How reliable is the National Institutes of Health Stroke Scale (NIHSS) when rated from telemedicine recordings?

Short title: Reliability of the National Institutes of Health Stroke Scale (NIHSS)

Alison McLoughlin[1]
Philippa Olive, Dr [1 2]
Catherine Elizabeth Lightbody, Professor [2]

[1] University of Central Lancashire, School of Nursing, Preston, Lancashire, England

[2] Lancashire Teaching Hospitals NHS Foundation Trust

Corresponding author: Alison McLoughlin, asrmcloughlin1@uclan.ac.uk, 01772 894950

Abstract
Background/Aims: The National Institutes of Health Stroke Scale (NIHSS) is widely used to measure stroke deficits and is deemed to be reliable when used by a range of professionals. This study aimed to establish the inter-rater reliability of the NIHSS when completed via telemedicine. Secondary aims were to explore if professional group, length of time since training and /or re-certification, frequency of use and reason for using the NIHSS influenced the inter-rater reliability. Methods: a total of 30 video clips, representing the equivalent of two whole patient assessments of the 15 NIHSS items, were analysed by a range of NIHSS certified clinical participants. Of which, ten were nurses and five were consultants. Kappa statistics were used to calculate the inter-rater reliability for each item, with additional data on the range of agreement of items. Data across group characteristics were compared to test hypotheses about factors that could impact on reliability. Findings: Overall, the inter-rater reliability was found to be lower than anticipated and there was a wide variation in ratings. Consultants tended to score better than nurses, and counter-intuitively stroke specialist staff and those who used the NIHSS more frequently tended to have poorer reliability than their counterparts. Total agreement on score was only achieved in five out of the 30 video clips (16.6%), with agreement better at either end of the scoring range (i.e. no deficit or worst deficit). These findings indicate that reliability of the NIHSS may be lower than anticipated. Conclusion: Further research is needed to better understand the poor reliability of the NIHSS as this has implications for care decisions and patient outcomes.

Introduction /Background

The National Institutes of Health Stroke Scale (NIHSS) is a 15-item ordinal measure developed in the 1980s as a research tool to allow consistent reporting of neurological deficits in acute-stroke studies, particularly the early trials of thrombolysis and putative neuroprotectants (Brott et al, 1989). The NIHSS has since become widely used for measuring stroke severity and functional deficit to help guide treatment decisions in clinical practice (Goldstein and Davis, 1989; Albanese et al, 1994; Spilker et al, 1997; Dewey et al, 1999; Josephson et al, 2006; Lyden et al, 2009). Despite its widespread use, variation exists in:

- Why clinicians apply the scale (for example, for clinical assessment, prognostication or research outcomes)
- How the scale is administered, at what time points and how often
- Training to administer the NIHSS, official NIHSS certified training, or in-house uncertified training, and whether refresher training is undertaken as recommended (every two years as a minimum).

Furthermore, the global shortage of stroke physicians has led to the introduction of telemedicine to remotely assess stroke patients. These remote assessments often use the NIHSS; however, the reliability of the NIHSS when being assessed via an audio-video link in comparison to face-to-face is not clear. Several studies have investigated the technological systems and processes of remote telemedicine assessment (Shafqat et al, 1999; Handschu et al, 2003; LaMonte et al, 2004; Meyer et al, 2005; Meyer et al, 2008; Berthier et al, 2012; Demaerschalk et al, 2012; Liman et al, 2012; Anderson et al, 2013; Berthier et al, 2013; Wu et al, 2014). However, the principal focus of these studies has been about ensuring specific remote technologies could be used, or that remote assessment could be completed in a timely manner. Some of these studies did test inter-rater reliability of remote NIHSS assessments, but the methodological quality was generally low, such as a low number of raters ranging from 2-10.

This study primarily aimed to establish the inter-rater reliability of the NIHSS when completed via telemedicine. Secondary aims were to explore if professional group, length of time since training and/or re-certification, frequency of use and reason for using the NIHSS influenced the inter-rater reliability.

Methods

This study involved the secondary analysis of video data collected as part of the Acute Stroke Telemedicine: Utility, Training and Evaluation (ASTUTE) project (French et al, 2013; Gibson et al, 2013). The ASTUTE project developed and tested an internet based Standardised Telemedicine Toolkit (STT) that included a training package for health staff, standardised assessments and a checklist to help doctors and nurses use telemedicine. Part of the ASTUTE project compared face-to-face and telemedicine acute stroke NIHSS assessments at a hospital in the Northwest of England. Stroke patient assessment videos recorded for the ASTUTE project required written consent from patients prior to the video recording. This consent stipulation meant that the videos were not real time acute assessments but were recorded after hospital admission.

In total, 22 patient videos were recorded and available to use. The range and severity of neurological deficits across the recruited patients was limited by research consent requirements, and some score options within the 15 items of the NIHSS were not represented. Where insufficient score options were not available from the ASTUTE patient videos, simulations were created. The equivalent of two full patient NIHSS assessment simulations were created and a total of 30 videos (two different options for each NIHSS item) were available for participants to score, which comprised both real patient and simulated assessments.

Selection and recruitment of participants

In line with COSMIN recommendations for sampling (Mokkink et al, 2019), the authors aimed to recruit 40 clinical members of staff from four groups

- Emergency department consultants
- Emergency department nurses

- Stroke and general physicians
- Stroke nurses.

Practitioners were eligible to participate if they had completed NIHSS training and certification. No time limit from certification was stipulated. Practitioners involved in the creation of the ASTUTE or simulated videos were ineligible to take part. Practitioners were provided information about the study and invited to participate via specialist conferences (for example, the UK Stroke Forum) and professional forums (such as, the National Stroke Nurses Forum (NSNF) and the Royal College of Emergency Medicine).

Data collection

The 30 video files were uploaded to a secure server, with security procedures in place to allow participants to access and undertake NIHSS scoring of the video clips. Table 1 outlines the questions asked of participants before commencing their assessments.

Table 1. Questions the participants were asked about their role and use of the National Institute for Health Stroke Scale.

Job Title
Staff Group
Speciality
When did you begin this job? dd/mm/yy
When did you first do the National Institutes of Health Stroke Scale (NIHSS) training? dd/mm/yy
Have you been re-certified? (Yes/No) If yes dd/mm/yy
How often do you use the NIHSS assessment?
When did you last use the NIHSS assessment?
What do you use the NIHSS for?

Using the NIHSS, each participant was asked to assess and score the 30 video files, independently and in one sitting. A database was created in PremiumSoft Navicat for MySQL, in which participants could access and score the videos. Scores were then exported into Microsoft Office excel and checked for consistency and completeness prior to analysis.

Data Analysis

Data were analysed to assess the inter-rater reliability of scores, and whether training and/or experience of using the NIHSS influenced the scoring. Groups were amalgamated and analysed with the following hypotheses:

- Consultants are more reliable than nurses
- Stroke specialist staff are more reliable than emergency department staff
- Those who completed NIHSS re-certification are more reliable than those who did not
- Those who used the NIHSS daily or weekly are more reliable than those who use it less often.

Inter-rater reliability between the two groups in each hypothesis were directly compared using kappa (k) statistics, which quantifies the agreement between examiners above what would be expected by chance (Harrison et al, 2013). Values for kappa can range from -1 (agreement less than chance) through to 0 (expected agreement by chance) and 1 (total agreement). There are a variety of techniques for calculating kappa statistics, but this study used the Fleiss (1971) method. This method was chosen as it can be used where the participants rating one scale point are not necessarily the same as those rating another. This meant all the data could be analysed even if not all video clips were scored by all participants. The results tables highlight where complete sets of scores were not available in groups.

STATA Version 13 data analysis and statistical software package was used to calculate k statistics. The Landis and Koch (1977) classification system was used to define reliability and highlight the differences between groups. Differences in classifications were also recorded to show the items where the biggest variation occurred between participants. The range of agreement or discrepancy in scores between participants was assessed using manual nearest neighbour analysis, which highlights the spread of scores by describing the extent to which a set of scores are clustered or spaced. The authors define a neighbour score as one above or below (i.e., next to each other) and a greater than neighbour score where scores are not next to each other.

Results

In total, 10 nurses and five consultants participated in the study. Participants were primarily UK-based, though one participant was from the USA. Only 12 of the 15 participants (80%) rated all 30 video clips, it is unknown why some did not complete all 30. Overall, k for individual items of the NIHSS only showed 'very good' reliability in one item (sensory) and 'good' reliability in three items (level of consciousness, commands and visual fields for all participants and participants by subgroups) (Table 2). All participants agreed on the same score for only five out of the 30 video clips (16.6%). Total agreement seemed to occur at either end of the scoring range (for example, no deficit or worst deficit). Overall, there was a wide variation in scoring with 50% (15) of the video clips having variation in scoring between two neighbour scores and 33% (10) beyond two neighbour points.

- **Professional groups.**

Table 2 shows the results comparing consultants and nurses. Both groups showed similar reliability across five items ('poor' in gaze, facial palsy, left arm and best language, and 'fair' in right arm). The consultants achieved better reliability in eight items. In four items (questions, commands, ataxia and dysarthria), the reliability was higher by one Landis and

Koch classification (1977). Level of consciousness, visual fields and left leg were two classifications higher and right leg three classifications higher. Reliability was poorer between consultants compared to nurses in two items (sensory reduced by one classification and extinction reduced by three classifications). Although not statistically significant, consultants appeared to rate more reliably than the nurses. However, as a group, consultants particularly struggled with the item of extinction.

Table 2. Kappa scores by item for all participants and consultants compared to nurses.

National Institutes of Health Stroke Scale items	Calculated kappa for all participants	Staff group: consultants (n=5)	Staff group: nurses (n=10, 3 not complete)	Consultants scored better than nurses	Consultants scored worse than nurses
Level of consciousness	0.6495	1	0.5699	↑↑	
Questions	0.3738	0.5833	0.244	↑	
Commands	0.7559	1	0.6469	↑	
Gaze	-0.0384	-0.2037	-0.0648		
Visual fields	0.6504	0.6552	0.2197	↑↑	
Facial palsy	0.0097	-0.0714	-0.0101		
Right arm	0.3182	0.2857	0.2593		
Left arm	0.0595	0.0517	-0.0151		
Right leg	0.4622	1	0.2624	↑↑↑	
Left leg	0.1937	0.5833	0.0065	↑↑	
Ataxia	0.2391	0.2424	0.146	↑	
Sensory	0.8564	0.6552	1		↓

Best language	0.1456	-0.0417	0.1429		
Dysarthria	0.2727	0.2857	0.1515	↑	
Extinction	0.311	0.0909	0.7083		↓↓↓

Each arrow represents direction of one category change in the Landis and Koch (1977) classification

Key to classification of kappa values as per Landis and Koch (1977) classification

kappa range	0.81-1	0.61-0.80	0.41-0.60	0.21-0.40	<0.20
Definition of agreement	Very good	Good	Moderate	Fair	Poor
key	🟩	🟨	🟧	🩷	🟥

Consultants showed the highest agreement on item scores across all the groups. They achieved total agreement on scores in 12 of the 30 (40%) videos. The nurses had total agreement on seven of the 30 videos (23.3%). The consultants also had the lowest number of greater than neighbour scores with only two out of 30 videos (6.6%) showing greater than neighbour score variation.

- **Speciality**

In total, eight participants stated they were a stroke specialist, six emergency department staff and one other. Table 3 shows the k results for comparison. In terms of speciality, emergency department and other staff generally had better reliability than those who classed themselves as a stroke specialist. In six items (gaze, facial palsy, left arm, best language, dysarthria and right arm), the reliability classification was similar between the stroke specialist staff and emergency department staff, with gaze, facial palsy, left arm, best language and dysarthria showing ‘poor’ reliability and right arm showing ‘fair’ reliability. Stroke specialists had better reliability in three items (level of consciousness and sensory improved by one classification, whereas extinction improved by three classifications). Poorer reliability scores for stroke specialists were seen in six items (visual fields and ataxia reduced by one classification, commands and left leg reduced by two classifications, questions reduced by three classifications and right leg reduced by four classifications).

Table 3. Kappa scores by item for all raters and speciality sub-groups.

National Institutes of Health Stroke Scale items	Calculated kappa (all participants)	Speciality: stroke (n=8, 2 not complete)	Speciality: emergency department or other (n=7, 1 not complete)	Stroke speciality scored better than emergency department or other	Stroke speciality scored worse than emergency department or other
Level of consciousness	0.6495	0.6391	0.6045	↑	
Questions	0.3738	0.109	0.7083		↓↓↓
Commands	0.7559	0.5694	1		↓↓
Gaze	-0.0384	-0.0667	-0.1447		
Visual fields	0.6504	0.2381	0.5157		↓
Facial Palsy	0.0097	-0.0714	-0.0282		
Right arm	0.3182	0.2381	0.3		
Left arm	0.0595	-0.0117	0		
Right leg	0.4622	0.1337	1		↓↓↓↓
Left leg	0.1937	-0.05	0.4815		↓↓
Ataxia	0.2391	0.0152	0.4074		↓

Sensory	0.8564	1	0.7455	↑	
Best language	0.1456	0.04	0.1086		
Dysarthria	0.2727	0.2	0.2		
Extinction	0.311	0.6571	0.1319	↑↑↑	

Each arrow represents direction of one category change in the Landis and Koch (1977) classification

Key to classification of kappa values as per Landis and Koch (1977) classification

kappa range	0.81-1	0.61-0.80	0.41-0.60	0.21-0.40	<0.20
Definition of agreement	Very good	Good	Moderate	Fair	Poor
key	🟩	🟨	🟧	🟪	🟥

Emergency department or other staff showed greater agreement than stroke specialists, with total agreement in item scores for 11 of the 30 videos (36.6%) compared with seven of the 30 videos (23.3%) respectively. The emergency department or other speciality group also recorded fewer greater than neighbour range of scores, with only four out of the 30 videos (13.3%) showing greater than neighbour score variation compared with six of the 30 videos (20%) for the stroke specialist group.

- **Recertification**

A total of eight participants (53%) reported having completed re-certification in the NIHSS, although dates were only provided by two. Of the eight participants who re-certified, seven (88%) were nurses and one an emergency department consultant.

Table 4 shows the results when comparing re-certified participants against none re-certified participants. For five items, there was similar reliability between re-certification sub-groups (gaze, facial palsy, left arm and best language showed 'poor' reliability in all participants and right arm showed ‘fair’ reliability in both groups). The reliability classification was better in five items (right leg and sensory improved by one classification, and commands, dysarthria and extinction improved by two classifications). Poorer reliability was found in five items (level of consciousness, questions and ataxia reduced by one classification and visual fields and left leg reduced by two classifications).

In both groups, the calculated inter-rater reliability was lower than the threshold for a clinical tool in practice, with only two items showing ‘very good’ and two showing ‘good’ reliability (Landis and Koch, 1977).

Table 4. Kappa scores by item for all participants and re-certification sub-groups

National Institutes of Health Stroke Scale item	Calculated kappa (all participants)	Calculated kappa re-certified (n=8 3 not complete)	Calculated kappa not-re-certified (n=7)	Re-certified scored better than not re-certified	Re-certified scored worse than not re-certified
Level of consciousness	0.6495	0.566	0.7455		↓
Questions	0.3738	0.2381	0.4909		↓
Commands	0.7559	1	0.5157	↑↑	
Gaze	-0.0384	0.109	-0.1429		
Visual fields	0.6504	0.1463	0.541		↓↓
Facial palsy	0.0097	-0.0946	-0.0126		
Right arm	0.3182	0.2381	0.3		
Left arm	0.0595	-0.0476	0.0994		
Right leg	0.4622	0.5241	0.3277	↑	

Left leg	0.1937	-0.0694	0.4286		↓↓
Ataxia	0.2391	0.0152	0.4074		↓
Sensory	0.8564	1	0.7455	↑	
Best language	0.1456	0.0741	0.0667		
Dysarthria	0.2727	0.5833	0.0278	↑↑	
Extinction	0.311	0.5833	0.1852	↑↑	

Each arrow represents direction of one category change in the Landis and Koch (1977) classification

Key to classification of kappa values as per Landis and Koch (1977) classification

kappa range	0.81-1	0.61-0.80	0.41-0.60	0.21-0.40	<0.20
Definition of agreement	Very good	Good	Moderate	Fair	Poor
key	🟩	🟨	🟧	🩷	🟥

There was a slight trend towards more agreement in the re-certified group as they all agreed on scores for nine out of the 30 videos (30%) compared with five of the 30 videos (16.6%) for the none re-certified group.

- **Frequency of use**

Data was amalgamated into daily or weekly use (nine) versus those who use it less often (six; three reported using the scale monthly and three less often). It was found that all those who re-certified (eight) reported using the NIHSS daily or weekly (Table 5). In seven items there was a similar reliability classification between the used daily or weekly and the used monthly or less often (gaze, facial palsy, right arm, left arm and best language showed ‘poor’ reliability, visual fields showed ‘fair’ reliability and commands showed ‘good’ reliability in all participants). Reliability in three items was better for those who use the NIHSS more often (sensory and dysarthria improved by one classification, whereas extinction improved by three classifications). Poorer reliability scores were seen in five items (level of consciousness, questions, right leg and left leg reduced by one classification, and ataxia reduced by two classifications). The used daily or weekly group has the largest number of 'poor' classifications in relation to the other groups.

Table 5. Kappa scores by item for all participants and frequency of use sub-groups.

National Institutes of Health Stroke Scale item	Calculated kappa (all participants)	Calculated kappa used daily or weekly (n=9 3 did not compete)	Calculated kappa used monthly or less often (n=6)	Used daily or weekly scored better than used monthly or less often	Used daily or weekly scored worse than used monthly or less often
Level of consciousness	0.6495	0.5909	0.7073		↓
Questions	0.3738	0.3077	0.4146		↓
Commands	0.7559	0.7978	0.6571		
Gaze	-0.0384	-0.0227	-0.2		
Visual fields	0.6504	0.2136	0.4783		
Facial palsy	0.0097	-0.0365	-0.1077		
Right arm	0.3182	0.1964	0.2		
Left arm	0.0595	-0.0109	0.0049		
Right leg	0.4622	0.4072	0.4667		↓
Left leg	0.1937	0.0204	0.3333		↓

Ataxia	0.2391	0.0955	0.4667		↓↓
Sensory	0.8564	1	0.7073	↑	
Best language	0.1456	0.2	-0.05		
Dysarthria	0.2727	0.4	0.04	↑	
Extinction	0.311	0.6571	0.1319	↑↑↑	

Each arrow represents direction of one category change in the Landis and Koch (1977) classification.

<u>Key to classification of kappa values as per Landis and Koch (1977) classification</u>

kappa range	0.81-1	0.61-0.80	0.41-0.60	0.21-0.40	<0.20
Definition of agreement	Very good	Good	Moderate	Fair	Poor
key	🟩	🟨	🟧	🟪	🟥

Those who used the NHISS daily or weekly showed a slight trend towards more agreement in item scores, with agreement for eight out of the 30 videos (26.6%) compared with five of the 30 videos (16.6%) for those who used the NIHSS monthly or less often. However, both groups showed similar numbers of greater than neighbour scoring (six out of 30 videos (20%) for the used daily or weekly group and five of the 30 videos (16.6%) for the used monthly or less often group).

<u>Discussion</u>

Overall, reliability was lower than expected from this group of experienced clinicians. In terms of professional groups, there was a tendency for consultants to score better than nurses, which seems reasonable given that medical staff receive more extensive training in neurological assessment than nurses. It was hypothesised that stroke specialist staff would perform better than emergency department or other members of staff and that those who use the NIHSS more frequently would have better reliability, based on the assumption that they would be more familiar and adept with the assessment. However, the findings did not support either hypothesis. It may be that non-stroke specific staff follow the criteria more stringently, whereas experienced raters may pick up bad habits or become complacent in scoring or adding variation into the assessment.

Drift when skills are not used regularly (Albanese et al, 1994; Goldstein and Samsa, 1997), did not seem to be a factor. However, the presence of drift, in both frequent and infrequent raters, and it's potential to reduce reliability warrants further exploration as it has important implications in clinical practice (LaMonte et al, 2004).

Participants were purposely sampled to be certified in the use of the NIHSS as this is considered by many as a requirement for reliable and valid use of the scale (Andre, 2002). It was expected that those who maintain certification would have better reliability, but the data did not support this. Although the re-certified group achieved more total agreement than the non-re-certified group, their calculated reliability was in fact poorer because of overall wider variation in their scores. This replicates earlier studies where improved interclass correlation coefficients with tighter confidence intervals have been reported for non-re-certified raters (Lyden et al, 2005; 2009).

The findings of this study are in parallel with previous research (Albanese et al, 1994; Schmulling et al, 1998) that illustrates certification alone is not sufficient to indicate competence in the performance of the scale (Hinkle, 2014). Additional training, as well as tighter definitions and scoring examples might be needed to reduce the level of variation between individuals' scores, and, therefore, increase reliability (Josephson et al, 2006; Lyden et al, 2005; 2009). Increased understanding of the education and training needed to achieve and maintain competency in testing, as well as reliability in scoring of the NIHSS is required.

It is important to minimise variation in scoring as it could impact patient care, especially where treatments are based on scoring above a certain threshold on the NIHSS. Additionally, reliance on total scores could result in changes in the patient's condition being missed. Raters can get the same total score based on different items within the NIHSS. To accurately assess and measure change in a patient's condition it is important to rate individual neurological deficits. Therefore, in this study, reliability was calculated for each individual item of the NIHSS, which ranged from k=1.0 to k=-0.2. This study concurred with others in that whether face-to-face or telemedicine assessment is conducted, no NIHSS items consistently have very good classifications of reliability, regardless of the kappa method used (Brott et al, 1989; Lyden et al, 1994 (reference added); Goldstein and Davis, 1989; Albanese et al, 1994; Schmulling et al, 1998; Shafqat et al, 1999; Dewey et al, 1999; Handschu et al, 2003; LaMonte et al, 2004; Lyden et al, 2005; Meyer et al, 2005; Josephson et al, 2006; Meyer et al, 2008; Lyden et al, 2009; Berthier et al, 2012; Demaerschalk et al, 2012; Liman et al, 2012; Anderson et al, 2013; Berthier et al, 2013; Wu et al, 2014).

Furthermore, there is a need to come to a consensus on the level of agreement acceptable in clinical practice. A kappa value of 0.41 is deemed statistically acceptable in the Landis and Koch classification (1977), but this could be too lenient if treatment decisions and stroke care quality data are linked to these assessments of the patient's condition (McHugh, 2012).

To improve reliability, some researchers have suggested the removal of the more unreliable items from the NIHSS for example, ataxia (Kasner et al, 1999; Lyden et al, 1999; Berthier et al, 2012). However, the removal of more items could erode the usefulness of the NIHSS to recognise and rate stroke specific deficits in practice. The NIHSS is already criticised for not representing all potential stroke deficits, particularly in posterior and right hemisphere lesions (Linfante et al, 2001; Gottesman et al, 2010). Rather than removing items, it might be useful to develop more in-depth training and competency assessments with the aim of improving agreement, but this would need further development and testing.

As well as the assessment itself, the population in which it is tested is important. Reliability should ideally be tested across a wide population with a full range of responses. Owing to the limited patient videos available, this study was not able to test across the full range of item scores. Despite this, this study showed that there was a tendency for more consistency in agreement in the extremes of the item scores. Total agreement was only present where no deficit or where worst possible deficit were recorded for an item. Therefore, there is potential for more variability when assessing patients with moderate deficits. Although presence or absence and worst score are important variables, for a scale to be useful it needs to be consistent across the whole range of stroke severities. Further large-scale studies are needed to measure reliability and agreement across a whole stroke population.

Limitations

The study had low numbers of both patient and clinician participants, as well as incomplete assessments across all videos by some participants. All these factors could affect the accuracy of kappa for assessing reliability. This study purposefully sampled for participants who are certified in the NIHSS. However, not all professionals using the NIHSS in practice are certified in its use, so variation could be greater than found here. Although the strength of the evidence may be limited, there were some interesting tendencies that indicate further assessment of the reliability of NIHSS items, and the factors that could affect it.

Owing to the number of useable videos available for secondary analysis, this study was restricted. Simulated videos were used to ensure that a minimum of two videos with different deficits were available to rate across each of the 15 items of the NIHSS. However, this does not represent all potential deficits within a whole stroke population. Use of videos minimised variation between rating conditions, but this might not reflect clinical practice. Real time assessment by telemedicine could be a more dynamic experience and factors, such as low light, noise levels, limitations in camera angles and audio-visual quality could potentially further impact the reliability.

Conclusions

This small study indicates that the reliability of the NIHSS could be lower in clinical practice than anticipated. There is variability in how raters score individual items. Clinically this could have significant implications in measuring the treatment effects of interventions or changes in a patient's condition early after a stroke. A continued focus on identifying and addressing issues that impact on inter rater reliability in clinical practice is needed in order to improve the accuracy of the NIHSS and to achieve optimal assessments.

Statement of Ethics

Participants gave their written informed consent for the videos to be used for the ASTUTE project and subsequent analysis. This ethical approval was granted by the National Health Service (NHS) Research Ethics Committee (REC), the Research and Development Committee at the hospital where the project was completed and The University of Central Lancashire's (UCLan) Built, Sport and Health (BuSH) Ethics Committee. This secondary analysis study protocol was also approved by UCLan's Science, Technology, Engineering, Medicine and Health (STEMH) ethics committee.

The authors have no ethical conflicts to disclose.

Funding Sources

This study formed part of a Masters dissertation and had no direct funding. The original Acute Stroke Telemedicine: Utility, Training and Evaluation (ASTUTE) project in which the videos were originally recorded was funded by the National Institute for Health Research (NIHR) Research for Patient Benefit Programme grant.

Keywords (5-6 searchable items)

Stroke, NIHSS, Reliability, Inter-Rater, Agreement

Key points

- The National Institutes of Health Stroke Scale (NIHSS) is a well-established tool in the assessment of stroke patients in both clinical and research practise.
- It is purported to be a reliable assessment when used by a range of professionals.
- The use of the NIHSS via telemedicine for remote assessment is being used in many centres. This research examined the inter-rater reliability of the NIHSS when completed via telemedicine and across different professional groups.
- The findings indicate that reliability of the NIHSS may be lower than anticipated. Further research is needed to better understand variation in stroke assessment because poor reliability and inconsistent scoring could have severe implications for care decisions and patient outcomes.

Reflective Questions

- Do you see variation in the assessment of stroke patients with the National Institutes of Health Stroke Scale (NIHSS)?
- Do you think the factors raised in this article effect the reliability and agreement of NIHSS assessments? Are there others you are aware of?
- What could services do to minimise this variation and ensure more reliability and agreement in patient assessment with the NIHSS?

References

Albanese MA, Clarke WR, Adams HR et al. 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(9):1746-1751. https://doi.org/10.1161/01.str.25.9.1746

Anderson ER, Smith B, Ido M, Frankel M. Remote assessment of stroke using the iPhone 4. J Stroke Cerebrovasc Dis. 2013;22(4):340-344. https://doi.org/10.1016/j.jstrokecerebrovasdis.2011.09.013

André C. The NIH stroke scale is unreliable in untrained hands. J Stroke Cerebrovasc Dis. 2022;11(1):43-46. https://doi.org/10.1053/jscd.2002.123974

Berthier E, Decavel P, Vuiller F et al. Review: Reliability of NIHSS by telemedicine. Eur. Res. in Telemed. 2012;1(3–4):111-114. https://doi.org/10.1016/j.eurtel.2012.09.001

Berthier E, Decavel P, Vuillier F et al. Reliability of NIHSS by telemedicine in non-neurologists. Int J Stroke. 2013;8(4):E11. https://doi.org/10.1111/j.1747-4949.2012.00965.x

Brott T, Adams HPJ, Olinger CP et al. Measurements of acute cerebral infarction: a clinical examination scale. Stroke. 1989;20(7):864-870. https://doi.org/10.1161/01.str.20.7.864

Demaerschalk BM, Vegunta S, Vargas BB et al. Reliability of real-time video smartphone for assessing national institutes of health stroke scale scores in acute stroke patients. Stroke. 2012;43(12):3271-3277. https://doi.org/10.1161/strokeaha.112.669150

Dewey HM, Donnan GA, Freeman EJ 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(6):323-327. https://doi.org/10.1159/000016006

Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bull. 1971;76(5):378-382. https://psycnet.apa.org/doi/10.1037/h0031619

French B, Day E, Watkins CL et al. The challenges of implementing a telestroke network: a systematic review and case study. BMC Med Inform Decis Mak. 2013;13(1):1472-6947. https://doi.org/10.1186/1472-6947-13-125

Gibson J, Day E, Fitzgerald J et al. Using telemedicine for acute stroke assessment. Nurs Times. 2013;109 (35):14-16

Goldstein LB, Davis J. Interrater reliability of the NIH stroke scale. Arch Neurol. 1989;46(6):660-662. https://doi.org/10.1001/archneur.1989.00520420080026

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(2):307-310. https://doi.org/10.1161/01.str.28.2.307

Gottesman RF, Kleinman JT, Davis C et al. The NIHSS-plus: improving cognitive assessment with the NIHSS. Behav Neurol. 2010;22(1-2):11–15. https://doi.org/10.3233/ben-2009-0259

Handschu R, Littmann R, Reulbach U et al. Telemedicine in emergency evaluation of acute stroke: interrater agreement in remote video examination with a novel multimedia system. Stroke. 2003;34(12):2842-2846. https://doi.org/10.1161/01.str.0000102043.70312.e9

Harrison JK, McArthur KS, Quinn T J. Assessment scales in stroke: clinimetric and clinical considerations. Clin Interv Aging. 2013;8:201–211. https://doi.org/10.2147/cia.s32405

Hinkle JL. Reliability and validity of the national institutes of health stroke scale for neuroscience nurses. Stroke. 2014;45(3):e32-e34. https://doi.org/10.1161/strokeaha.113.004243

Josephson SA, Hills NK, Johnston SC. NIH stroke scale reliability in ratings from a large sample of clinicians. Cerebrovasc Dis. 2006;22(5-6):389-395. https://doi.org/10.1159/000094857

Kasner SE, Chalela JA, Luciano JM et al. Reliability and validity of estimating the NIH stroke scale score from medical records. Stroke. 1999;30(8):1534–1537. https://doi.org/10.1161/01.str.30.8.1534

LaMonte MP, Xiao Y, Hu PF et al. Shortening time to stroke treatment using ambulance telemedicine: TeleBAT. J Stroke Cerebrovasc Dis. 2004;13(4):148-154. https://doi.org/10.1016/j.jstrokecerebrovasdis.2004.03.004

Landis J, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174

Liman TG, Winter B, Waldschmidt C et al. Telestroke ambulances in prehospital stroke management: concept and pilot feasibility study. Stroke. 2012;43(8):2086-2090. https://doi.org/10.1161/strokeaha.112.657270

Linfante I, Llinas RH, Schlaug G et al. Diffusion-weighted imaging and national institutes of health stroke scale in the acute phase of posterior-circulation stroke. Arch Neurol. 2001;58(4):621–628. https://doi.org/10.1001/archneur.58.4.621

Lyden P, Brott T, Tilley B et al. Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. Stroke. 1994;25(11):2220–2226. https://doi.org/10.1161/01.str.25.11.2220

Lyden P, Lu M, Jackson C 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(11):2347–2354. https://doi.org/10.1161/01.str.30.11.2347

Lyden P, Raman R, Liu L et al. NIHSS training and certification using a new digital video disk is reliable. Stroke. 2005;36(11):2446-2449. https://doi.org/10.1161/01.str.0000185725.42768.92

Lyden P, Raman R, Liu L rt al. National institutes of health stroke scale certification is reliable across multiple venues. Stroke. 2009;40(7):2507-2511. https://doi.org/10.1161/strokeaha.108.532069

McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.

Meyer BC, Lyden PD, Al-Khoury L et al. Prospective reliability of the STRokE DOC wireless/site independent telemedicine system. Neurology. 2005;64(6):1058-1060. https://doi.org/10.1212/01.wnl.0000154601.26653.e7

Meyer BC, Raman R, Chacon MR et al. Reliability of site-independent telemedicine when assessed by telemedicine-naive stroke practitioners. J Stroke Cerebrovasc Dis. 2008;17(4):181-186. https://doi.org/10.1016/j.jstrokecerebrovasdis.2008.01.008

Mokkink LB, Prinsen CAC, Patrick DL et al. COSMIN study design checklist for patient-reported outcome measurement instruments. 2019. https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf (accessed 30 August 2022)

Schmulling S, Grond M, Kiencke JRP. Training as a prerequisite for reliable use of NIH stroke scale. Stroke. 1998;29(6):1258-1259. https://doi.org/10.1161/01.str.29.6.1258

Shafqat S, Kvedar JC, Guanci MM et al. Role for telemedicine in acute stroke: Feasibility and reliability of remote administration of the NIH stroke scale. Stroke. 1999;30(10):2141-2145. https://doi.org/10.1161/01.str.30.10.2141

Spilker J, Kongable G, Barch C et al. Using the NIH stroke scale to assess stroke patients. J Neurosci Nurs. 1997;29(6):384-392. https://doi.org/10.1097/01376517-199712000-00008

Wu T, Nguyen C, Ankrom C et al. Prehospital utility of rapid stroke evaluation using in-ambulance telemedicine: A pilot feasibility study. Stroke. 2014;45(8):2342-2347. https://doi.org/10.1161/strokeaha.114.005193