National Institutes of Health Stroke Scale Certification Is Reliable Across Multiple Venues

Patrick Lyden, MD; Rema Raman, PhD; Lin Liu, PhD; Marian Emr; Margo Warren; John Marler, MD

Background and Purpose—National Institutes of Health Stroke Scale certification is required for participation in modern stroke clinical trials and as part of good clinical care in stroke centers. A new training and demonstration DVD was produced to replace existing training and certification videotapes. Previously, this DVD, with 18 patients representing all possible scores on 15 scale items, was shown to be reliable among expert users. The DVD is now the standard for National Institutes of Health Stroke Scale training, but the videos have not been validated among general (ie, nonexpert) users.

Methods—We sought to measure interrater reliability of the certification DVD among general users using methodology previously published for the DVD. All raters who used the DVD certification through the American Heart Association web site were included in this study. Each rater evaluated one of 3 certification groups.

Results—Responses were received from 8214 raters overall, 7419 raters using the Internet and 795 raters using other venues. Among raters from other venues, 33% of all responses came from registered nurses, 23% from emergency department MD/other emergency department/other physicians, and 44% from neurologists. Half (51%) of raters were previously National Institutes of Health Stroke Scale-certified and 93% were from the United States/Canada. 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. In addition, agreement in this study was compared with the agreement obtained in the original DVD validation study to determine if there were differences between novice and experienced users. Kappas ranged from 0.15 (ataxia) to 0.81 (Item 1c, Level of Consciousness-commands [LOCC] questions). Of 15 items, 2 showed poor, 11 moderate, and 2 excellent agreement based on kappa scores. Agreement was slightly lower to that obtained from expert users for LOCC, best gaze, visual fields, facial weakness, motor left arm, motor right arm, and sensory loss. The intraclass correlation coefficient for total score was 0.85 (95% CI, 0.72 to 0.90). Reliability scores were similar among specialists and there were no major differences between nurses and physicians, although scores tended to be lower for neurologists and trended higher among raters not previously certified. Scores were similar across various certification settings.

Conclusions—The data suggest that certification using the National Institute of Neurological Disorders and Stroke DVDs is robust and surprisingly reliable for National Institutes of Health Stroke Scale certification across multiple venues. (Stroke. 2009;40:2507-2511.)

Key Words: clinimetrics ■ reliability ■ scales ■ stroke

Neurologists who care for patients with stroke are required to certify in use of the National Institutes of Health Stroke Scale (NIHSS) now that Disease-Specific Specialty Designation as a Primary Stroke Center is available from the Joint Commission.1,2 The NIHSS is a widely used stroke deficit assessment tool used in nearly all large clinical stroke trials to document baseline and outcome severity.3-5 A training and certification process exists to assure that raters use the NIHSS in a uniform manner6-7; videotapes were used for training and certification from 1988 to 2006. To update the training and certification process, the National Institute of Neurological Disorders and Stroke produced a DVD in 2006 that is distributed widely by the American Academy of Neurology, the American Heart Association, and the National Stroke Association. Originally the DVD was validated in 3 select stroke centers to obtain a best-case impression of how the DVD patients should be scored among expert users.8 The DVD was designed, however, for a nonexpert single user to view at home or in an office and the use among nonexperts has not been validated. In addition, the DVD certification in group settings is not validated. Also, scores may not be generally applicable when novice users view the training DVD and then attempt certification. Hence, we collected scores from

Received July 23, 2008; accepted August 8, 2008.
From the Departments of Neurosciences (P.L., R.R.) and Family and Preventive Medicine (R.R., L.L.), University of California–San Diego School of Medicine, San Diego, Calif; the Department Neurology (P.L.), Veteras Administration Medical Center, San Diego, Calif; and the National Institute of Neurological Disorders and Stroke (M.E., M.W., J.M.), Bethesda, Md.
Correspondence to Patrick Lyden, UCSD Stroke Center, OPC Third Floor, Suite #3, 200 W Arbor Drive, San Diego CA 92103. E-mail plyden@ucsd.edu
© 2009 American Heart Association, Inc.

Stroke is available at http://stroke.ahajournals.org

DOI: 10.1161/STROKEAHA.108.532069

2507
single use, group use, and a web site to determine the reliability of the DVD certification outside of experienced centers and across multiple venues.

Methods

The training DVD includes 18 patients divided into 3 groups balanced for severity and stroke side. Raters were asked to certify using one of the 3 patient groups. Details on the DVD and the certification method have been described.8

We obtained certification scores from users in the following venues: single user (home or desktop), small groups, large groups, and a web site. Single users took the DVD home or to an office, watched the training video, and then watched the certification video cases. Small group certifications occurred at single sites where the training video was shown and then no more than 12 users watched the certification video and marked score sheets individually. Large group certifications occurred at multiple sites where rater teams participating in a variety of clinical trials; the training video was shown and then certification patients were shown. In the large group settings, each user marked their own score sheet without discussion among other users. From all venues, score sheets were faxed to the University of California–San Diego Stroke Clinical Trial Coordinating Center for scoring using the published algorithm.7 The training/certification web site is sponsored by the American Heart Association. Users were encouraged to watch the training video over the Internet before certifying on one of the 3 certification groups; scores were recorded on the web site and then raw data were transmitted to the University of California–San Diego.

Descriptive analysis was performed on all data in the data set. The number of raters who certified using this DVD was tabulated by setting (individual, small group, investigator meeting, and web site) as well as specialty (RN, emergency department MD, neurology, other emergency department, other), prior certification status (yes, no), and country (US/Canada, others), if collected. Summaries of the individual item score as well as the total NIHSS were generated.

Reliability was assessed for the individual items of the NIHSS as well as the overall score. Scores of the individual items were tabulated. Agreement occurred individually among raters was assessed using the unweighted kappa statistic (κ) for multiple raters9 with a 95% CI obtained using the bootstrap resampling technique with 1000 replicates. The methods used here are similar to the methods used in the original DVD validation study to allow comparison between the 2 studies.8 In this study, the bootstrap technique was used instead of the jackknife technique because there are several instances when the jackknife technique was not appropriate.10 Agreement between this study and the original DVD study was considered to be statistically different if the estimated κ in the original study did not fall into the 95% CI for κ in this study. Using similar methods, reliability of the individual items was assessed separately for the subgroups of patients by setting as well as specialty, certification status, and country, if available. Comparison of κ statistics across subgroups was done using the bootstrap technique for correlated data.11 Ninety-five percent CIs for differences in κ between 2 subgroups were calculated. The Bonferroni correction was used to adjust for multiple comparisons within each subgroup comparison. In addition, the scatterplot of the item scores appeared similar among the subgroups, namely, sites, specialties, and prior NIHSS certification status.

Results

We received score sheets from 379 single users, 178 small group users, 238 large group users, and 7419 web users. Among the 49 284 expected responses (8214×6), we received 49 272 ratings (99.9% completion rate). Responses were received from 8214 individual raters (4796 raters scored patients in Group A, 2762 in Group B, and 656 in Group C) who each rated between 3 and 6 patients. As a result, each patient had somewhere between 655 and 4796 ratings (unequal cluster sizes). Among the raters who provided demographic information, 33% of all responses came from registered nurses, 23% from emergency department/other physicians, and 44% from neurologists. Most of the raters (93%) were from the United States and half of the raters on an average (51%) were previously NIHSS-certified. Item responses were tabulated, scoring performed as described previously, and agreement measured with unweighted κ coefficients for individual items and an ICC for the overall score.

Table 1 indicates the range of values obtained on each item over all 18 patients. The mean NIHSS total score was 8.0±6.6 (median, 7; range, 0 to 41). The spread of responses in individual items and total scores appeared similar among the subgroups, namely, sites, specialties, and prior NIHSS certification status.

Table 2 compares the agreement obtained using the unweighted κ from the current data set with that of the original DVD study.1 The agreements ranged from 0.15 (ataxia) to 0.81 (Item 1c, Level of Consciousness-commands [LOCC]) using the current data set. The agreements obtained from this group of raters were similar to that of the original DVD study on all items of the NIHSS except for 7 items with lower agreement (LOCC, best gaze, visual fields, facial weakness, motor left arm, motor right arm, and sensory loss).

Among all 18 certification patients, the agreement was similar across all subgroups and among all venues. Results were remarkably similar to the results in the original DVD validation study except for some small inconsistent differences across certain subgroups (data not shown). Agreement in 4 fields (LOCQ, LOCC, visual fields, and motor left leg) was higher in other countries compared with the United States/Canada. Among specialties, emergency department MDs had higher agreement in motor right leg compared with nurses; in LOCC, motor right leg and sensory loss compared with neurologists and in motor left leg and motor right leg compared with other specialties; nurses showed greater agreement in dysarthria compared with neurologists and in motor left arm and motor left leg when compared with other specialties. Agreement in LOCQ was higher in noncertified
raters than that in certified raters. Comparing venues, individual users showed higher agreement in extinction/neglect compared with the large group setting and higher agreement in visual fields and motor left arm compared with web users; in the large group setting, scores showed lower agreement in extinction/neglect compared with the web setting; the small group setting showed higher agreement in motor left arm than web users. There is no significant difference in agreement across 3 certification groups.

Table 3 lists the intraclass correlation coefficient for the overall total NIHSS score and total NIHSS by subgroup. There continues to be very good agreement in the total NIHSS score across all venues and subgroups (overall ICC of 0.85; 95% CI, 0.72 to 0.90). There are no statistically

<table>
<thead>
<tr>
<th>Item</th>
<th>Total Responses on This Item</th>
<th>Level of Responses, N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a LOC</td>
<td>49272</td>
<td>43564 (88)</td>
</tr>
<tr>
<td>1b LOC questions</td>
<td>49272</td>
<td>28395 (58)</td>
</tr>
<tr>
<td>1c LOC command</td>
<td>49272</td>
<td>45815 (93)</td>
</tr>
<tr>
<td>2 Gaze</td>
<td>49271</td>
<td>43908 (89)</td>
</tr>
<tr>
<td>3 Visual fields</td>
<td>49269</td>
<td>39378 (80)</td>
</tr>
<tr>
<td>4 Facial weakness</td>
<td>49272</td>
<td>23114 (47)</td>
</tr>
<tr>
<td>5a Motor left arm</td>
<td>49258</td>
<td>34310 (70)</td>
</tr>
<tr>
<td>5b Motor right arm</td>
<td>49261</td>
<td>36828 (75)</td>
</tr>
<tr>
<td>6a Motor left leg</td>
<td>49261</td>
<td>27665 (56)</td>
</tr>
<tr>
<td>6b Motor right leg</td>
<td>49264</td>
<td>27086 (55)</td>
</tr>
<tr>
<td>7 Ataxia</td>
<td>49243</td>
<td>29828 (61)</td>
</tr>
<tr>
<td>8 Sensory</td>
<td>49264</td>
<td>4645 (9.4)</td>
</tr>
<tr>
<td>9 Aphasia</td>
<td>49264</td>
<td>27877 (57)</td>
</tr>
<tr>
<td>10 Dysarthria</td>
<td>49256</td>
<td>27222 (55)</td>
</tr>
<tr>
<td>11 Extinction</td>
<td>49256</td>
<td>33012 (67)</td>
</tr>
</tbody>
</table>

LOC indicates level of consciousness.

*The agreement (unweighted \( \kappa \) and 95% CI) among all raters for 15 NIHSS items using new training and certification DVD. For comparison, the agreement on individual items in original DVD validation study is given. To assess the effect of subgroups, we used a pairwise comparison with a Bonferroni adjustment to account for the multiple comparisons.
significant differences in mean NIHSS scores by country and prior NIHSS certification status. There was a statistically significant interaction between specialty and setting in mean NIHSS scores ($P=0.046$); however, there were no clinically significant differences. Although there were slight differences in ICC across covariates, in all cases, the agreement remained very high. Agreement was lower among raters from the United States/Canada compared with the raters from other countries. The ICC was slightly lower among neurologists compared with the nurses, emergency department MDs, other MDs, and other physicians. Similarly, the raters with prior certification had slightly lower agreement than those who were not certified previously. The ICC was slightly lower in the case of small group setting as compared with individual, investigator meeting setting, or web users. The ICCs for certification Groups A and B were slightly lower than Group C.

**Discussion**

Our data show that NIHSS training and certification using the DVD is valid and reliable among general users. The certification process showed remarkable consistency across widely differing venues, including single users, small groups, large groups, and certification data from the American Heart Association web site. The individuals in this study included novice users—who viewed the training video and then attempted certification—as well as previously certified users. The reliability assessments of this certification DVD among these novice users were similar to what was found using the experienced stroke centers, indicating that the DVD is a surprisingly valid and reliable replacement for the previous videotapes. The agreement among the items was similar whether it was used by a single user or in a group setting.

We found no differences in the ICC of the total NIHSS when the DVD was used by neurologists, emergency department 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. Agreement among those identifying themselves as neurologists was slightly lower than individuals identifying themselves as registered nurses, emergency department/other MDs, or other specialties, but the results were statistically similar and generally excellent. Agreement across various settings was similar and generally moderate to excellent.

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. 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 most of the raters were from the United States and Canada. We were able to determine that the scoring sheet works well for novice as well as experienced users in North America. However, these scores may not be generally applicable for non-English-speakers or raters in other countries. Therefore, we continue to collect scores from the web site to determine if the same scoring sheet generally works well outside of North America. 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. It is likely that web-based video training and certification will become more widespread, because the cost efficiencies are significant. Finally, the web site does not require viewing of the training video before attempted certification, so an unknown number of novice users could have tried to certify without proper training.

Due to the unbalanced group sizes, small cells for item statistics, unweighted $\kappa$ scores may underestimate agreement, yet in this study, the unweighted $\kappa$ scores were comparable to the unweighted scores obtained in the primary DVD study and the weighted scores obtained in previous videotape studies. Therefore, the agreement among the viewers was at least as good and likely better than that seen previously with the videotapes. Agreement using the DVD continues to be surprisingly good and consistent among experienced as well as novice users.

**Acknowledgments**

We acknowledge the diligent effort and expertise of Ms Alyssa Chardi and Karen Rapp, RN.
Sources of Funding
This work was supported by National Institute of Neurological Disorders and Stroke P50 NS044148 and the Veterans Affairs Medical Research Service.

Disclosures
None.

References
National Institutes of Health Stroke Scale Certification Is Reliable Across Multiple Venues
Patrick Lyden, Rema Raman, Lin Liu, Marian Emr, Margo Warren and John Marler

Stroke, published online June 11, 2009;
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2009 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/early/2009/06/11/STROKEAHA.108.532069.citation

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/