ORIGINAL ARTICLE
J Clin Neurol 2012;8:177-183

Print ISSN 1738-6586 / On-line ISSN 2005-5013
http://dx.doi.org/10.3988/jcn.2012.8.3.177

Open Access

Validity and Reliability of a Korean Version of the National Institutes of Health Stroke Scale

Mi Sun Oh,[a] Kyung-Ho Yu,[a] Ju-Hun Lee,[b] San Jung,[c] Im-Suck Ko,[d] Joon-Hyun Shin,[b] Soo-Jin Cho,[e] Hui-Chul Choi,[f] Hyang Hee Kim,[g] Byung-Chul Lee[a]

[a]Department of Neurology, Hallym University Sacred Heart Hospital, Hallym University College of Medicine, Anyang, Korea
[b]Department of Neurology, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea
[c]Department of Neurology, Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea
[d]Department of Neurology, National Medical Center, Seoul, Korea
[e]Department of Neurology, Hangang Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Korea
[f]Department of Neurology, Chuncheon University Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, Korea
[g]Department of Neurology, Yonsei University College of Medicine, Seoul, Korea

Background and Purpose The National Institutes of Health Stroke Scale (NIHSS) is a clinical assessment tool that is widely used in clinical trials and practice to evaluate stroke-related neurological deficits. The aim of this study was to determine the validity and reliability of the Korean version of the NIHSS (K-NIHSS) for evaluating Korean stroke patients.

Methods The K-NIHSS was translated and adapted with regard to cultural and linguistic peculiarities. To examine its content validity, we quantified the Content Validity Index (CVI), which was rated by 11 stroke experts. The validity of the K-NIHSS was assessed by comparison with the Glasgow Coma Scale (GCS), the modified Rankin Scale (mRS), and the Barthel Index. The reliability of the K-NIHSS was evaluated using the unweighted kappa statistics for multiple raters and an intraclass correlation coefficient (ICC).

Results The CVI of the K-NIHSS reached 0.91-1.00. The median K-NIHSS score at baseline was 3 (interquartile range, 2-7), and the mean±SD score was 6.0±6.6. The baseline K-NIHSS had a significantly negative correlation with the GCS at baseline and the Barthel Index after 90 days. The K-NIHSS also had a significantly positive correlation with the mRS after 90 days. Facial paresis and dysarthria had moderate interrater reliability (unweighted kappa, 0.41-0.60); kappa values were substantial to excellent (unweighted kappa, >0.60) for all the other items. The ICC for the overall K-NIHSS score was 0.998. The intrarater reliability was acceptable, with a median kappa range of 0.524-1.000.

Conclusions The K-NIHSS is a valid and reliable tool for assessing neurological deficits in Korean acute stroke patients. **J Clin Neurol 2012;8:177-183**

Key Words NIH Stroke Scale, validity, reliability.

Received September 8, 2011
Revised March 20, 2012
Accepted March 20, 2012

Correspondence
Byung-Chul Lee, MD
Department of Neurology,
Hallym University Sacred Heart Hospital, Hallym University College of Medicine,
22 Gwanpyeong-ro 170 beon-gil,
Dongan-gu, Anyang 431-070,
Korea
Tel +82-31-380-3741
Fax +82-31-380-4659
E-mail ssbrain@hallym.ac.kr

Introduction

The National Institutes of Health Stroke Scale (NIHSS) is the most widely used clinical assessment tool for evaluating stroke-related neurological deficits.[1-4] The scale can be easily learned, performed quickly, and used by nonneurologists as well as neurologists with appropriate training. The scale has high reliability and is well validated for use in prospective clinical research.[2,5-9] Moreover, the baseline NIHSS score is well known to accurately predict short- and long-term outcomes of stroke patients in clinical trials and in clinical practice.[3,4,10,11]

Training and certification in the use of the NIHSS are in-

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

creasingly required as prerequisites for participation in multicenter stroke clinical trials. A training and certification program using a validated English video or DVD has been developed.[5,7] In addition, online NIHSS certification has been available since 2005 (http://nihss-english.trainingcampus.net).[12] However, several items used to examine aphasia and dysarthria in the original NIHSS might not be appropriate for languages other than English as a result of cultural peculiarities. Consequently, several non-English-language versions of the NIHSS have been translated and validated in the literature.[13-17] With regard to the Korean language in particular, in addition to cultural peculiarities there are linguistic differences between Korean and English. A valid and reliable Korean version of the NIHSS (K-NIHSS) is needed to enable Korean-speaking physicians to assess Korean-speaking patients. In this report we describe our development and validation of such a tool.

Methods

Development of the K-NIHSS

Two stroke experts translated the original English-language NIHSS (E-NIHSS) into an initial Korean version by a consensus review. A linguistic expert was consulted to build an adequate list of sentences, pictures, and words, enabling translation and adaptation of the items describing aphasia and dysarthria in the E-NIHSS. The linguistic expert selected the drawing and reading items from the Korean Boston Naming Test[18] and adapted them for assessment of aphasia and dysarthria in the K-NIHSS. The main rationale for the translation and adaptation was to replace the English items with terms that were more culturally and linguistically appropriate, while maintaining phonemes, word lengths, and numbers of syllables similar to those in the original NIHSS items (Supplemental Table 1 and 2 detail the modification of each item). For example, to evaluate aphasia, pictures of a hammock, a feather, and a kitchen were replaced in the K-NIHSS with pictures of a traffic light, a crutch, and a road [Appendix 1 (http://www.thejcn.com/journal/view.php?myear=2012&vol=8&no=3&spage=177); Figs. 1 and 2]. The final, revised K-NIHSS was evaluated by a panel of four stroke neurologists and five stroke research nurses who had not participated in the development of the initial version of the K-NIHSS. They independently reviewed the initial version of the K-NIHSS and, by consensus, resolved any discrepancies between the original NIHSS and the K-NIHSS.

To examine the content validity of the final version of the K-NIHSS regarding whether each of its item accurately reflected the original version, 11 experts (comprising 6 stroke neurologists and 5 stroke research nurses) were recruited from 5 university hospitals. Each of the experts had more than 5 years of experience in the care of acute stroke patients and had not participated in the development and revision of the K-NIHSS. Each expert independently reappraised the content validity of the K-NIHSS. To quantify its content validity, we used the Content Validity Index (CVI), which is the proportion of items receiving a rating by the experts higher than 3 points on a 4-point ordinal rating scale; a rating of 1 denotes an irrelevant item, and a rating of 4 denotes an extremely relevant item. Items with a CVI of at least 0.78 were accepted as valid.[19,20]

Assessing the validity of the K-NIHSS

Subjects

For the prospective validity study, 207 patients with acute ischemic stroke were enrolled from 4 large training hospitals between March 2007 and October 2007. All of the patients were admitted within 5 days of the onset of symptoms, and the responsible acute infarction corresponding to the stroke index was confirmed by brain magnetic resonance imaging within 24 hours of the patient's admission. The study protocol to assess the validity of the K-NIHSS was approved by the local ethics committee of Hallym University Sacred Heart Hospital, and all the subjects provided written informed consent to participate before the commencement of the study.

Construct validity

The construct validity of the K-NIHSS was determined by comparison with the Glasgow Coma Scale (GCS). The GCS provides a global rating of a patient's status and is a reliable tool for estimating status in patients with acute neurological diseases.[21,22] One stroke neurologist at each site used the K-NIHSS and the GCS to assess each patient within 48 hours of admission. We calculated the Spearman correlation coefficient between the K-NIHSS and the GCS.

Predictive validity

The predictive validity of the baseline K-NIHSS was assessed by comparison with the 90-day outcome, as measured by the modified Rankin Scale (mRS)[23] and the Barthel Index (BI).[24] At 90 days after the onset of symptoms, two investigators performed the mRS and the BI measurements. One stroke neurologist at each site performed the primary 90-day assessment through a face-to-face interview, while a second assessment was performed through a telephone interview by an investigator who was well trained in neurological examinations and blind to the clinical findings and the rating of the primary investigator. In the statistical analyses we used the Spearman correlation coefficient to estimate the association between the overall K-NIHSS score at baseline and the total BI and mRS scores after 90 days.

Assessing the reliability of the K-NIHSS

Subjects

A total of 30 patients included in the K-NIHSS videos represented every choice in every scale item to cover all stroke severities. We placed 11 patients in the training video and 19 patients in the certification video. Power analysis was calculated by the intraclass correlation coefficient (ICC) for the overall NIHSS score and was performed using the PASS program for sample size estimation; an ICC value in the range of 0.5-0.6 indicates moderate agreement.[25] A sample size of 19 subjects, with 21 raters per subject, was estimated to achieve 82% power to detect an ICC of at least 0.6 under the alternative hypothesis, assuming a null hypothesis correlation of 0.39 with a significance level of 0.05. Consents for videotaping and use of the videotapes for the training and certification program were obtained from all of the patients.

The training and certification videos using the K-NIHSS were produced by a professional video production team using a set equipped with appropriate lighting and recording capabilities built at Hallym University Sacred Heart Hospital. The appropriate documentation was acquired by filming the patients using two cameras.[5]

Interobserver and intraobserver agreement

Twenty-one raters, comprising 7 stroke experts, 12 neurology residents, and 2 stroke nurses, were selected as raters to determine agreement among the raters (interrater reliability) and between 2 training and certification sessions (intrarater reliability). Two training and certification workshops using the K-NIHSS video were held in May 2010. The raters reviewed the training video and scored the 19 patients in the certification video on the same day of the first workshop. At a second workshop held 2 weeks later, the same 21 raters again reviewed the training video and then scored the 19 patients in the certification video.

The level of agreement for the individual items of the K-NIHSS among the 21 raters was evaluated using unweighted kappa statistics for multiple raters,[26,27] with 95% confidence intervals (CIs) obtained using the bootstrap resampling technique with 1000 replicates.[5,7] To facilitate the comparison of our results with those of previous studies, our statistical methodology was similar to that employed in those studies.[5,7,17,28] The agreement was defined as excellent if the unweighted kappa was >0.80, substantial if it was between 0.61 and 0.80, moderate if it was between 0.41 and 0.60, fair if it was between 0.21 and 0.40, and slight or poor if it was <0.20.[29] The agreement on the overall score of the K-NIHSS was assessed with an ICC for a continuous variable obtained using a one-way random effects model for repeated measurements with continuous outcomes.[30] The bootstrap resampling technique was used to obtain 95% CIs for the ICC. An ICC of 1 suggests perfect reliability, and an ICC of 0.8 is generally considered to represent excellent reliability.[25]

Intrarater reliability was calculated by comparing the 2 repeated sets of scores for the 19 patients that the 21 raters individually assessed 2 weeks apart.[2] The medians of the individual kappa and ICC values were calculated to obtain the overall kappa and ICC values for each type of analysis.

Results

Subject characteristics for the validity study

During an 8-month period, 207 patients with acute ischemic stroke were prospectively recruited. The patients were aged 67.1±11.6 years (mean±SD), and 94 (45%) were women. Among all the 207 patients, the median K-NIHSS score at baseline was 3 (interquartile range, 2-7; range, 1-35), and the score was 6.0±6.6.

Seven patients (3.3%) were lost to follow-up. For 97 patients (47%), we performed both a face-to-face interview and a telephone interview to assess the functional outcome measures (by BI and mRS) at 90 days after the onset of symptoms. The remaining patients (n=103, 49.7%) could be assessed only by a telephone interview at the 90-day time point.

Content validity

On a 4-point rating scale the mean scores of the CVI for each item ranged from 3.46 to 3.73. Only one item (the visual field defect) was scored as a 2 by one rater. Based on the appraisal of the 11 experts, the individual K-NIHSS CVIs were 0.91-1.00 (Table 1).

Construct validity

The correlation coefficient between the K-NIHSS and the GCS was significantly negative (Spearman rho=-6.71, p<0.001), indicating that the construct validity of the K-NIHSS was acceptable.[24]

Predictive validity

A follow-up performed 90 days after the telephone interviews with 200 patients revealed that 13 (6.5%) of the patients had died and 19 patients (9.5%) had an mRS score of 0, while 94 (47.0%), 51 (25.5%), and 23 (11.5%), had mRS scores of 1 or 2, 3 or 4, and 5, respectively. The median mRS was 2.0 (interquartile range, 1-4; 2.5±1.7). The BI was 78.0±16.9.

In a face-to-face interview, the Spearman correlation coefficient between the K-NIHSS at baseline and the mRS at the 90-day time point was significantly positive (Spearman rho= 0.600; p<0.001). The K-NIHSS also showed a significantly

positive correlation with the mRS, as determined by telephone interview after 90 days (Spearman rho=0.708, $p<0.001$). The correlation with the BI at 90 days was significantly negative (Spearman rho=-0.647, $p<0.001$).

Subject characteristics for the reliability study

Thirty patients (age, 69.2±7.1 years; 8 women, 42.1%) were enrolled to produce the training and certification videos for the K-NIHSS. The median time between symptom onset and videotaping was 9 days (range, 5-16 days). The median K-NIHSS score was 8.0 (with an interquartile range of 4.0-15.0; range, 1-20; 8.8±5.6).

We obtained 5985 responses from the 21 experts who took part in the training and certification workshop (15 items×19 patients×21 raters). There were no missing data (100% response rate; Table 2). The range of scores acquired for each item for the 19 patients is given in Table 2. We obtained nearly all possible responses on the individual scale scores, except for a score of 4 on the level of consciousness and a score of 3 on the visual field.

Table 1. CVI of the individual items of the K-NIHSS, as evaluated by 11 experts

Items		Mean of scores	Number rating above 3 points	CVI
1a	LOC	3.55	11	1
1b	LOC questions	3.55	11	1
1c	LOC command	3.45	11	1
2	Gaze	3.55	11	1
3	Visual fields	3.36	10	0.91
4	Facial weakness	3.64	11	1
5a	Motor left arm	3.73	11	1
5b	Motor right arm	3.73	11	1
6a	Motor left leg	3.73	11	1
6b	Motor right leg	3.73	11	1
7	Ataxia	3.45	11	1
8	Sensory	3.55	11	1
9	Aphasia	3.55	11	1
10	Dysarthria	3.73	11	1
11	Extinction	3.45	11	1

CVI: Content Validity Index, K-NIHSS: Korean version of the National Institutes Health Stroke Scale, LOC: level of consciousness.

Interrater agreement

The values indicating interobserver agreement for each item of the K-NIHSS, obtained using the unweighted kappa for multiple raters, are given in Table 3. The unweighted kappa scores were lowest for facial paresis (0.439) and dysarthria (0.465), and highest for the level of consciousness commands (0.950) and sensory function (0.911). The ICC for the overall score on the K-NIHSS was 0.998 (95% CI, 0.910-0.999). The reliability of the K-NIHSS was equal to or superior to that of the E-NIHSS (Table 3).

Intrarater agreement

Table 4 lists the results for the intrarater agreement for the K-NIHSS. The median scores of intrarater reliability for each item of the K-NIHSS ranged from 0.524 to 1.00. The ICC of the overall score on the K-NIHSS was 0.969 (with an interquartile range of 0.957-0.974).

The intrarater reliability showed moderate-to-perfect agreement between two repeated ratings.

Table 2. Distribution of responses for individual items of the K-NIHSS

Item		Total responses on this tem	Level of Responses, *n* (%)					
			0	1	2	3	4	UN
1a	LOC	399	349 (87.5)	40 (10.0)	10 (2.5)	0 (0)	-	
1b	LOC questions	399	191 (47.9)	145 (36.3)	63 (15.8)	-	-	
1c	LOC command	399	387 (97.0)	10 (2.5)	2 (0.5)	-	-	
2	Gaze	399	336 (84.2)	62 (15.5)	1 (0.3)	-	-	
3	Visual fields	399	316 (79.2)	52 (13.0)	31 (7.8)	0 (0)	-	
4	Facial weakness	399	93 (23.3)	135 (33.8)	168 (42.1)	3 (0.8)	-	
5a	Motor left arm	399	230 (57.6)	21 (5.3)	30 (7.5)	51 (12.8)	67 (17.8)	
5b	Motor right arm	399	281 (70.4)	75 (18.8)	6 (1.5)	23 (5.8)	14 (3.5)	
6a	Motor left leg	399	230 (57.6)	22 (5.5)	23 (5.8)	84 (21.0)	40 (10.0)	
6b	Motor right leg	399	250 (62.6)	106 (26.6)	20 (5.0)	17 (4.3)	6 (1.5)	
7	Ataxia	399	277 (69.4)	55 (13.8)	46 (11.5)	-	-	21 (5.3)
8	Sensory	399	169 (42.4)	210 (52.6)	20 (5.0)	-	-	
9	Aphasia	399	251 (62.9)	89 (22.3)	49 (12.3)	10 (2.5)	-	
10	Dysarthria	399	47 (11.8)	316 (79.2)	36 (9.0)	-	-	
11	Extinction	399	319 (77.7)	36 (9.0)	53 (13.3)	-	-	

K-NIHSS: Korean version of the National Institutes Health Stroke Scale, LOC: level of consciousness, UN: untestable.

Table 3. Inter-observer agreement for individual items and overall scores for the K-NIHSS compared with the E-NIHSS

Item		K-NIHSS 21 raters, 19 patients		E-NIHSS (Lyden, 2005) 51 raters, 18 patients	
		Kappa	95% CI	Kappa	95% CI
1a	LOC	0.891	0.867-0.914	0.460	0.39-0.53
1b	LOC questions	0.882	0.858-0.905	0.770	0.64-0.90
1c	LOC command	0.950	0.921-0.982	0.920	0.75-1.0
2	Gaze	0.613	0.582-0.643	0.700	0.39-1.0
3	Visual fields	0.615	0.591-0.639	0.720	0.57-0.87
4	Facial weakness	0.439	0.418-0.462	0.380	0.27-0.49
5a	Motor left arm	0.875	0.857-0.894	0.650	0.51-0.79
5b	Motor right arm	0.829	0.807-0.851	0.720	0.54-0.79
6a	Motor left leg	0.899	0.880-0.918	0.640	0.51-0.77
6b	Motor right leg	0.899	0.880-0.918	0.640	0.53-0.72
7	Ataxia	0.655	0.634-0.676	0.210	0.12-0.30
8	Sensory	0.911	0.781-0.990	0.730	0.53-0.93
9	Aphasia	0.641	0.617-0.664	0.640	0.53-0.75
10	Dysarthria	0.465	0.441-0.489	0.560	0.39-0.73
11	Extinction	0.806	0.782-0.830	0.570	0.40-0.74
Overall scores		ICC	95% CI	ICC	95% CI
		0.998	0.997-0.999	0.940	0.84-1.00

CI: confidence interval, E-NIHSS: English-language National Institutes Health Stroke Scale, ICC: intraclass correlation coefficient, K-NIHSS: Korean version of the National Institutes Health Stroke Scale, LOC: level of consciousness.

Discussion

In this study we developed a valid and reliable K-NIHSS and a training and certification program to enable the K-NIHSS to be applied to Korean stroke patients. The scale had good validity, correlated well with the GCS scale at baseline and outcome scales (mRS and BI) after 90 days, and showed high interrater and intrarater reliability.

A previous Korean version of the NIHSS has been published by rehabilitation physicians, but its predictive validity was not assessed, and appropriate statistical methods were not used to evaluate its reliability.[31] The interrater and intrarater reliability were evaluated in that study using the Spearman correlation coefficient, and only chronic stroke patients were enrolled-the scale has not been applied to patients with acute stroke.[31] Moreover, direct translation does not correctly account for any cultural peculiarities that might affect scoring on a particular scale. Unlike many other languages whose word order or articulation are similar to that of English, there are substantial linguistic differences between Korean and English. Regarding articulation, for the K-NIHSS we selected Korean words that are approximately as difficult to articulate as those in the English version, possibly because they are pronounced using the same vocal structures (the tongue, palate, and lips), and that have the same word length or number of syllables. Regarding the differences in word order between the two languages, Korean sentences were chosen to match the length and structure of those in the English version. Our study's strength lies in the K-NIHSS being constructed in a culturally and linguistically more appropriate format and is therefore of good quality.

A CVI of 0.91 was obtained for the visual field defects assessment item; the CVI was 1 for all of the other items. However, according to Waltz and Bausell,[19] the CVI of a measure evaluated by seven experts must be at least 0.80 to be considered acceptable.[20] The baseline K-NIHSS was strongly correlated with the patient's global status, as assessed by the GCS within 7 days after the onset of symptoms. In addition, the K-NIHSS exhibited an adequate-to-excellent correlation with the outcome scales (the mRS and BI) at the 90-day follow-up. The findings of our study establish that the K-NIHSS not only effectively measures the degree of acute neurological deficit, but also strongly predicts the outcome at 90 days, in keeping with the properties of the original NIHSS.[4,9]

Substantial-to-excellent values of agreement were established on 13 items of the K-NIHSS (unweighted kappa, >0.60), 2 items (facial paresis and dysarthria) showed moderate agreement (unweighted kappa, 0.41-0.60), and no item showed a fair or poor agreement (unweighted kappa, <0.40). Facial paresis and dysarthria also yielded low values of kappa in previous studies,[5,7,28] consistent with the finding that facial paresis and dysarthria are difficult to score because videography cannot portray these items as accurately as direct examination.[32] The interobserver agreement obtained for the K-NIHSS was higher than that of the original NIHSS on all but three items (best gaze, visual field, and dysarthria).[1,7,10,12] The higher reli-

Table 4. Intra-observer agreement for individual items and overall scores on the K-NIHSS

Item		K-NIHSS	
		Kappa, median	Kappa, interquartile range
1a	LOC	1.000	1.000, 0.974
1b	LOC questions	0.827	0.659, 0.895
1c	LOC command	1.000	1.000, 1.000
2	Gaze	0.524	0.428, 0.618
3	Visual fields	0.616	0.510, 0.814
4	Facial weakness	0.546	0.408, 0.589
5a	Motor left arm	0.878	0.825, 1.000
5b	Motor right arm	0.891	0.781, 0.974
6a	Motor left leg	0.912	0.829, 0.916
6b	Motor right leg	0.802	0.711, 0.907
7	Ataxia	0.726	0.630, 0.861
8	Sensory	0.947	0.799, 1.000
9	Aphasia	0.627	0.548, 0.776
10	Dysarthria	0.559	0.406, 0.778
11	Extinction	0.852	0.710, 1.000
		ICC, median	ICC, Interquartile range
Overall scores		0.969	0.957, 0.974

ICC: intraclass correlation coefficient, K-NIHSS: Korean version of the National Institutes Health Stroke Scale, LOC: level of consciousness.

ability of the K-NIHSS relative to the E-NIHSS might be due to our intensive examiner training. There have been some reports of intensive training positively affecting the reliability of the NIHSS.[5,33] In addition, training might influence interrater reliability because the raters were gathered in the same place to assess the patients. The ICC of the overall score did not differ between the K-NIHSS and E-NIHSS.[5,7] The intraobserver agreement for each item and the overall score on the K-NIHSS were similar to those obtained in previous studies of the E-NIHSS.[1,2,5] Our study demonstrates a high level of reliability for the K-NIHSS that is comparable with other-language versions of the NIHSS.[7,13,15-17,28]

There are several limitations to our study. First, we enrolled acute ischemic stroke patients within 5 days after symptom onset, and hence our results might not be applicable to other stroke types, such as hemorrhagic stroke. Second, two stroke-research nurses were the only nonneurologists who participated in the workshop. We could not register nonneurologists from various related fields, including the physical medicine and rehabilitation, emergency, and nursing departments. In previous studies using the NIHSS, the interobserver agreement was similar between neurologists and nonneurologists.[7,12,28,34] Thus, further study is needed to verify that the K-NIHSS may be generally used by nonneurologists and nonphysicians. Finally, as noted by Lyden et al.,[7,12] video technology has inherent limitations and is a poor substitute for direct examination.

In conclusion, the K-NIHSS was proven to be a valid and reliable tool for examining Korean stroke patients when used by trained neurologists. We hope that the K-NIHSS will become widely used in clinical practice by Korean-speaking physicians examining Korean-speaking patients with acute stroke and in acute stroke clinical trials on Korean stroke patients. The instruction and testing materials for the K-NIHSS are provided in Appendix 1. The Web site for training and certification in the use of the K-NIHSS is http://www.stroke-crc.or.kr/nihss, which is sponsored by the Clinical Research Center for Stroke.

Conflicts of Interest

The authors have no financial conflicts of interest.

REFERENCES

1. Goldstein LB, Bertels C, Davis JN. Interrater reliability of the NIH stroke scale. *Arch Neurol* 1989;46:660-662.
2. Albanese MA, Clarke WR, Adams HP Jr, Woolson RF. Ensuring reliability of outcome measures in multicenter clinical trials of treatments for acute ischemic stroke. The program developed for the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Stroke* 1994;25:1746-1751.
3. Kasner SE, Chalela JA, Luciano JM, Cucchiara BL, Raps EC, McGarvey ML, et al. Reliability and validity of estimating the NIH stroke scale score from medical records. *Stroke* 1999;30:1534-1537.
4. Adams HP Jr, Davis PH, Leira EC, Chang KC, Bendixen BH, Clarke WR, et al. Baseline NIH Stroke Scale score strongly predicts outcome after stroke: A report of the Trial of Org 10172 in Acute Stroke Treatment (TOAST). *Neurology* 1999;53:126-131.
5. Lyden P, Brott T, Tilley B, Welch KM, Mascha EJ, Levine S, et al. Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. *Stroke* 1994;25:2220-2226.
6. Goldstein LB, Samsa GP. Reliability of the National Institutes of Health Stroke Scale. Extension to non-neurologists in the context of a clinical trial. *Stroke* 1997;28:307-310.
7. Lyden P, Raman R, Liu L, Grotta J, Broderick J, Olson S, et al. NIHSS training and certification using a new digital video disk is reliable. *Stroke* 2005;36:2446-2449.
8. Hacke W, Donnan G, Fieschi C, Kaste M, von Kummer R, Broderick JP, et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet* 2004;363:768-774.
9. Lyden P, Lu M, Jackson C, Marler J, Kothari R, Brott T, et al. Underlying structure of the National Institutes of Health Stroke Scale: results of a factor analysis. NINDS tPA Stroke Trial Investigators. *Stroke* 1999;30:2347-2354.
10. Lyden PD, Lu M, Levine SR, Brott TG, Broderick J; NINDS rtPA Stroke Study Group. A modified National Institutes of Health Stroke Scale for use in stroke clinical trials: preliminary reliability and validity. *Stroke* 2001;32:1310-1317.
11. Schlegel D, Kolb SJ, Luciano JM, Tovar JM, Cucchiara BL, Liebeskind DS, et al. Utility of the NIH Stroke Scale as a predictor of hospital disposition. *Stroke* 2003;34:134-137.
12. Lyden P, Raman R, Liu L, Emr M, Warren M, Marler J. National Institutes of Health Stroke Scale certification is reliable across multiple venues. *Stroke* 2009;40:2507-2511.
13. Nilanont Y, Phattharayuttawat S, Chiewit P, Chotikanuchit S, Limsriwilai J, Chalernpong L, et al. Establishment of the Thai version of

National Institute of Health Stroke Scale (NIHSS) and a validation study. *J Med Assoc Thai* 2010;93 Suppl 1:S171-S178.
14. Sun TK, Chiu SC, Yeh SH, Chang KC. Assessing reliability and validity of the Chinese version of the stroke scale: scale development. *Int J Nurs Stud* 2006;43:457-463.
15. Domínguez R, Vila JF, Augustovski F, Irazola V, Castillo PR, Rotta Escalante R, et al. Spanish cross-cultural adaptation and validation of the National Institutes of Health Stroke Scale. *Mayo Clin Proc* 2006; 81:476-480.
16. Cincura C, Pontes-Neto OM, Neville IS, Mendes HF, Menezes DF, Mariano DC, et al. Validation of the National Institutes of Health Stroke Scale, modified Rankin Scale and Barthel Index in Brazil: the role of cultural adaptation and structured interviewing. *Cerebrovasc Dis* 2009;27:119-122.
17. Pezzella FR, Picconi O, De Luca A, Lyden PD, Fiorelli M. Development of the Italian version of the National Institutes of Health Stroke Scale: It-NIHSS. *Stroke* 2009;40:2557-2559.
18. Kang Y, Kim H, Na DL. Parallel short forms for the Korean-Boston Naming Test (K-BNT). *J Korean Neurol Assoc* 2000;18:144-150.
19. Waltz CF, Bausell RB. *Nursing Research: Design, Statistics, and Computer Analysis*. Philadelphia: Davis, 1981.
20. Lynn MR. Determination and quantification of content validity. *Nurs Res* 1986;35:382-385.
21. Rowley G, Fielding K. Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users. *Lancet* 1991;337: 535-538.
22. Fischer J, Mathieson C. The history of the Glasgow Coma Scale: implications for practice. *Crit Care Nurs* Q 2001;23:52-58.
23. Wilson JT, Hareendran A, Hendry A, Potter J, Bone I, Muir KW. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke* 2005;36:777-781.
24. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J* 1965;14:56-61.
25. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* 2nd ed. Upper Saddle River: Prentice Hall Health, 2000;560-567.
26. Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York: John Wiley & Sons, 1981.
27. Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioral Sciences.* 2nd ed. New York: McGraw-Hill, 1988.
28. Cheung RT, Lyden PD, Tsoi TH, Huang Y, Liu M, Hon SF, et al. Production and validation of Putonghua- and Cantonese-Chinese language National Institutes of Health Stroke Scale training and certification videos. *Int J Stroke* 2010;5:74-79.
29. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363-374.
30. Shoukri MM. *Measures of Interobserver Agreement.* Boca Raton: Chapman and Hall/CRC, 2004.
31. Lee KM, Jang YH, Kim YH, Moon SK, Park JH, Park SW, et al. Reliability and validity of Korean version of National Institutes of Health Stroke Scale: multi-center study. *J Korean Acad Rehabil Med* 2004; 28:422-435.
32. Hills NK, Josephson SA, Lyden PD, Johnston SC. Is the NIHSS certification process too lenient? *Cerebrovasc Dis* 2009;27:426-432.
33. Schmülling S, Grond M, Rudolf J, Kiencke P. Training as a prerequisite for reliable use of NIH Stroke Scale. *Stroke* 1998;29:1258-1259.
34. Dewey HM, Donnan GA, Freeman EJ, Sharples CM, Macdonell RA, McNeil JJ, et al. Interrater reliability of the National Institutes of Health Stroke Scale: rating by neurologists and nurses in a community-based stroke incidence study. *Cerebrovasc Dis* 1999;9:323-327.