Reliability and Validity of Estimating the NIH Stroke Scale Score from Medical Records

Scott E. Kasner, MD; Julio A. Chalela, MD; Jean M. Luciano, RN; Brett L. Cucchiara, MD; Eric C. Raps, MD; Michael L. McGarvey, MD; Molly B. Conroy, MD; A. Russell Localio, MS

Background and Purpose—The aim of our study was to determine whether the National Institutes of Health Stroke Scale (NIHSS) can be estimated retrospectively from medical records. The NIHSS is a quantitative measure of stroke-related neurological deficit with established reliability and validity for use in prospective clinical research. Recently, retrospective observational studies have estimated NIHSS scores from medical records for quantitative outcome analysis. The reliability and validity of estimation based on chart review has not been determined.

Methods—Thirty-nine patients were selected because their NIHSS scores were formally measured at admission and discharge. Handwritten notes from medical records were abstracted and NIHSS scores were estimated by 6 raters who were blinded to the actual scores. Estimated scores were compared among raters and with the actual measured scores.

Results—Interrater reliability was excellent, with an intraclass correlation coefficient of 0.82. Scores were well calibrated among the 6 raters. Estimated NIHSS scores closely approximated the actual scores, with a probability of 0.86 of correctly ranking a set of patients according to 5-point interval categories (as determined by the area under the receiver-operator characteristic curve). Patients with excellent outcomes (NIHSS score of ≤5) could be identified with sensitivity of 0.72 and specificity of 0.89. There were no significant differences between these parameters at admission and discharge.

Conclusions—For the purposes of retrospective studies of acute stroke outcome, the NIHSS can be abstracted from medical records with a high degree of reliability and validity. (Stroke. 1999;30:1534-1537.)

Key Words: neuropsychological tests ■ outcome assessment ■ reproducibility of results

Clinical outcomes after stroke are measured quantitatively in clinical trials but are often recorded only qualitatively in clinical practice. Retrospective stroke studies that rely on chart review are therefore limited to clinical narratives of the neurological examination that are difficult to utilize in formal data analyses and compare to published data.

The National Institutes of Health Stroke Scale (NIHSS) is a quantitative measure of stroke-related neurological deficit that spans key aspects of the neurological examination: level of consciousness, language function, neglect, visual fields, eye movements, facial symmetry, motor strength, sensation, and coordination.1,2 The examination can be performed quickly, and the NIHSS score can be assessed by neurologists and nonneurologists after appropriate training.3 The scale has proven intrarater and interrater reliability and has predictive validity for stroke outcome.2,4 Consequently, the NIHSS is used in nearly every current acute stroke study in the United States as a measure of the initial and final neurological deficit.

Because of the widespread support for the NIHSS, recent retrospective studies have estimated the NIHSS from medical records to quantify baseline deficits and outcomes.5,6 Others have avoided this approach7 because of concerns that the reliability and validity of score estimation has not been proved, and the interpretation of such retrospective studies may be limited by the methodological constraints of information bias. We evaluated the ability to extract the NIHSS score from medical records.

Subjects and Methods

This study was performed at an academic university hospital after review and approval by our institutional review board. Patients were selected for this study if they had suffered an acute ischemic stroke leading to enrollment in an experimental stroke protocol. To be included, patients must have had a formal measurement of the NIHSS score (as part of the experimental stroke protocol) at the time of admission and discharge. At our center, this information was collected prospectively during 3 separate clinical trials, providing 39 patients for analysis. The medical monitors for these clinical trials permitted the use of their case report forms to determine the actual NIHSS scores for this study. For each eligible patient, the handwritten notes from the day of admission and the day of discharge were photocopied, edited to remove any reference to the actual measured NIHSS score and participation in the clinical trial, and then photo-
copied again for distribution to the raters. The raters estimated each component of the NIHSS for each patient on admission and discharge and calculated the total score. For each patient evaluation, raters also characterized the quality of data in the record as complete, incomplete, poor, absent, or illegible. If data were absent or illegible for a patient evaluation, a situation which occurred only in discharge notes, the NIHSS score was assumed to be unchanged from admission, as the last observation carried forward. The admission and discharge notes were paired so that the raters could make inferences for a given patient about the discharge score based in part on the admission score, even if documentation was incomplete in the discharge note. This approach to handling poor or missing information was chosen because it closely approximates the actual process of retrospective chart abstraction.

Six raters of varying levels of experience—a stroke specialist attending physician, stroke fellow, senior neurology resident, junior neurology resident, nurse coordinator, and fourth-year medical student—were selected to review the records. All were previously trained and certified in the administration of the NIHSS. All were blinded to the actual NIHSS score and to the ratings of the others.

Statistical analysis was performed using STATA version 5.0 (Stata Corporation). The scores were compared among raters, and interrater reliability was determined by calculation of an intraclass correlation coefficient (ICC) with use of ANOVA. The ICC reflects the proportion of the total variance that is due to the “true” variance among patients and is calculated as $s_e^2/(s_r^2+s_e^2)$, where $s_r^2$ is the variance component for subjects, $s_e^2$ is the variance component for the raters, and $s_{r,e}^2$ is the variance component for residual error.\(^5\) The ICC can be interpreted as a weighted $k$ statistic: ICC$=\kappa$ suggests perfect reliability, and ICC$<0.8$ is generally considered to represent excellent reliability.\(^6\) Pairwise comparisons between raters were also assessed with the ICC. Criterion validity was determined by comparison of the estimated NIHSS scores with the actual NIHSS score as recorded on the case report forms of the clinical trials in which these patients had been enrolled. To calculate sensitivity and specificity, the NIHSS was categorized into 5-point intervals (0–5, 6–10, 11–15, 16–20, 21–25, and 26–30). This 5-point categorization has been used in prior attempts to determine the NIHSS in a retrospective manner and represents a clinically relevant score threshold.\(^5,6\) Means and 95% CIs for sensitivity and specificity were calculated using a clustered bootstrap percentile method (1000 repeated samples with the patient serving as the cluster).

<table>
<thead>
<tr>
<th>NIHSS Score</th>
<th>No. Below Threshold*</th>
<th>Sensitivity (95% CI)†</th>
<th>No. Above Threshold*</th>
<th>Specificity (95% CI)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤5</td>
<td>20</td>
<td>0.72 (0.58–0.84)</td>
<td>56</td>
<td>0.89 (0.82–0.95)</td>
</tr>
<tr>
<td>≥10</td>
<td>43</td>
<td>0.90 (0.81–0.97)</td>
<td>33</td>
<td>0.71 (0.56–0.82)</td>
</tr>
<tr>
<td>≥15</td>
<td>62</td>
<td>0.93 (0.87–0.98)</td>
<td>14</td>
<td>0.46 (0.29–0.62)</td>
</tr>
<tr>
<td>≥20‡</td>
<td>74</td>
<td>0.99 (0.97–1.00)</td>
<td>2</td>
<td>0.25 (0.17–0.33)</td>
</tr>
</tbody>
</table>

*No. of patient assessments above and below the threshold scores, based on the actual measured NIHSS. There were 39 patient records evaluated at admission, of which 37 were evaluated at discharge and 2 patients were deceased, for a total of 76 patient records.

†95% CIs were determined by a clustered bootstrap percentile method (1000 repeated samples with the patient serving as the cluster).

‡For the opposite threshold, NIHSS >20, these parameters are reversed such that sensitivity is 0.25 and specificity 0.99.

Results

Admission notes were available on all 39 patients, of whom 37 were evaluated at discharge and 2 were deceased, for a total of 76 patient records. Notes were written by at least 1 neurologist (attending or resident physician) for all patients on admission and 34 of the 37 patients on discharge. Complete or near-complete neurological examinations were documented in 38 of 39 admission notes but in only 15 of 37 of discharge notes. In 3 of 37 discharge notes, the examination was entirely absent or illegible.

The actual NIHSS scores ranged from 0 to 23, with a mean±SD score of 9.7±5.4 and a median of 8. Estimated NIHSS scores are summarized by rater and by admission/discharge evaluation in Figure 1. Overall, there were no clinically or statistically significant differences among estimated mean scores according to the 6 raters (by ANOVA: overall $P=0.15$, admission $P=0.28$, discharge $P=0.59$). Median NIHSS scores and distributions were also similar for all raters, suggesting excellent calibration. Overall interrater reliability was excellent, as determined by an ICC of 0.82 (variance components $s_r^2=37.6$, $s_e^2=0.9$, and $s_{r,e}^2=7.1$). There was little difference between reliability at admission (ICC=0.83; variance components $s_r^2=19.0$, $s_e^2=0.64$, and $s_{r,e}^2=3.3$) and at discharge (ICC=0.81; variance components $s_r^2=21.8$, $s_e^2=0.40$, and $s_{r,e}^2=4.7$). Agreement between pairs of raters was also very good to excellent, with ICCs ranging from 0.70 to 0.89. For all pairs of raters, over 90% of the estimated NIHSS scores were within 5 points at both admission and discharge.

Estimated NIHSS scores were good approximations of the actual NIHSS. The differences between actual and estimated scores are summarized by rater and admission/discharge evaluation in Figure 2. Overall, 88% of the estimated scores deviated by ≤5 points from the actual scores at both admission and discharge. After categorization of the scores into 5-point intervals, the sensitivity and specificity were calculated for each threshold. These results are summarized in the Table. ROC curves were generated, and the area under the ROC curve indicates the proportion of all pairs of ratings per patient for which the patient with the higher estimated NIHSS also has the higher actual NIHSS ranking.\(^8\) Overall, the area under the ROC curve was 0.86, and it was 0.88 at admission and 0.84 at discharge. Likewise, there was little difference between the sensitivities and specificities when admission and discharge scores were compared. Patients with little or no neurological deficit (NIHSS ≤5) could be distin-
guished with a sensitivity of 0.72 (95% CI 0.58 to 0.84) and specificity of 0.89 (95% CI 0.82 to 0.95). Conversely, patients with severe deficits (NIHSS >20) could be identified with a sensitivity of 0.25 (95% CI 0.17 to 0.33) and specificity of 0.99 (95% CI 0.97 to 1.00).

Discussion

Our results suggest that the NIHSS can be estimated from the review of medical records with a high degree of reliability and validity. Prior studies have demonstrated the value of the NIHSS in prospective stroke research. Observational retrospective cohort and case-control studies cannot substitute for well-designed prospective trials but are often important for hypothesis generation and for situations in which randomized clinical trials and prospective cohorts are not feasible. The findings from the present study will be most useful for those retrospective studies in which information about stroke-related neurological deficits must be abstracted qualitatively and then transformed into a quantitative format for analysis. The vast majority of patient records can be estimated within 5 points of the actual NIHSS score, but scores for patients with relatively mild strokes or residual deficits can be assessed more accurately than for those with severe strokes.

There were no major differences between the assessment of the NIHSS score at admission and discharge, despite the relatively lower quality of the information available at discharge. It appears that the raters were able to estimate the neurological deficit at the time of discharge based at least partially on their prior knowledge of the baseline deficit and other nonspecific qualitative information in the handwritten notes, such as “no change,” “ready for discharge home,” “only needs outpatient speech therapy,” or “expect prolonged rehabilitation course.”

Estimated NIHSS scores were comparable for all 6 raters. Although all received identical training in the use of the NIHSS as a direct patient assessment tool, their level of clinical neurological experience varied. The similarity of the scores among raters attests to the effectiveness of the training videotapes, to the wide applicability and interrater reliability of the NIHSS, or to both. Further, estimated scores closely approximated the actual measured NIHSS scores.

These data were obtained from a single academic university setting in which the vast majority of patients were evaluated by neurologists. These results may not be generalizable to other settings such as community hospitals, where neurologists may not be available for patient examination either at admission or
discharge and documentation of the neurological deficit may be lacking. Therefore, the utility of chart-based estimation in a nonacademic setting remains to be determined.

These data support the use of estimated NIHSS scores in retrospective studies. However, the NIHSS is quick and easy to obtain prospectively, whereas chart abstraction is difficult, time-consuming, and only an approximate method. Standardized measurement and documentation of the NIHSS for all stroke patients could be routinely added to daily hospital notes with minimal additional effort and provide a wealth of data for future stroke research.

Acknowledgments
The authors wish to thank Ms Colleen E. Walsh for data coordination and project support.

References


Reliability and Validity of Estimating the NIH Stroke Scale Score from Medical Records

Stroke. 1999;30:1534-1537
doi: 10.1161/01.STR.30.8.1534

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/30/8/1534