NIHSS Training and Certification Using a New Digital Video Disk Is Reliable

Patrick Lyden, MD; Rema Raman, PhD; Lin Liu, PhD; James Grotta, MD; Joseph Broderick, MD; Scott Olson, MD; Sandi Shaw, RN; Judith Spilker, RN; Brett Meyer, MD; Marian Emr; Margo Warren; John Marler, MD

Background and Purpose—NIH Stroke Scale certification is required for participation in modern stroke clinical trials and as part of good clinical care in stroke centers. The existing training and certification videotapes, however, are more than 10 years old and do not contain an adequate balance of patient findings.

Methods—After producing a new NIHSS training and demonstration DVD, we selected 18 patients representing all possible scores on 15 scale items for a new certification DVD. Patients were divided into 3 certification groups of 6 patients each, balanced for lesion side, distribution of scale item findings, and total score. We sought to measure interrater reliability of the certification DVD using methodology previously published for the original videotapes. Raters were recruited from 3 experienced stroke centers. Each rater watched the new training DVD and then evaluated one of the 3 certification groups.

Results—Responses were received from 112 raters: 26.2% of all responses came from stroke nurses, 34.1% from emergency departments/other physicians, and 39.6% from neurologists. One half (50%) of raters were previously NIHSS-certified. Item responses were tabulated, scoring performed as previously published, and agreement measured with unweighted $\kappa$ coefficients for individual items and an intraclass correlation coefficient for the overall score. $\kappa$ ranged from 0.21 ± 0.05 (ataxia) to 0.92 ± 0.09 (LOC-C questions). Of 15 items, 2 showed poor, 11 moderate, and 2 excellent agreement based on $\kappa$ scores. The intraclass correlation coefficient for total score was 0.94 (95% confidence interval, 0.84 to 1.00). Reliability scores were similar among specialists and centers, and there were no differences between nurses and physicians. $\kappa$ scores trended higher among raters previously certified.

Conclusions—These certification DVDs are reliable for NIHSS certification, and scoring sheets have been posted on a web site for real-time, online certification. (Stroke. 2005;36:2446-2449.)

Key Words: clinometrics ■ reliability ■ scales ■ stroke

The National Institutes of Health Stroke Scale (NIHSS) is a widely used stroke deficit assessment tool. Nearly all large clinical stroke trials—whether diagnostic or therapeutic—require a baseline and outcome severity assessment using the NIHSS. Nonstroke specialists who care for stroke patients are certifying in increasing numbers, especially now that Disease-Specific Specialty Designation as a Primary Stroke Center is available from the Joint Commission on Accreditation of Health Care Organization. A training and certification process exists to assure that raters use the NIHSS in a uniform manner1: videotapes were produced for training and certification in 1988 and have been distributed widely by the American Academy of Neurology, the American Heart Association, and the National Stroke Association. These tapes contain certain flaws that continue to affect the effective training of new raters. First, the video technology was state-of-the-art at the time, but some subtle findings are poorly visualized—reliability of some items is poor.2–7 Second, the patients were selected as a convenience sample from a single medical center; several key scale items could not be illustrated in either the training or the certification videos. Third, DVDs are replacing videotape as the medium of choice. Rather than simply transfer flawed video from the tapes to DVD, we sought to create new video that would address the previous flaws.

Methods

Experts were convened by the National Institute of Neurological Disorders and Stroke (NINDS) to review the previous videotapes, craft a new shooting script, and design the new DVD. A professional video production team was selected and a set built in the UCSD.
television studio. Videotaping occurred between February 17 and 27, 2003. Patients were selected to assure that every choice in every scale item was illustrated at least once. Patients were drawn from the UCSD Outpatient Stroke Clinics, the Veteran’s Affairs Medical Center, San Diego Stroke Clinic, the UCSD Medical Center, the San Diego Rehabilitation Institute, and the Grossmont Medical Center Rehabilitation Institute. To assure proper illustration of findings, patients were brought to the television studio for best possible lighting; as before, 2 cameras were used to facilitate illustration of the findings and the proper technique. However, some patients were filmed as inpatients using a single portable camera. Consent for videotaping was obtained from all patients.

A total of 26 patients were selected. To cover all possible scale scores, 18 were used for certification and 8 were used for training. The training patients were used during the instruction section; the script for this instruction section was edited and approved by the expert panel. The 18 certification patients were divided into 3 groups of 6 patients each, balanced for severity and stroke side.

To assess the reliability of the certification patients and confirm the use of the training video, DVDs were sent to 3 stroke centers known for prior use of the NIHSS: UCSD, University of Texas–Houston, and University of Cincinnati. To assure sufficient power for assessing overall NIHSS score reliability, each center was sent 51 DVDs with scoring sheets. The power analysis is based on the intraclass correlation coefficient for the overall NIHSS score and was performed using the software PASS. A sample size of 18 subjects with 51 observations per subject achieves 80% power to detect an intraclass correlation coefficient of at least 0.60 under the alternative hypothesis, assuming a null hypothesis correlation of 0.39 and a significance level of 0.05. The DVDs were distributed to raters, including neurologists, emergency department (ED) or other physicians, and nurses; previously uncertified examiners were encouraged to participate. Whether previously certified or not, each rater was asked to view the training video, and then score one of the 3 certification groups; each DVD envelope was labeled to indicate which patient group to use and groups were assigned at random among the 51 DVDs for each site. To avoid bias, only 2 UCSD staff involved in videotaping was obtained from all patients.

Reliability was studied for both the individual items of the NIHSS and the overall score. Scores for each of the individual items were tabulated. Agreement was assessed with the unweighted κ statistic (κ) for the case of multiple raters. Jackknife technique was used to compute the standard error of κ (estimate of κ) and 95% confidence intervals for κ was computed by the standard formula, $\kappa \pm 1.96 \text{ standard error (SE)} \times \sqrt{\frac{N}{4}}$. These methods are similar to the analytic methods described in the initial reliability paper and previously published findings. Agreement was assessed with the weighted κ statistic (κw) for the case of multiple raters. Pairwise comparisons were made using Fisher’s exact test for the case of multiple raters. The weighted κ statistic is calculated using weighted κ measures for each item item and adjusting for multiple comparisons using Bonferroni correction. In addition, the scatterplot of item scores for each subject was used to visually compare and confirm the reliability graphically and the consistency of the item score by subgroups.

Agreement on the overall NIHSS score was assessed with the intraclass correlation coefficients (ICC) obtained using a one-way random-effects regression model (model 1) for clustered data (with ratings nested within patients). The model used included a random effect for patients and assumed that the within-patient variance was the same across all patients. The ICC is calculated by expressing the between-patient variance as a percentage of the total variance (between-patient variance). It represents the correlation between the scores from 2 randomly chosen raters within a randomly chosen patient. Values of the ICC close to 1.0 indicate excellent agreement among the measurements within a patient. Standard error and 95% confidence intervals of the ICC were calculated using jackknife methods with the resampling done at the patient level.

To assess the effect of covariates (specialty, certification status, group) on the variability (and hence the ICC), we fit random-effects regression models for clustered data in which the within-patient and between-patient variance was allowed to vary across the subgroups of the covariate of interest (model 2). Comparing models 1 and 2 using the likelihood ratio test allows us to determine if the assumption of constant variance, and hence of constant ICC, among the subgroups of raters was valid. Similar to individual items, the scatterplot of overall score for each subject was used to visualize the variation of the total score by subgroups.

Results

We received 756 responses from an expected 918 (18 patients×51 raters), despite multiple requests to each test site. Among the 756 responses, there were 32 records containing missing data on at least one individual item (and hence the NIHSS overall score), so the statistical analysis is based on 724 complete records (79% response rate). Responses were received from 112 individual raters who each rated between 3 and 6 patients. As a result, each patient had somewhere between 33 and 52 ratings (unequal cluster sizes). Among the raters, 26% of all responses came from stroke nurses, 34% from ED/other physicians, and 40% from neurologists. One half (50%) of raters were previously NIHSS-certified. Item responses were tabulated, scoring performed as described, and agreement measured with unweighted κ coefficients for individual items and an intraclass correlation coefficient for the overall score.

Table 1 indicates the range of values obtained on each item over all 18 patients. In confirmation of the intended patient selection design, the table documents the presence of all possible responses on all individual scale items. The mean NIHSS total score was 8.8 ± 6.6 (median, 7; range, 1 to 33). The spread of responses in individual items and total scores looks similar among the subgroups, namely, sites, specialties, and prior NIHSS certification status.

Table 2 compares the agreement obtained using the unweighted κ from the 3 different studies: the current DVD study (18 patients), the first certification videotape (5 patients), and the second certification videotape with 6 patients. The agreements ranged from 0.21 (ataxia) to 0.92 (LOCC) using the DVD and the values were remarkably similar to those obtained previously using the videotapes, although weighted κ was used in the previous article. The agreements obtained from the DVD were closer to those obtained from tape 2 than tape one except for 2 items with poor concordance (facial weakness and ataxia). There are 2 reasons for this: (1) the video approach on tape 2 more closely resembled the approach used on the DVD; tape one used a significantly different approach; and (2) tape 2 was generally viewed 6 months after passing tape one, so all viewers of tape 2 were recently certified.

Among all 18 certification patients, the agreement was similar across all subgroups. There were no differences among nurses, ED physicians; among the study sites, or between novice raters and previously certified raters (data not shown).

Table 3 lists the intraclass correlation coefficient for the overall total NIHSS score and total NIHSS by subgroup. There was very good agreement in the total NIHSS score across all ratings (overall intraclass correlation coefficient of 0.94; 95% confidence interval, 0.84 to 1). There are no statistically significant differences in mean NIHSS scores by prior NIHSS certification status, site, specialty or group.
Although there were slight differences in ICC across covariates, in all cases, the agreement still remained very high. The ICC was slightly lower among nurses compared with the neurologists and other/ED physicians. Similarly, the raters from Cincinnati had slightly lower agreement scores compared with the raters in UCSD and Houston. However, none of these differences are considered to be clinically significant. These scores have been provided to the American Heart Association for scoring their online NIHSS training and certification procedure.

**Discussion**

We created a new certification and training DVD for the NIHSS using modern videography, a television studio set to...
TABLE 3. Intraclass Correlation Coefficients for NIHSS Total Score*

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>0.94</td>
<td>(0.84 to 1)</td>
</tr>
<tr>
<td><strong>Site</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Diego</td>
<td>0.94</td>
<td>(0.85 to 1)</td>
</tr>
<tr>
<td>Houston</td>
<td>0.96</td>
<td>(0.91 to 1)</td>
</tr>
<tr>
<td>Cincinnati</td>
<td>0.90</td>
<td>(0.71 to 1)</td>
</tr>
<tr>
<td><strong>Specialty</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nurse</td>
<td>0.89</td>
<td>(0.69 to 1)</td>
</tr>
<tr>
<td>Neurologist</td>
<td>0.96</td>
<td>(0.89 to 1)</td>
</tr>
<tr>
<td>ED/other MD</td>
<td>0.95</td>
<td>(0.89 to 1)</td>
</tr>
<tr>
<td><strong>Certification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certified</td>
<td>0.92</td>
<td>(0.79 to 1)</td>
</tr>
<tr>
<td>Not certified</td>
<td>0.95</td>
<td>(0.87 to 1)</td>
</tr>
<tr>
<td><strong>Group</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.88</td>
<td>(0.68 to 1)</td>
</tr>
<tr>
<td>B</td>
<td>0.98</td>
<td>(0.84 to 1)</td>
</tr>
<tr>
<td>C</td>
<td>0.91</td>
<td>(0.7 to 1)</td>
</tr>
</tbody>
</table>

*The table presents the ICCs for total score by overall, by site, specialty, certification status, and group. The ICCs for subgroups are calculated by model 1 (see Materials and Methods).

enlarge accurate examination recordings, and diligent attention to proper neurologic technique during the taping and editing. We carefully selected patients to span the range of all possible responses on all items, as verified in Table 1. The reliability of the new certification DVD among previously certified and noncertified users was essentially the same as found for the previous videotapes, indicating that the DVD is a valid and reliable replacement for the videotapes.

We found no differences when the DVD was used by neurologists, ED physicians, and nurses, suggesting that the NIHSS may be appropriate for use in clinical research trials, as well as in daily communication among healthcare providers, but the study may be underpowered to detect a subtle difference. In previous studies, agreement among stroke nurses and neurologists was generally good to excellent.6 Agreement among ED physicians is generally somewhat less than among stroke neurologists, but in our data, there was no difference. Larger trials of the DVD certification will be needed to discern more subtle differences among subgroups.

The DVD format has some advantages over videotape. The digital images can be loaded onto a web site, and the American Heart Association successfully implemented a web-based training campus using our images. This web site allows raters to view the training and certification patient videos online (http://www.asatrainingcampus.net). The DVD technology is more widely available now than videotapes, so NIHSS certification should be possible for many more years, even if videotapes become obsolete.

This study contains certain limitations, the most important of which is that the validation process was done in selected stroke centers. We chose 3 very experienced centers to obtain a best-case impression of how the DVD patients should be scored. However, these scores may not be generally applicable when novice users view the training DVD and then attempt certification. For example, agreement among the 3 stroke center directors was significantly higher than among all other groups. Therefore, we continue to collect scores to determine if the same scoring sheet generally works well outside of experienced centers. This DVD was designed for a single user to view at home or in an office. Its use in group settings is not validated, although this study is underway.

Another inherent limitation is that video technology is a poor substitute for direct examination. In the absence of widespread proctored certification, however, no other option is available. Video certification is now widely used in many disciplines with reasonable validity and reliability.2 It is likely that web-based video training and certification will become more widespread, because the cost efficiencies are significant.

As a result of the unbalanced group sizes, we could not use weighted κ statistics, as has been used in previous trials. Unweighted κ scores may underestimate agreement, yet in this study, the unweighted κ scores were comparable to the weighted scores obtained in previous studies. Therefore, the agreement among the viewers was at least as good, and likely better, than that seen previously with the videotapes.

Acknowledgments

This work was supported by the NINDS P50 NS044148 and the Veterans Affairs Medical Research Service. The authors thank the kind staff at San Diego Rehabilitation Institute, Dr Lance Stone, Medical Director, and the Grossmont Rehabilitation Institute, Dr Sherry Brabheny, Medical Director. The authors are especially grateful to the patients and caregivers who volunteered to participate in this project.

References

NIHSS Training and Certification Using a New Digital Video Disk Is Reliable
Patrick Lyden, Rema Raman, Lin Liu, James Grotta, Joseph Broderick, Scott Olson, Sandi Shaw, Judith Spilker, Brett Meyer, Marian Emr, Margo Warren and John Marler

*Stroke*. 2005;36:2446-2449; originally published online October 13, 2005;
doi: 10.1161/01.STR.0000185725.42768.92

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 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/36/11/2446

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org/subscriptions/