Shortening the NIH Stroke Scale for Use in the Prehospital Setting

David L. Tirschwell, MD; W.T. Longstreth, Jr, MD; Kyra J. Becker, MD; Richard E. Gammans, Sr, PhD; LuAnn A. Sabounjian, RN, BSN; Scott Hamilton, PhD; Lewis B. Morgenstern, MD

Background and Purpose—Prehospital stroke scales should identify stroke patients and measure stroke severity. The goal of this study was to identify a subset of the 15 items in the National Institutes of Health Stroke Scale (NIHSS-15) that measures stroke severity and predicts outcomes.

Methods—Using 2 distinct data sets from acute stroke clinical trials, we derived and validated shortened versions of the NIHSS (sNIHSS). Stepwise logistic regression and bootstrap techniques were used in selection of NIHSS-15 items. Areas under the receiver operator characteristic curve (C statistics) were used to compare predictive performance of logistic models incorporating differing versions of the NIHSS.

Results—The derivation analyses suggested the 8 NIHSS-15 items that were most predictive of “good outcome” 3 months after stroke, in order of decreasing importance: right leg item, left leg, gaze, visual fields, language, level of consciousness, facial palsy, and dysarthria. The sNIHSS-8 comprises all 8 and the sNIHSS-5, the first 5. In the validation models, C statistics were NIHSS-15=0.80, sNIHSS-8=0.77, and sNIHSS-5=0.76. Statistical comparisons suggested that the NIHSS-15 had better predictive performance than the sNIHSS-8 or the sNIHSS-5; the absolute difference in C statistics was small. There was no significant difference between the sNIHSS-8 and the sNIHSS-5.

Conclusions—Much of the predictive performance of the full NIHSS-15 was retained with a shortened scale, the sNIHSS-5. Shortening the NIHSS-15 will facilitate its use during prehospital evaluations. The sNIHSS severity information may be useful to triage acute stroke patients in communities and to provide a baseline stroke severity for prehospital acute stroke trials. (Stroke. 2002;33:2801-2806.)

Key Words: cerebrovascular accident ■ emergency medical services

A cuted stroke therapy is time dependent. Intravenous recombinant tissue plasminogen activator (rtPA) is almost 4 times more effective at 1 hour than at 3 hours. In a community, prompt triage to medical centers prepared to expeditiously treat acute stroke patients may reduce long-term disability. Similarly, to conserve resources, those patients not eligible for therapy and those who will improve without aggressive management may be well served at the nearest local hospital. A prehospital stroke scale leading to early recognition of stroke and an immediate estimation of prognosis is therefore a public health priority. Prehospital stroke scales may also be critical to determine enrollment criteria and balance of treatment groups in prehospital acute stroke treatment trials. Neuroprotective substances with minimal risk may be ideal candidates for such trials and may allow an ultra-early treatment time.

An optimal prehospital stroke scale must be brief and easily administered in the field. The full version of the National Institutes of Health Stroke Scale (NIHSS-15) measures stroke severity and predicts outcome from stroke. It comprises 15 neurological examination test items, making it cumbersome for use in the prehospital setting. The goal of this study was to identify a minimal subset of the 15 items in the full version of the NIHSS that continues to measure stroke severity and predicts outcomes after stroke. A shortened scale based on a subset of items from the NIHSS could comprise the prognostic portion of a prehospital stroke scale. The shortened versions of the NIHSS (sNIHSS) developed in this study were based on data from the placebo groups of acute stroke clinical trials and validated on data from another acute stroke clinical trial.

Subjects and Methods
The first phase of the study involved derivation of the sNIHSSs of the NIHSS-15 and the second phase, validation of these sNIHSSs.
Derivation Phase

Patients
The derivation sample was selected from 3 trials of citicoline for acute ischemic stroke. Only patients from the placebo-treated arms of the trials were available for analysis. The 3 trials had similar entry criteria, enrolling adult ischemic stroke patients with a middle cerebral artery syndrome within 24 hours of symptom onset. In addition, the NIHSS-15 was required to be $\geq 5$, with at least 2 points related to extremity motor weakness. A total of 223 patients from the 3 trials had complete information on presentation NIHSS-15, comorbid risk factors, and 3-month outcomes (data courtesy of Interneuron Pharmaceuticals, Lexington, Mass).

Outcome Measures
In these analyses, a “global outcome” similar to that used in the National Institute of Neurological Disorders and Stroke (NINDS) tPA trial was used as the outcome of interest. This global outcome comprised 3 separate measures of good outcome at the 12-week follow-up, including dichotomized NIHSS-15 ($\leq 1$), Rankin score ($\leq 1$), and Barthel Index ($\geq 95$). Glasgow Outcome Score (GOS) data were not collected. The global good outcome is dichotomous and can be interpreted as a multiple scale, simultaneous assessment of whether a patient has little or no residual neurological deficit, disability, or handicap.

Statistical Analysis
Forward and backward stepwise logistic regression, in which each NIHSS-15 item was entered as a separate, continuous variable, screened the NIHSS-15 for items independently predictive of global good outcome. All 3 good outcome variables for each patient were entered into the logistic models, with statistical correction for within-person correlation of outcomes. Items were retained for further analysis when $P \leq 0.20$ (a generous threshold) to ensure that predictive items were not excluded early in the analysis.

Bootstrap resampling techniques use random sampling with replacement from an original study cohort. Analyses performed on large numbers of bootstrap samples from a data set are thought to better represent the results as might be derived from the underlying population. Bootstrap techniques allow approximation of the relative importance of the items retained by the stepwise logistic models. Five hundred bootstrap iterations were performed with the items retained by the forward and backward stepwise logistic regression models; relative importance was approximated by the percentage of logistic models when $P \leq 0.10$ for the individual item.

Validation Phase

Patients
The validation sample was selected from the ATLANTIS trial of rtPA for acute ischemic stroke. Both tPA-treated and placebo-treated patients were available for analysis. The entry criteria included adult ischemic stroke patients with a measurable neurological deficit between 3 and 5 hours after symptom onset. A total of 280 patients were randomized; 131 (95%) of these patients had complete information on presentation NIHSS-15, comorbid risk factors, and 3-month outcomes (data courtesy of Genentech Inc).

Outcome Measures
A global outcome measure was again used. In this patient sample, global outcome comprised 4 separate measures of good outcome at a 90-day follow-up, including dichotomized NIHSS-15 ($\leq 1$), Rankin score ($\leq 1$), GOS ($= 1$), and Barthel Index ($\geq 95$).

Statistical Analysis
sNIHSS scores were computed for each patient on the basis of the NIHSS-15 items retained from the derivation phase analyses. Spearman’s $\rho$ was computed between the NIHSS-15 and each sNIHSS version as a measure of concurrent validity. Logistic regression was used to model good outcome based on the NIHSS-15, the sNIHSS versions, and base models. Base models used only age, gender, and comorbidity to predict outcome. A comorbidity score was created by using forward stepwise logistic regression. Age and gender were forced into the model, and then comorbid risk factors were entered into a model when $P < 0.2$, with global good outcome as the dependent variable. The natural logarithms of the odds ratios for each comorbid condition present in each patient were summed to form the comorbidity score. The score had a normal distribution and was strongly associated with global good outcome (data not shown).

The area under the receiver operator characteristic curve (also called the C statistic) is a summary measure of discriminative performance of a logistic regression model. The C statistic can range from 0.5 (prediction no better than chance) to 1.0 (perfect prediction). C statistics were used to compare predictive performance of logistic models that incorporated differing versions of the NIHSS and base models. Differing models within the derivation or the validation phase and similar models between the 2 phases were compared by their C statistics. Correction for the within-person correlation of outcomes was not available when C statistics were statistically compared, so probability values for comparisons are reported separately for each of the dichotomous outcomes (Table 4).

Patient participation in all of these randomized trials required written, informed consent and local institutional review board approval of the protocol. All analyses were performed with STATA statistical software (version 7, Stata Corp).

Results

Derivation Phase
Patient characteristics and percentages of good outcome for each of the good outcome measures are shown in Table 1. The results of the forward and backward stepwise logistic regressions led to retention of 8 of the NIHSS-15 items, the summed score of which comprises the sNIHSS-8 (Table 2). Five hundred logistic regressions on bootstrap samples were then performed with the 8 items retained by the stepwise logistic regressions. In Table 3, the relative importance of these sNIHSS-8 items, approximated by the percentage of models where $P \leq 0.10$, is shown. The items were thus ranked in order of decreasing importance: right leg item, left leg, best gaze, visual fields, best language, level of consciousness, facial palsy, and dysarthria. Retaining only those items where $P < 0.10$ in $\geq 20\%$ of the logistic models led to the removal of 3 additional items: level of consciousness, facial palsy, and dysarthria. Summing the 5 remaining items formed the sNIHSS-5 (Table 2). The percentages in Table 3 suggested that the greatest amount of prognostic information was related to motor leg scores; thus, an sNIHSS-1 score was computed as the score from the more severely affected leg.
(Table 2). Nonparametric correlations of the sNIHSS versions with the NIHSS-15 are shown in Table 2.

The estimated C statistics for logistic models that predicted global good outcome, including the base models and each version of the NIHSS, are shown in Table 4. The derivation phase models with any of the NIHSS items (models 3 through 6) had significantly greater C statistics than the base models. However, of the 9 comparisons between derivation models with NIHSS items (models 3 through 6), only 1 reached statistical significance (model 4 versus model 3, with \( \text{NIHSS}^\text{H11349} \) as the dichotomous outcome); the sNIHSS versions retained almost all of the predictive performance of the NIHSS-15.

**Validation Phase**

Patient characteristics and percentages of good outcome for each good outcome measure are shown in Table 1. Patients from the validation phase were more commonly male, younger, had less severe strokes, and better outcomes. The sNIHSS scores were computed for each patient (Table 2). The C statistics for each model were lower in the validation phase than the derivation phase and decreased with removal of items from the NIHSS-15 (Table 4). Again, the (validation) models with any of the NIHSS items (models 3 through 6) had significantly greater C statistics than the base models. In addition, C statistics were significantly different between model 4 (sNIHSS-5, C=0.76) and model 3 (sNIHSS-1, C=0.72) in 3 of 4 comparisons and between model 6 (NIHSS-15, C=0.80) and model 5 (sNIHSS-8, C=0.77) in all 4 comparisons; none of 4 comparisons were significantly different between model 5 (sNIHSS-8, C=0.77) and model 4 (sNIHSS-5, C=0.76).

When the C statistics from the derivation and validation phases were compared, the differences were most significant for models 3 through 5, which incorporated the versions of the sNIHSS. Also, the C statistics for model 6, which incorporated the (full) NIHSS-15, were significantly different in 2 of 3 comparisons.

### Table 2. NIHSS-15 Items, Their Inclusion in Shortened Scales, Mean Scores, and Nonparametric Correlations With NIHSS-15

<table>
<thead>
<tr>
<th>NIHSS-15 Item*</th>
<th>sNIHSS-8</th>
<th>sNIHSS-5</th>
<th>sNIHSS-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a. Level of consciousness</td>
<td>(0–3)</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1b. Orientation</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c. Commands</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Gaze</td>
<td>(0–2)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3. Visual fields</td>
<td>(0–3)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4. Facial paresis</td>
<td>(0–3)</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5a. Motor–arm–right</td>
<td>(0–4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5b. Motor–arm–left</td>
<td>(0–4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6a. Motor–leg–right</td>
<td>(0–4)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6b. Motor–leg–left</td>
<td>(0–4)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7. Limb ataxia</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Sensory</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Language</td>
<td>(0–3)</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10. Dysarthria</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Extinction</td>
<td>(0–2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Possible scores</th>
<th>Derivation phase (mean, SD)</th>
<th>Validation phase (mean, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–42</td>
<td>7.4 (3.9)</td>
<td>6.2 (3.4)</td>
</tr>
<tr>
<td>0–24</td>
<td>4.7 (3.0)</td>
<td>3.7 (2.6)</td>
</tr>
<tr>
<td>0–16</td>
<td>2.2 (1.2)</td>
<td>1.7 (1.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correlation with NIHSS-15 score (Spearman’s ( \rho ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivation phase</td>
</tr>
<tr>
<td>Validation phase</td>
</tr>
</tbody>
</table>

*Numbers in parentheses next to NIHSS-15 items indicate the possible score.
TABLE 4. C Statistics for Models Predicting 3-Month “Global Good Outcome” (Statistical Comparisons Made for Each “Good Outcome” Separately)

<table>
<thead>
<tr>
<th>Model</th>
<th>Derivation C Statistic</th>
<th>Validation C Statistic</th>
<th>Derivation vs Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Age, gender</td>
<td>0.65</td>
<td>0.60</td>
<td>†‡§</td>
</tr>
<tr>
<td>(2) = (1)+ comorbidity score</td>
<td>0.72‡‡‡</td>
<td>0.65‡‡‡</td>
<td>†‡‡‡</td>
</tr>
<tr>
<td>(3) = (2)+ sNIHSS-1</td>
<td>0.84††††</td>
<td>0.72††††</td>
<td>††††††</td>
</tr>
<tr>
<td>(4) = (2)+ sNIHSS-5</td>
<td>0.86†††</td>
<td>0.76†††</td>
<td>†††††</td>
</tr>
<tr>
<td>(5) = (2)+ sNIHSS-8</td>
<td>0.87†††</td>
<td>0.77†††</td>
<td>†††††</td>
</tr>
<tr>
<td>(6) = (2)+ sNIHSS-15</td>
<td>0.86†††</td>
<td>0.80††††</td>
<td>†††††</td>
</tr>
</tbody>
</table>

*Footnote symbols in Derivation or Validation columns refer to a vertical statistical comparison of a C statistic with the C statistic immediately above. In the Derivation vs Validation column, footnotes refer to a horizontal comparison of the C statistics in the Derivation vs the Validation phase of the study. All comparisons are done separately for each of the dichotomized “good outcomes” as described below.

- §NIHSS≤1 outcome; †P<0.05, ††P<0.01, †††P<0.005, ††††P<0.0005.
- ‡Rankin score=1 outcome; †P<0.05, ††P<0.01, †††P<0.005, ††††P<0.0005.
- §Barthel Index=95 outcome; §P<0.05, §§P<0.01, §§§P<0.005, §§§§P<0.0005.
- ||Glasgow Outcome Score=1 outcome; ||P<0.05, |||P<0.01, ||||P<0.005, |||||P<0.0005.

Conclusions

In this study, we derived and validated a shortened stroke severity scale, based on the NIHSS-15, which may be easier to apply in the prehospital setting than the full NIHSS-15. The analyses performed in this study suggest that much of the predictive performance of the full NIHSS-15 can be retained with the sNIHSS-5 scale.

The NIHSS-15 has been shown to be an excellent predictor of stroke outcome,17–20 but not all of its items are equally important. One review addressed the reliability, or reproducibility, of individual stroke scale items from 9 studies of 7 different stroke severity scales. The most reliable items were language function and motor power in the arm and leg. The next most reliable items were facial weakness and level of consciousness. Least reliable items included sensory function, speech, cerebellar signs, visuospatial dysfunction, and visual fields.21 Studies of the interrater reliability of the individual items of NIHSS-15 have found that arm and leg strength evaluations had the greatest interrater reliability and that facial movement, limb ataxia, neglect, level of consciousness, and dysarthria had the least reliability.19,22–24 In 1 report, removing items with poor reliability from the NIHSS-15 (level of consciousness, facial weakness, dysarthria, and ataxia) allowed retention of the predictive performance of the scale. This modified scale performed better than the NIHSS-15 and has recently been validated in an independent data set.25,26

Poorly reliable scale items are unlikely to be independent predictors of outcome because of random noise resulting from variable interrater scoring. The consistency between the aforementioned observations from the literature and the items removed in our analyses supports the position that removal of items from the NIHSS-15 can be done while retaining predictive performance. In the derivation models, little difference was evident between models with 15, 8, or 5 NIHSS items in their predictive performance as measured by C statistics. This type of effect is expected, because stepwise selection procedures during model derivation tend to overfit the model to the derivation data, thus leading to optimistic C statistics.27 In the validation models with 15, 8, and then 5 items, the observed decreases in the C statistics demonstrate why such models need to be validated in distinct patient populations. Another part of the difference in the predictive performance of the derivation versus validation models may have to do with differences in the patient populations. The citicoline patients were older and had more severe strokes; also, they were required to have had a middle cerebral artery syndrome with at least 2 NIHSS points related to the motor items. These criteria may have led to an emphasis on the motor elements in the sNIHSS-8 and sNIHSS-5 in the derivation phase. The ATLANTIS patients did not have specific artery or NIHSS element requirements and had less severe strokes. Also, in these analyses, leg strength was by far the most important predictor of the global good outcome. This observation emphasizes the importance of motor function in outcome assessment but also points out that good outcomes, as defined here, are strongly correlated with mobility.

The Los Angeles (LAPSS) and Cincinnati (CPSS) Prehospital Stroke Screens were designed to identify stroke patients in prehospital settings and assist with rapid triage. These scales were not designed with prediction of outcomes in mind and do not measure stroke severity. Both screens have been validated and rely heavily on the finding of motor asymmetry to identify stroke patients.28–31 In comparison with these validated screening tools, the sNIHSS validated in this study rates stroke severity only and has not been tested in its ability to identify stroke patients. However, there is overlap between sNIHSS items and those in the LAPSS and CPSS. To develop an NIHSS-based prehospital stroke evaluation tool best able to identify stroke patients, it may be necessary to add back some of the NIHSS items that are not predictive of outcome (eg, facial and arm weakness). Recently, a severity measure based on the LAPSS (LAPSS motor scale, or LAMS) has been proposed and preliminarily shown to have predictive performance similar to the NIHSS-15 and the sNIHSS.32 The
LAMS is a pure motor scale that uses face, arm, and grip strength and thus may have suboptimal content validity,\(^3\) ignoring possible stroke symptoms involving language, visual fields, and gaze (all captured in the sNIHSS-5).

A prehospital stroke evaluation tool will be useful in prehospital triage. For example, if patients with stroke onset within the last 1 to 2 hours could be identified in the field, they might be transported to the nearest hospital equipped to give intravenous tPA (which might not be the nearest hospital). Additional triage decisions might be made on the basis of a prehospital measure of stroke severity. If the deficits appeared minor and a good outcome was likely without aggressive intervention, diversion to a hospital with more resources for stroke care might not be necessary. If deficits were more severe and the patient or family wanted aggressive care, a hospital with a higher level of stroke care resources might be the appropriate destination regardless of time elapsed since symptom onset.

Reliably predicting good or poor outcomes and thus, the indication (or not) for diversion to a hospital with greater stroke care resources, is difficult. In a study of consecutive patients admitted to a stroke unit that compared the predictive performance of 4 stroke impairment scales, the NIHSS-15 performed best, with a cutoff of 13 or more predicting poor outcome (alive in a care facility or dead) with a sensitivity of 71%, a specificity of 90%, and a positive predictive value of 82%.\(^3\) Recently, in a placebo arm analysis of the NINDS tPA data, it was suggested that an initial NIHSS-15 $\geq 17$ with either atrial fibrillation or an altered level of consciousness was a predictor of poor outcome (Rankin $\geq 3$), with high specificity and positive predictive value but low sensitivity.\(^3\)

Reliably predicting a good outcome may be more important, because it suggests low utility of risky interventions or transfer to a hospital with greater stroke care resources. In a randomized trial of acute ischemic stroke patients, $\approx 30\%$ of patients with an NIHSS-15 score $\geq 15$ and $\approx 90\%$ of patients with an initial NIHSS-15 of 4 to 6 had a good or excellent outcome (patients who had a GOS score of 1 or 2 and a Barthel Index score of 12 to 20) at 3 months.\(^4\) All of these results suggest that even the full 15-item NIHSS cannot reliably predict outcome for the individual patient. Also, the described NIHSS-15 cutoffs have not been validated in distinct patient data sets; thus, these cutoffs are likely optimistic in their estimates of predictive power.

Limitations of this study include the fact that the sNIHSSs may be specific for predicting these specifically dichotomized outcomes. If a different outcome scale were used, such as a quality-of-life scale, a different subset of NIHSS-15 items might be predictive. Recent research suggests the importance of quality-of-life outcomes for stroke.\(^3\) Nonetheless, most acute stroke clinical trials use the Barthel Index and the NIHSS-15 as major outcome measures.\(^3\) That a shortened version of the NIHSS-15 can be reliably performed by prehospital providers remains unproven. Lending optimism are that prehospital severity assessments have been shown to reliably predict outcome after stroke,\(^3\) that identification of stroke by prehospital providers remains unproven. Lending optimism is the importance of quality-of-life outcomes for stroke.\(^3\) Nonethe-

The sNIHSS-5 provides a brief scale that can serve as a valid predictor of outcome after stroke. This abbreviated scale was developed and validated on independent data sources. Shortening the scale, as was done here, will greatly facilitate its use during an emergency prehospital evaluation. Integrating these findings with those of other prehospital stroke studies should allow the creation of a broadly applicable prehospital stroke evaluation tool, brief in duration, reliable in identification, and valid in prediction of outcomes. The severity information derived from the shortened scale may be useful to triage acute stroke patients in communities and to provide a baseline stroke severity assessment for prehospital acute stroke trials.

### Acknowledgments

Dr Tirschwell’s efforts were supported in part by NINDS grant 1 K23 NS02119-01 and by the Medic One Research Foundation, Seattle, Washington. The authors also thank Interneuron Pharmaceuticals and Genentech for allowing us access to the data from their sponsored clinical trials in the hope that such access for secondary analyses will become standard practice.

### References


Shortening the NIH Stroke Scale for Use in the Prehospital Setting

Stroke. 2002;33:2801-2806; originally published online November 14, 2002;
doi: 10.1161/01.STR.0000044166.28481.BC

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/33/12/2801