluated as outpatients (n=27). In these patients, it was possible to obtain accurate functional outcome scale measurements (BI and MRS) to assess validity.

**Results**

Table 1 presents the baseline characteristics of the 45 patients. In the present study, 23 (51%) patients were white, 7 (16%) patients were African American, and 8 (18%) patients identified themselves as Hispanic. There were 30 (67%) male patients with the mean age of 65 years (range 42 to 88 years). Ischemic stroke represented 42 (93%) and intracerebral hemorrhage represented 3 (7%) of the total patients. There were no patients with transient ischemic attack (TIA) evaluated. There were 18 (40%) acute inpatient evaluations and 27 (60%) outpatient clinic evaluations. Time since symptom onset ranged from 1 to 13 days (mean 4.72 days) for the inpatients and 0 to 12 410 days (mean 531.6 days) for the outpatients. One patient was seen 34 years after the onset of symptoms. Excluding this patient from analysis, the range was 0 to 478 days (mean 74.77 days).

Other demographic data obtained are summarized in Table 1. Forty-four percent of the patients had a history of prior stroke or TIA. Seventy-three percent of the patients had hypertension, 42% had coronary artery disease, 47% had diabetes, 49% had hyperlipidemia, and 18% had atrial fibrillation. A family history of stroke or TIA was present in 22% of the patients. Tobacco use was acknowledged in 18%, and current alcohol use was found in 22%. Forty-seven percent of the patients had only mild neurological deficits at the time of examination.

The NIHSS scores obtained ranged from 0 to 24 (median 5). The mNIHSS scores ranged from 0 to 20 (median 3). Total NIHSS scores did not differ between examiners by >4 points, whereas total mNIHSS scores did not differ between examiners by >2 points.

**Reliability**

The NIHSS and mNIHSS were prospectively evaluated for reliability on the basis of weighted \(\kappa\) scores. The specific \(\kappa\) values for each item on the NIHSS and mNIHSS are shown (Table 2). Regarding the original NIHSS, 10 (66.67%) items (LOC questions, LOC commands, visual fields, left arm motor, right arm motor, left leg motor, right leg motor, sensory, aphasia, and neglect) displayed excellent agreement beyond chance alone. Four (26.67%) NIHSS items (LOC, gaze palsy, facial palsy, and ataxia) displayed good agreement beyond chance alone. One (6.67%) NIHSS item (dys-
TABLE 2. $\kappa$ Values, Item Reliability Table: Item Name, Number of Comparisons per Item, Weighted $\kappa$ and 95% CIs for NIHSS and mNIHSS (between examiners)

<table>
<thead>
<tr>
<th>Number</th>
<th>Item Name</th>
<th>Comparisons per Item, n</th>
<th>Weighted $\kappa$ Value</th>
<th>95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIHSS</td>
<td>LOC</td>
<td>45</td>
<td>0.457*</td>
<td>0.164–0.749</td>
</tr>
<tr>
<td></td>
<td>LOC questions</td>
<td>45</td>
<td>0.937*</td>
<td>0.645–1.229</td>
</tr>
<tr>
<td></td>
<td>LOC commands</td>
<td>45</td>
<td>0.943*</td>
<td>0.652–1.235</td>
</tr>
<tr>
<td></td>
<td>Gaze</td>
<td>45</td>
<td>0.662*</td>
<td>0.369–0.954</td>
</tr>
<tr>
<td></td>
<td>Visual fields</td>
<td>45</td>
<td>0.876*</td>
<td>0.589–1.163</td>
</tr>
<tr>
<td></td>
<td>Facial palsy</td>
<td>45</td>
<td>0.742*</td>
<td>0.460–1.024</td>
</tr>
<tr>
<td></td>
<td>Left arm motor</td>
<td>45</td>
<td>0.971*</td>
<td>0.678–1.263</td>
</tr>
<tr>
<td></td>
<td>Right arm motor</td>
<td>45</td>
<td>0.959*</td>
<td>0.669–1.249</td>
</tr>
<tr>
<td></td>
<td>Left leg motor</td>
<td>45</td>
<td>0.947*</td>
<td>0.654–1.239</td>
</tr>
<tr>
<td></td>
<td>Right leg motor</td>
<td>45</td>
<td>0.975*</td>
<td>0.684–1.267</td>
</tr>
<tr>
<td></td>
<td>Limb ataxia</td>
<td>45</td>
<td>0.690*</td>
<td>0.407–0.973</td>
</tr>
<tr>
<td></td>
<td>Sensory</td>
<td>45</td>
<td>0.892*</td>
<td>0.601–1.183</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>45</td>
<td>0.841*</td>
<td>0.555–1.127</td>
</tr>
<tr>
<td></td>
<td>Dysarthria</td>
<td>45</td>
<td>0.289</td>
<td>–0.005–0.583</td>
</tr>
<tr>
<td></td>
<td>Neglect</td>
<td>45</td>
<td>0.891*</td>
<td>0.599–1.183</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>45</td>
<td>0.969*</td>
<td>0.678–1.261</td>
</tr>
<tr>
<td>mNIHSS</td>
<td>LOC questions</td>
<td>45</td>
<td>0.937*</td>
<td>0.645–1.229</td>
</tr>
<tr>
<td></td>
<td>LOC commands</td>
<td>45</td>
<td>0.943*</td>
<td>0.652–1.235</td>
</tr>
<tr>
<td></td>
<td>Gaze</td>
<td>45</td>
<td>0.661*</td>
<td>0.369–0.954</td>
</tr>
<tr>
<td></td>
<td>Visual fields</td>
<td>45</td>
<td>0.876*</td>
<td>0.589–1.163</td>
</tr>
<tr>
<td></td>
<td>Left arm motor</td>
<td>45</td>
<td>0.971*</td>
<td>0.678–1.263</td>
</tr>
<tr>
<td></td>
<td>Right arm motor</td>
<td>45</td>
<td>0.959*</td>
<td>0.669–1.249</td>
</tr>
<tr>
<td></td>
<td>Left leg motor</td>
<td>45</td>
<td>0.947*</td>
<td>0.654–1.239</td>
</tr>
<tr>
<td></td>
<td>Right leg motor</td>
<td>45</td>
<td>0.975*</td>
<td>0.684–1.267</td>
</tr>
<tr>
<td></td>
<td>Limb ataxia</td>
<td>45</td>
<td>0.690*</td>
<td>0.407–0.973</td>
</tr>
<tr>
<td></td>
<td>Sensory</td>
<td>45</td>
<td>0.892*</td>
<td>0.601–1.183</td>
</tr>
<tr>
<td></td>
<td>Language</td>
<td>45</td>
<td>0.841*</td>
<td>0.555–1.127</td>
</tr>
<tr>
<td></td>
<td>Neglect</td>
<td>45</td>
<td>0.891*</td>
<td>0.599–1.183</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>45</td>
<td>0.988*</td>
<td>0.696–1.280</td>
</tr>
</tbody>
</table>

*P < 0.05.

Validity

The NIHSS has been shown to be a valid clinical deficit scale.1 In retrospective analyses,14 the mNIHSS showed high correlation with other scales, similar to the NIHSS, and was a valid predictor of outcome. In prospective evaluation, we also found construct validity, in that the mNIHSS performed similarly to the NIHSS. The Spearman correlation coefficient between NIHSS and mNIHSS (for both examiners) was high (0.947, 0.941), with an overall value of 0.944. As a measure of concurrent validity, the NIHSS and mNIHSS were compared with functional outcome measures (BI and MRS). The coefficients for examiner 1 for NIHSS versus BI and MRS were −0.166 and −0.169, respectively. The coefficients for mNIHSS versus BI and MRS were −0.230 and 0.281, respectively. The absolute Spearman values were improved with the use of the mNIHSS, although values were not statistically significant. Examiner 2 and combined examiners revealed similar trends. Spearman coefficients comparing NIHSS and mNIHSS with BI and MRS are shown in Table 4.

Discussion

The mNIHSS is a revised stroke scale primarily developed for clinical research and is simpler and easier to administer than the previous NIHSS. The main prerequisites to the ideal stroke-scoring system are reliability, consistency, and validi-
Many stroke scales have been developed in the past, but few have been thoroughly tested for interobserver reliability and validity. A few well-known scales have shown high interobserver reliability (International Classification of Diseases score of 10), high reliability across items (NIHSS, the Canadian Neurological Scale, and the European Stroke Scale), and highly reliable measures of disability (BI). The present study prospectively evaluated a retrospectively constructed mNIHSS; we found much improved reliability and consistent validity when it was compared with the original scale. Prior studies have retrospectively proved the reliability, validity, and responsiveness of the mNIHSS. The items removed from the scale were LOC, facial weakness, ataxia, and dysarthria. The sensory item was collapsed to 2 responses. Items were removed because of poor reliability or redundancy in prior clinimetric studies. The resulting scale continued to have validity, and improved reliability was noted. However, the mNIHSS had not previously been prospectively studied with the use of an independently collected cohort of patients.

In the present study, the NIHSS and mNIHSS were tested prospectively. As in the retrospective evaluation, the same unreliable items (LOC, facial palsy, limb ataxia, and dysarthria) were again confirmed (Table 2). These were the same items that were removed from the final version of the mNIHSS. In the present study, the gaze-palsy item also had only good reliability.

**Reliability**

Reliability may be tested with $\kappa$ statistics. In a previous retrospective study, reliability data were collected on the 11 remaining items from the original NIHSS scoring sheet (15 total items). Theoretically, this may have allowed for inadvertent biasing of the scores. To exclude this possibility, the 11-item scale was tested prospectively. Previous studies have used a difference of $\pm 4$ points on the NIHSS to reflect a clinically significant change beyond interrater variability. Total NIHSS scores did not differ between examiners by $>4$ points, whereas total mNIHSS scores did not differ by $>2$ points. The mNIHSS score was more consistent and reproducible.

The reliability of the mNIHSS compares favorably with that of the NIHSS reported by Brott et al. with 10 (90.91%) of 11 mNIHSS items having $\kappa$ values in the excellent range compared with only 10 (66.67%) of 15 items on the NIHSS. As predicted, the mNIHSS is more reliable by excluding NIHSS items with low $\kappa$ values and is, therefore, an improved scale.

**Validity**

Construct validity is measured by using correlation coefficients to test a newly developed scale against a previously used scale. In the present study, the mNIHSS was found to be a valid predictor of the original NIHSS. The correlation coefficients between the total NIHSS score and the total mNIHSS score were excellent (Table 4). When prospectively compared with the original scale, the nearly identical correlation coefficients confirm that the mNIHSS behaves similarly to the original.

As a measure of concurrent validity, the mNIHSS was compared with the BI and MRS. The absolute Spearman values were higher for the mNIHSS than for the NIHSS; however, the trend was not statistically significant. In the present study, the majority of the patients had only mild clinical deficits. This clustering of patients with mild deficits has made it difficult to draw further conclusions regarding concurrent validity of this new stroke scale, especially at the higher end of the scale.

The mNIHSS, which is easier to administer because of fewer and less complicated items, was prospectively shown to be a more reliable and accurate stroke deficit scale. The mNIHSS can potentially be used in lieu of the original NIHSS in similar settings. Accurate assessment of clinical deficit, inclusion and exclusion criteria for clinical trial enrollment, and potential guidelines for safe thrombolytic use are all areas for potential use.

Other indications, capitalizing on strengths of the mNIHSS, can be further evaluated. NIHSS data abstraction has previously been found to be a reliable and valid method for the estimation of the NIHSS from medical records. This means of data collection has been adopted, in several studies, to evaluate increased numbers of patients when an initial NIHSS was not performed at the time of admission. If the mNIHSS can also be abstracted from medical records with a high degree of reliability and validity, a record of initial patient presentation that is more accurate and easier to obtain may be found. This may make clinical trial data analysis more efficient and accurate, allowing for increased patient numbers included in study trials. A prospective analysis of medical record abstraction using the mNIHSS may be indicated.

Previous studies have evaluated the role of telemedicine in the evaluation of stroke. One study tested whether NIHSS agreement would persist if performed over a telemedicine link. A good interrater correlation coefficient was seen (0.97, $P<0.001$). Although 4 items had excellent weighted $\kappa$ correlations, certain items (LOC, ataxia, and commands) had poor reliability. If the improved mNIHSS were implemented, this application of telemedicine could be potentially expanded, allowing for broader-reaching evaluation of stroke patients in areas previously not staffed by NIHSS-certified examiners. A prospective evaluation using the mNIHSS could be performed.

The potential benefits of the present study must be viewed in light of its limitations. First, the small study size must be taken into consideration. Although the results are statistically significant, further evaluation in a setting with larger patient numbers and multiple examiners should be considered. The validity assessments were performed only on outpatients (n = 27) to accurately assess functional outcome scales, thus limiting the conclusions that can be drawn. Further validity assessments are planned with a larger number of patients. Second, the relatively low stroke scale values obtained in the present study may not be representative of the stroke population as a whole. Most patients had mild to moderate strokes (median NIHSS 5, maximum 24, and minimum 0). This is a 7- to 9-point lower median than is found in other studies. A broader range of deficits must be evaluated to fully test reliability. This clustering of mild deficits may have negatively affected the accurate assessment of scale validity. Further validity assessments are planned with a larger number of patients. Third, the present study did not correlate mNIHSS with imaging findings of stroke size, nor did it assess the stroke scale as a predictor of
outcome after a stroke. As previously shown, baseline NIHSS strongly predicts outcomes after stroke, with a score of ≥16 being associated with a high probability of death or severe disability and a score of ≥6 being associated with a good recovery.\textsuperscript{13} The mNIHSS was retrospectively shown to be a predictor of stroke outcome. This should be prospectively addressed in the future. The correlation of outcome with stroke scale score can be a measure of validity. Because the NIHSS, mNIHSS, BI, and MRS were all obtained at the same time, a true measure of predictive validity cannot be assessed by the present study. A prospective follow-up of these patients, or another study, could help to strengthen the concurrent or predictive validity of this scale.

Previous findings revealed that the NIHSS favors left/dominant hemisphere strokes, with right hemisphere events being consistently larger than left hemisphere events.\textsuperscript{31} In the NIHSS, 7 of 42 points are related to language function, whereas only 2 of 42 points are attributed to neglect functions. By dropping the dysarthria question from the mNIHSS, the balance may be shifted more toward minimizing the lateralization bias. Subsequently, the mNIHSS may be a more accurate representation of the true clinical deficit. This should be further evaluated with studies on the NIHSS and mNIHSS with regard to potential hemispheric bias.

The clinical utility of the mNIHSS has yet to be evaluated outside of the clinical research setting or with nonneurologist physicians or nurse coordinators. A simplified mNIHSS may allow for improved and increased use to provide the physician with a reliable and accurate assessment of patient deficit at 1 specific time point, with possible far-reaching implications for patient outcome. The mNIHSS has initially been tested by a specialized group of neurologists previously certified in use of the NIHSS. This may limit the applicability of the results to other groups, and further evaluation in a broader setting with nonneurologist physicians or nurse coordinators should be performed.

The mNIHSS is not the ideal stroke-scoring scale. However, it is an improvement over many of the scales used in the past. Overall, the improved reliability and the preserved validity of the mNIHSS make it a very attractive clinical deficit scale for the evaluation of stroke patients in the clinical research setting and beyond.

Acknowledgments

This study was supported in part by the Veterans’ Affairs Medical Research Division and a Grant-in-Aid from the American Heart Association.

References

Modified National Institutes of Health Stroke Scale for Use in Stroke Clinical Trials: Prospective Reliability and Validity
Brett C. Meyer, Thomas M. Hemmen, Christy M. Jackson and Patrick D. Lyden

Stroke. 2002;33:1261-1266
doi: 10.1161/01.STR.000015625.87603.A7
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 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/33/5/1261