Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Inter-Rater Agreement on Cincinnati Prehospital Stroke Scale (CPSS) and Prehospital Acute Stroke Severity Scale (PASS) Between EMS Providers, Neurology Residents and Neurology Consultants

Martin Gude[1], Hans Kirkegaard[1,2], Rolf Blauenfeldt[3], Anne Behrndtz[3], Jeppe Mainz[3], Ingunn Riddervold[4], Claus Z Simonsen[2,3], Niels Hjort[2,3], Søren P Johnsen[5], Grethe Andersen[2,3], Jan Brink Valentin[5]

[1]Department of Research and Development, Prehospital Emergency Medical Services, Central Denmark Region; and Aarhus University Hospital, Aarhus, Denmark; [2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [3]Danish Stroke Center, Department of Neurology, Aarhus University Hospital, Aarhus, Denmark; [4]Norwegian Air Ambulance Foundation, Oslo, Norway; [5]Danish Center for Clinical Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, Denmark

Correspondence: Martin Gude, Prehospital Emergency Medical Services, Central Denmark Region and Aarhus University Hospital, Aarhus, Denmark, Email gude@dadlnet.dk

Objective: To examine the agreement between emergency medical service (EMS) providers, neurology residents and neurology consultants, using the Cincinnati Prehospital Stroke Scale (CPSS) and the Prehospital Acute Stroke Severity Scale (PASS).

Methods: Patients with stroke, transient ischemic attack (TIA) and stroke mimic were included upon primary stroke admission or during rehabilitation. Patients were included from June 2018 to September 2019. Video recordings were made of patients being assessed with CPSS and PASS. The recordings were later presented to the healthcare professionals. To determine relative and absolute interrater reliability in terms of inter-rater agreement (IRA), we used generalisability theory. Group-level agreement was determined against a gold standard and presented as an area under the curve (AUC). The gold standard was a consensus agreement between two neurology consultants.

Results: A total of 120 patient recordings were assessed by 30 EMS providers, two neurology residents and two neurology consultants. Using the CPSS and the PASS, a total of 1,800 assessments were completed by EMS providers, 240 by neurology residents and 240 by neurology consultants. The overall relative and absolute IRA for all items combined from the CPSS and PASS score was 0.84 (95% CI 0.80; 0.87) and 0.81 (95% CI 0.77; 0.85), respectively. Using the CPSS, the agreement on a group-level resulted in AUCs of 0.83 (95% CI 0.78; 0.88) for the EMS providers and 0.86 (95% CI 0.82; 0.90) for the neurology residents when compared with the gold standard. Using the PASS, the AUC was 0.82 (95% CI 0.77; 0.87) for the EMS providers and 0.88 (95% CI 0.84; 0.93) for the neurology residents.

Conclusion: The high relative and absolute inter-rater agreement underpins a high robustness/generalisability of the two scales. A high agreement exists across individual raters and different groups of healthcare professionals supporting widespread applicability of the stroke scales.

Plain Language Summary: Early stroke identification is pivotal to enable faster treatment. To aid this identification, many symptom-based stroke scales have been constructed for both stroke screening and severity assessment. In the prehospital environment, several scales have been evaluated on performance, but only few studies have evaluated the agreement between the ambulance personnel (emergency medical service (EMS) providers) and stroke physicians when interpreting the assessed symptoms in the scales. It is of great importance to know how EMS providers interpret symptoms seen in connection with the use of the scales to focus the continuous training of the EMS providers but also to aid the decision on which scale to implement in ambulances. From previous studies, we know that complex stroke scales are used to a considerably lesser extent than more simple scales, which could be caused by difficulties interpreting specific symptoms. In this study, a variety of methods was

Received: 23 April 2023
Accepted: 30 August 2023
Published: 7 September 2023

applied to determine the inter-rater agreement for two simple stroke scales using dichotomously evaluated symptoms. High inter-rater agreement between EMS providers and Stroke Neurologists exists both between individual raters and between raters grouped according to their profession and seniority. Previous studies have also found high inter-rater agreement for simple stroke scales but lesser agreement for more advanced scales. In conclusion, simple stroke scales seem to produce the highest agreement.

Keywords: observer variation, prehospital emergency care, emergency medical service provider, neurologists, stroke

Introduction

In prehospital stroke management, the use of symptom-based stroke scales is recommended by international guidelines for initial stroke screening and subsequent stroke severity assessment to identify stroke from large-vessel occlusion (LVO).[1,2] In the Central Denmark Region, a combination of the Cincinnati Prehospital Stroke Scale (CPSS)[3,4] and the Prehospital Acute Stroke Severity Scale (PASS)[5] was implemented in the prehospital setting in late 2017 and became the nationwide score in Denmark from mid-2021. The combined use of the CPSS and the PASS is called the Prehospital Stroke Score (PreSS), and its performance has been evaluated prospectively in the prehospital environment.[6]

The CPSS is based on three symptoms (face, arm and speech), and it is one of the most evaluated stroke screening tools in the prehospital setting.[7] Many stroke-severity scales have been constructed in recent years to identify LVO, all based on various combinations of single items of the National Institute of Health Stroke Scale (NIHSS).[8] These scales, including the full NIHSS scale, seem to perform fairly similarly depending on the cohort employed.[9–13] While NIHSS was constructed for in-hospital use by physicians,[8,14] other stroke severity assessment tools were constructed for use in the prehospital setting by emergency medical service (EMS) providers.[9,10] Many studies on prehospital stroke scales have focused on the diagnostic performance based only on included patients and not on all eligible patients.[10,15–17] A prospective study on EMS providers prehospital simultaneous use of different stroke severity scales gave a rare insight into the scales' prehospital feasibility.[9] The EMS providers assessed putative stroke patients following a list of single items. Different combinations of these items were later used to construct seven prehospital stroke severity scales evaluated in the study. The later reconstruction of a specific scale was only possible if all items constituting that scale were used in the initial assessment of the patient by the EMS provider in the prehospital environment. After the assessment of all patients, the reconstruction rate for the full scales was 76.4% for PASS but only 57.2% for the slightly better performing RACE (Rapid Arterial Occlusion Evaluation). This highlights the importance of evaluating how well symptom-based stroke scales are used by EMS providers to establish if some items are so difficult to interpret that the EMS providers tend to avoid them. Only few studies have evaluated the inter-rater agreement (IRA) between the EMS providers and stroke neurologists for whom the NIHSS was originally designed.[4,18–21] PASS consists of three NIHSS items (level of consciousness questions, best gaze and motor arm) with a simple, dichotomous construction, which makes it highly feasible for use in the prehospital environment.[10]

Our objective was to examine the agreement between EMS providers, residents in neurology and consultants in neurology when using the CPSS and PASS by assessing video recordings of patients with stroke symptoms.

Materials and Methods

Setting

Patients were included between June 2018 and September 2019. Video-recorded assessments using the CPSS and PASS were made by research staff (MFG and JM) using a standardised setup (Supplemental Document 1). The videos were viewed and assessed using the CPSS and PASS by 30 EMS providers, two residents in neurology and two consultants in neurology during November and December 2019.

Study Population

Patients

The study population consisted of patients suspected of stroke which included stroke, TIA and stroke mimics. The stroke diagnosis included acute ischemic stroke and intracerebral haemorrhage. No healthy controls were used in the study. The patients were either included upon primary stroke admission (stroke or stroke mimics) or during rehabilitation (stroke) at

the Department of Neurology, Aarhus University Hospital, or at the Department of Rehabilitation, Skive Neurorehabilitation Centre, both located in the Central Denmark Region. Patients with stroke or stroke mimic were identified by physician staff from the two departments who were not related to the study. Patients were included by research staff (MFG and JM) when available at the departments and included not based on their diagnosis. The research staff (MFG and JM) registered that all score items were represented as the patients were included, and the only inclusion where specific symptom/items were requested applied to three patients with severe symptoms (including gaze palsy/deviation) to ensure that all items from the stroke scales were sufficiently represented.

Healthcare Personnel

The healthcare personnel volunteered to participate in the study, and the EMS providers were paid an hourly-based salary. The neurology physicians were employed at the Department of Neurology, Aarhus University Hospital, and they were purposefully selected; the consultants because they worked almost exclusively with stroke both in their clinical work and in research and were therefore considered "gold standard"; the residents, because they (like the consultants) worked as on-duty physicians receiving telephone conference calls from prehospital personnel regarding patients suspected of stroke in the prehospital environment. The group of EMS providers was recruited from advertising bulletins posted in the region's 36 ambulance stations. From a total of 670 EMS providers (including 36 paramedics) employed in the region, 28 emergency medical technician intermediates (EMT-I) and two paramedics volunteered to participate in the study. Both the EMS providers and the neurology physicians used the symptom-based PreSS in their routine clinical work. As part of their basic training, they had completed an e-learning module covering how to assess a patient using the CPSS and PASS, and they had completed a small case-based test. All neurology physicians were NIHSS certified.

The CPSS and PASS

The CPSS comprised three classic symptoms: facial droop, arm weakness and speech impairment.[3] PASS also comprised three symptoms: arm drift (re-used from the CPSS), level of consciousness (month/age) and gaze palsy/deviation.[5] The cut-off values used in this study were pre-defined from the original studies on the CPSS[3] and PASS,[5] and the cut-off values were used in a study prospectively evaluating the performance of PreSS in the prehospital environment.[6] All items/symptoms were dichotomously assessed and positive scores represented one point each. Hence, a maximum score of five points was possible because of the use of arm weakness by both stroke scales. Further information on the use of the CPSS and PASS taught in the E-learning material may be found in Supplemental Document 2. In this study, the data were analysed both as dichotomous outcomes with the predefined cut-offs, but also as three- and five-point ordinal scales.

Video Recordings

Following a pilot phase, a predefined standardised setup was used for the CPPS and PASS video recordings, including the use of the same words for patient instruction and the same item order. Furthermore, the camera was placed on a telescope tripod with a fixed position and with a fixed camera angle (Supplemental Document 1).

Videos Assessments

Each healthcare professional spent one day assessing the video recordings displayed on a laptop. One laptop was assigned to each healthcare professional to ensure that videos were assessed uninfluenced by the other raters. Prior to the assessments, an approx. 20 minutes long e-learning module on the CPSS and PASS was completed (Supplemental Document 2). Every neurologist (resident or consultant) assessed the complete set of videos. In contrast, each of the EMS providers assessed half the videos. Thus, the first and second half of the recordings were assessed by different EMS providers. While EMS providers were randomly divided into two groups, all raters assessed the video recordings in a fixed order. Assessments were recorded in a research database.

Gold Standard

The gold standard for the CPSS and PASS assessments was agreement between the two neurology consultants (CZS and NH) achieved after their individual assessments. All five-item-assessments with complete agreement between the two stroke experts were the gold standard, also referred to as the "consensus assessment". In case of disagreement on single

items, the specific video recording was subsequently reviewed, and a consensus agreement was achieved between the two stroke experts. Thus, included assessments performed by the neurology consultants consisted of both their individual assessments and their consensus assessment.

Statistical Analysis

Using generalisability theory, we estimated the relative and absolute IRA on the five-point total score (CPSS plus PASS) as well as three-point CPSS and PASS scores separately.[22–25] The relative IRA (rIRA) is comparable to the conventional intra-class correlation, while the absolute IRA (aIRA) also accounts for any general displacement between rater-assessment scores.[26] Relative agreement is often referred to as consistency, while absolute agreement is sometimes referred to as agreement.

Bland-Altman plots were applied on the five-point total score (CPSS plus PASS) comparing EMS providers, neurology residents and neurology consultants to the gold standard "the consensus assessment". The 95% limits of agreement (LoA) and the mean difference were calculated and displayed. Limits of agreement are placed at the 2.5 and 97.5 percentile centred at the mean difference and under assumption of normal distributed residuals. In addition, we estimated the slope of the trend line, which is estimated using linear regression.[27]

To assess the agreement between personnel groups (EMS providers, neurology residents and consultants) and the consensus assessment regarding the dichotomous outcomes of the CPSS, PASS and all five single items, we determined the diagnostic performance by calculating the concordance index (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).[28] In this particular case, the AUC represents sensitivity plus specificity divided by two. These quantities provide an estimate of how well EMS personnel as well as neurology residents can distinguish cases from non-cases. To account for violation of the independent observation assumption, the AUC, sensitivity, specificity, PPV and NPV were estimated using a random crossed effects model with random intercepts on raters and patients where applicable. Eg, the sensitivity is estimated as the marginal mean outcome for a positive consensus score using a linear mixed effect regression model with the dichotomous outcome as dependent variable and the dichotomous consensus score as independent variable. The specificity is estimated as one minus the marginal mean outcome for a negative consensus score using the same model.

All results are presented with 95% confidence intervals (CI). Data were analysed using Stata 16 (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC) and R (R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.)

Ethics Approval and Informed Consent

The study was evaluated by the local ethical committee in the Central Denmark Region, which waived the need for ethical approval (case no. 1-10-72-439-17). Informed consent was achieved from the patients or from their nearest relatives in cases where the patients could not fully comprehend the information/provide informed consent. Data storage and management were approved by the Danish Data Protection Agency (no. 1-16-02-701-17). The study was performed in accordance with the ethical standards in the 1964 Declaration of Helsinki and its later amendments.

Data Availability Statement

Anonymised data may be shared upon reasonable request and in pursuance of Danish legislation.

Results

In the Department of Neurology at Aarhus University Hospital, 52 patients were included in 16 separate days. The median time from admission to video recording was 1.0 day (IQR 0.6; 1.3). The department's mean daily admission rate was 10.8 patients (95% CI 9.5–12.1). In the Department of Rehabilitation, Skive Neurorehabilitation Centre (a 27-bed department), 68 patients were included in 12 separate days. The median time from admission to video recording was 13.3 days (IQR 1.0; 41.0).

Based on the total 120 video recordings, 1,800 assessments with CPSS and PASS were performed in the EMS provider group (n=30), 240 in the group of neurology residents (n=2) and 240 in the group of neurology consultants (n=2).

For the gold standard assessment, a perfect agreement between the two neurology consultants was observed in 91 cases for all five items used in the CPSS and PASS. In 27 cases, disagreement existed on one item and in two cases on two items. A consensus agreement was reached for all cases. The consensus assessments included all possible total scores and covered all items (Table 1).

Table 1 Patient Characteristics and Outcomes Based on the Consensus Assessment on the Total 120 Video Recordings

Patient Characteristics	
Age, median (IQR)	68.5 years (60; 75)
Female, n (%)	39 (32.5)
Pathology at admission	
AIS, n (%)	87 (72.5)
ICH, n (%)	12 (10)
TIA, n (%)	9 (7.5)
Stroke mimics, n (%)	12 (10)
Consensus assessment (gold standard)	
Video recordings (total)	
n	120
Total score	
0 points, n (%)	47 (39.2)
1 point, n (%)	35 (29.2)
2 points, n (%)	20 (16.7)
3 points, n (%)	10 (8.3)
4 points, n (%)	7 (5.8)
5 points, n (%)	1 (0.8)
Total test outcomes	
CPSS ≥ 1 point, n (%)	73 (60.8)
PASS ≥ 2 points, n (%)	17 (14.2)
Single items (abnormal/positive scores)	
Arm weakness, n (%)	61 (50.8)
Facial droop, n (%)	29 (24.2)
Speech impairment, n (%)	26 (21.7)
Level of consciousness, n (%)	13 (10.8)
Gaze palsy/deviation, n (%)	9 (7.5)

Notes: The stroke mimic diagnosis included patients with non-stroke related aphasia, dysphasia and amnesia, brain tumours and venous thrombosis.

Abbreviations: CPSS, Cincinnati Prehospital Stroke Scale; PASS, Prehospital Acute Severity Scale; AIS, Acute ischemic stroke; ICH, Intracerebral haemorrhage; TIA, Transient ischemic attack.

A positive CPSS was present in 73 cases (≥1 point), equalling 60.8%. A positive PASS was present in 17 cases (≥2 points), equalling 14.2% (Table 1).

The rIRA and aIRA for the five-point total score (CPSS plus PASS) was 0.84 (95% CI: 0.80; 0.87) and 0.81 (95% CI: 0.77; 0.85). The rIRA and aIRA for the three-point CPSS was 0.77 (95% CI: 0.72; 0.81) and 0.73 (95% CI: 0.68; 0.78), while for PASS the rIRA and aIRA was 0.83 (95% CI: 0.79; 0.86) and 0.81 (95% CI: 0.77; 0.85). The underlying variance contributions of the five-point total score (used to estimate the rIRA and aIRA) are shown in Supplemental Table S1.

Figure 1 and Supplemental Figure S1 show the Bland-Altman plots with trend constrained to zero, whereas Bland-Altman plots without constraints on trend are shown in Figure 2 and Supplemental Figure S2. In the group of neurology consultants, a total of 240 single assessments were plotted and as expected from the construction of the consensus assessment category, they equalled the corresponding consensus assessment in most cases, illustrated by the large dots on the 0-line of the y-axis (Supplemental Figure S1) and a mean difference at 0.01 points. Only few assessments (small dots) deviated and only by a maximum of one point with the LoA from −0.69 to 0.72 with 12.9% outside the LoA.

In the group of neurology residents, a total of 240 single assessments were plotted (Figure 1 right). As expected, more deviations were seen and with a maximum of two points with LoA from −1.60 to 1.14 points and 5.8% outside the LoA. The deviations generally accounted for higher total scores, which resulted in a mean difference in total scores of −0.229 points (dashed line). Hence, the neurology residents scored the patients slightly higher than the consensus on average (Figure 1 right).

In the group of EMS providers, most total scores were still equal to the consensus assessments, which is visible by large dots on the 0-line of the y-axis; but many total scores also deviated (Figure 1 left). Extensive deviation was seen for a few assessments (more than four points), but most only deviated by one point, supported by a LoA ranging from −1.22

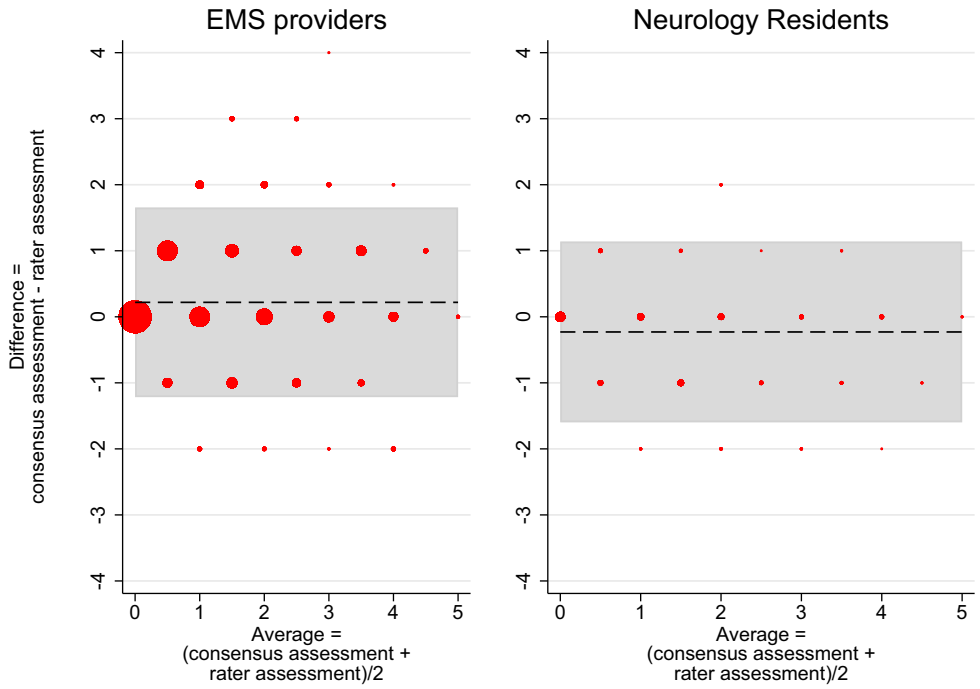

Figure 1 Bland-Altman plots of EMS providers and neurology residents versus consensus assessment.

Notes: Bland Altman plot of the 1,800 total scores from the group of EMS providers (left) and the 240 total scores from the group of neurology residents (right) versus the corresponding consensus assessments. Dashed line = mean difference between raters and the consensus, grey zone = limits of agreement (95%). Marker sizes represent the number of assessments. X-axis = average values of each rater assessment and the corresponding consensus assessment, Y-axis = deviation of the rater score from the consensus assessment. Limits of agreement are placed at the 2.5 and 97.5 percentile centred at the mean difference and under the assumption of normal distributed residuals. Left: Mean difference = 0.22, lower limit of agreement = −1.22, upper limit of agreement = 1.65. 5.39% of the observations are placed outside the limit of agreement. Right: Mean difference = −0.23, lower limit of agreement = −1.60, upper limit of agreement = 1.14. 5. 83% of the observations are placed outside the limit of agreement.

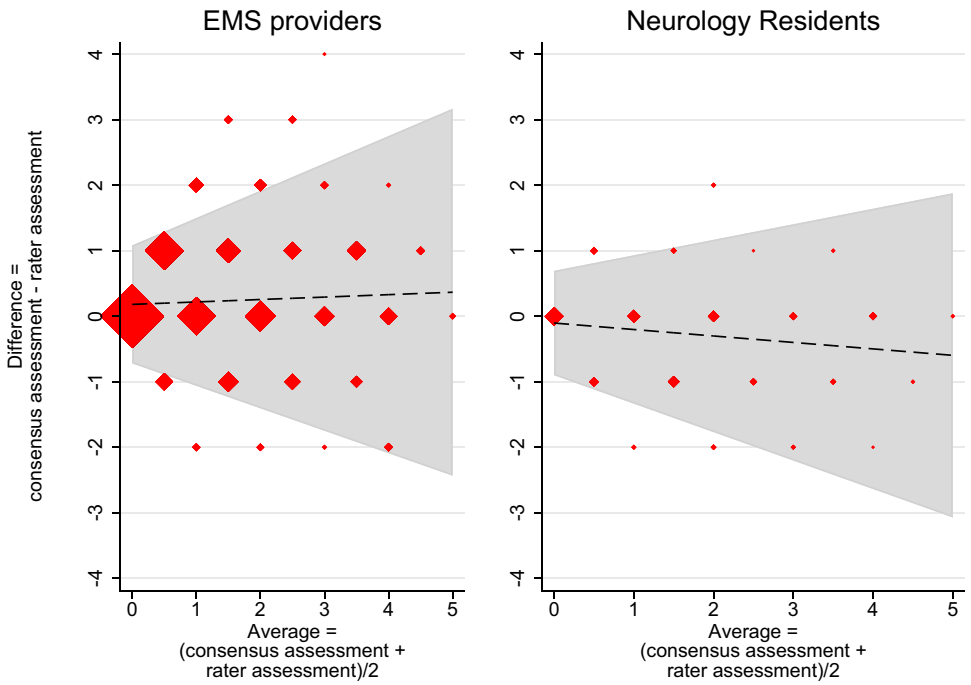

Figure 2 Bland-Altman plots of EMS providers and neurology residents versus consensus assessment – without constrains on trend.

Notes: Bland Altman plot of the 1,800 total scores from the group of EMS providers (left) and the 240 total scores from the group of neurology residents (right) versus the corresponding consensus assessments. Dashed line = trendline of the mean difference between raters and the consensus, grey zone = limits of agreement (95%). Marker sizes represent the number of assessments. X-axis = average values of each rater assessment and the corresponding consensus assessment, Y-axis = deviation of rater score from the consensus assessment. The mean difference trendline is estimated using linear regression, while the limits of agreement are estimated using linear regression on the residuals of the trendline. Limits of agreement represent the 2.5 and 97.5 percentile observations centred at the mean difference and under the assumption of normal distributed residuals. Left: Mean difference trendline = 0.18 + 0.04 × average. Limits of agreement = mean difference ± 2.46 × (0.37 + 0.15 × average). 6.78% of the total scores are placed outside the limit of agreement. Right: Mean difference trendline = −0.10 + (−0.10 × average). Limits of agreement = mean difference ± 2.46 × (0.32 + 0.14 × average). 7.92% of the total scores are placed outside the limit of agreement.

to 1.65 points with 5.39% outside the LoA. On average, the EMS providers slightly underscored the patients, resulting in a mean difference of 0.219 points (Figure 1 left).

Figures 2 and S2 show the Bland-Altman plots without constraints on trend. The trend for neurology consultants and EMS personnel was negligible (Figures 2 left and S2) and statistically insignificant (results not shown), which was to be expected for the neurology consultants. However, the slope of the trend for the neurology residents was −0.10 (95% CI −0.17; −0.03), which indicated that these raters mainly overrated patients with multiple symptoms (Figure 2 right). The clinical implications of the groups' mean total scores and LoAs are described by the AUCs for the dichotomous outcome of CPSS, PASS and single items.

Comparing the dichotomized scores to the consensus assessment, we found that the use of the CPSS resulted in an AUC of 0.95 (95% CI 0.93; 0.98) with balanced and high sensitivity and specificity in the group of neurology consultants, an AUC of 0.86 (95% CI 0.82; 0.90) in the group of neurology residents and an AUC of 0.83 (95% CI 0.78; 0.88) in the group of EMS providers (Table 2). Whereas the neurology residents had a high sensitivity of 92.5% (95% CI 79.4; 100.0) compared with only 72.0% (95% CI 65.1; 79.0) in the EMS provider group, specificity was characterized by opposite measures with a 94.3% specificity (95% CI 85.9; 100.0) in the EMS provider group compared with only 79.8% (95% CI 66.2; 93.4) in the resident group – corresponding to the mean differences observed in the Bland-Altman plots (Figure 1).

The AUCs related to PASS were: 0.94 (95% CI 0.91; 0.96) for the neurology consultants, 0.88 (95% CI 0.84; 0.93) for the neurology residents and 0.82 (95% CI 0.77; 0.87) for the EMS providers (Table 2). The sensitivity was relatively low even for the consultants who had a sensitivity of 88.2% (95% CI 83.6; 92.9). The sensitivities were 79.4% (95% CI 71.1; 87.7) and 65.9% (95% CI 57.0; 74.7) in the group of neurology residents and EMS providers. However, the specificity was generally high (>97%) in all groups.

Table 2 Personnel Groups' Performance Against Consensus (Gold Standard)

Performance Measures (95% CI)	EMS Providers (n = 30, Assessments = 1,800)	Residents in Neurology (n = 2, Assessments = 240)	Consultants in Neurology (n = 2, Assessments = 240)
Cincinnati Prehospital Stroke Scale (CPSS)			
AUC	0.83 (0.78; 0.88)	0.86 (0.82; 0.90)	0.95 (0.93; 0.98)
Sensitivity	72.0% (65.1; 79.0)	92.5% (79.4; 100.0)	95.9% (91.3; 100.0)
Specificity	94.3% (85.9; 100.0)	79.8% (66.2; 93.4)	94.7% (89.5; 100.0)
PPV	95.6% (92.8; 98.4)	88.5% (78.2; 98.8)	96.6% (92.4; 100.0)
NPV	68.8% (66.2; 71.4)	88.7% (77.4; 100.0)	93.8% (88.9; 98.7)
Prehospital Acute Severity Scale (PASS)			
AUC	0.82 (0.77; 0.87)	0.88 (0.84; 0.93)	0.94 (0.91; 0.96)
Sensitivity	65.9% (57.0; 74.7)	79.4% (71.1; 87.7)	88.2% (83.6; 92.9)
Specificity	97.6% (94.0; 100.0)	97.6% (93.3; 100.0)	99.5% (97.6; 100.0)
PPV	82.0% (78.5; 85.4)	84.4% (77.0; 91.9)	96.8% (91.7; 100.0)
NPV	94.5% (93.3; 95.8)	96.6% (93.6; 99.6)	98.1% (96.2; 100.0)
Single items			
Arm weakness			
AUC	0.81 (0.76; 0.86)	0.89 (0.85; 0.93)	0.95 (0.93; 0.98)
Sensitivity	62.3% (54.9; 69.7)	85.2% (72.3; 98.2)	95.9% (90.7; 100.0)
Specificity	99.4% (91.9; 100.0)	92.4% (79.4; 100.0)	94.9% (89.6; 100.0)
Facial droop			
AUC	0.81 (0.76; 0.85)	0.83 (0.77; 0.88)	0.96 (0.93; 0.99)
Sensitivity	73.8% (65.2; 82.3)	91.4% (69.1; 100.0)	94.8% (89.7; 100.0)
Specificity	88.1% (82.6; 93.5)	73.6% (53.0; 94.2)	96.7% (93.5; 99.9)
Speech impairment			
AUC	0.79 (0.75; 0.82)	0.91 (0.86; 0.96)	0.93 (0.90; 0.96)
Sensitivity	57.5% (50.6; 64.3)	92.3% (82.5; 100.0)	86.5% (81.6; 91.5)
Specificity	99.7% (96.0; 100.0)	90.4% (84.3; 96.6)	98.9% (96.3; 100.0)
Level of consciousness (month/age)			
AUC	0.98 (0.97; 0.99)	0.95 (0.92; 0.98)	1.00 (0.98; 1.00)
Sensitivity	96.9% (94.9; 98.9)	92.3% (86.3; 98.3)	100.0% (97.5; 100.0)
Specificity	99.2% (98.5; 99.9)	98.1% (96.1; 100.0)	99.5% (98.7; 100.0)
Gaze palsy/deviation			
AUC	0.92 (0.88; 0.97)	0.96 (0.91; 1.00)	0.92 (0.89; 0.94)
Sensitivity	88.9% (80.0; 98.0)	94.4% (85.7; 100.0)	83.3% (78.6; 88.0)
Specificity	95.9% (93.1; 98.8)	97.3% (94.7; 99.9)	100.0% (98.7; 100)

Abbreviations: AUC, area under the receiver-operating curve; PPV, positive predictive value; NPV, negative predictive value; EMS, emergency medical service; (95% CI), all measures with 95% confidence intervals in parentheses.

With respect to single items, the EMS providers had the lowest performance regarding arm weakness, facial droop and speech impairment. The AUCs for the three items ranged from 0.79 (95% CI 0.75; 0.82) to 0.81 (95% CI 0.76; 0.86) (Table 2). The differences in AUCs between various rater groups are shown in Supplementary Table S2.

Discussion

While the ultimate goal of CPSS and PASS is early identification of stroke patients, the first step in the process is accurate recognition of symptoms. For this purpose, the agreement between neurologists and EMS providers was shown to be high when measured by generalisability theory and by AUCs. By defining the gold standard as the consensus assessment based only on the two neurology consultants (stroke experts) as a group, it was possible to evaluate the agreement between stroke physicians with different seniority (residents versus consultants). By evaluating this agreement (residents versus consultants), it is possible to provide a nuanced picture against which the performance of EMS providers should be measured. Many studies have used a setup in which the performance of individual stroke physicians was chosen as the gold standard against which the performance of EMS providers was measured.[4,18–21] As seen in our study, the inter-rater agreement among stroke physicians is not perfect, which is mirrored by other studies applying the full NIHSS.[8,29–31] Hence, the performance of EMS providers might be underestimated when compared with individual neurologists as gold standard.

In this study, the performance of the EMS providers should be compared with the performance of the group of neurology residents because they did not contribute to the consensus assessment, and they still represent on-duty stroke physicians, certified in the use of NIHSS. The differences in AUCs between neurology residents and EMS providers were insignificant when using the CPSS and significant but relatively small when using the PASS (Table S2). These results fit well with the high inter-rater agreement shown by the generalisability theory and visualized in the Bland-Altman plots.

The agreement on the total scores (shown in the Bland-Altman plots) cannot be directly interpreted as clinically meaningful because the total score consists of both the CPSS and the PASS. However, a relatively narrow LoA was seen in all groups, and only small mean differences were found with a general tendency to score patients lower in the EMS provider group and higher in the group of neurology residents. The tendency to score patients lower in the EMS provider group was also seen in the lower sensitivity for the CPSS, PASS and among several single items, generally with high specificities. The small mean differences on the Bland-Altman plots coincide with the small differences found between rIRA and aIRA. Both analyses indicate high agreement, and the main driver for the difference between the rIRA and aIRA is that the EMS providers generally score lower than the resident neurologists.

The clinical implication of the lower sensitivity is not clear. The tendency could reflect that EMS providers made conservative/restrictive assessments to avoid mistakes when participating in an unfamiliar study setup. The EMS providers have previously shown a high sensitivity identifying stroke and TIA using CPSS in the prehospital setting in our region,[6] which is in line with other studies conducted in a real-life setting.[20]

Previous studies on the agreement among physicians using the full NIHSS have identified the items with the lowest agreement measured by weighted Cohen's kappa (K_w) to be facial palsy and dysarthria.[21,29,32] This fits well with our findings where these single items had the lowest AUC in all groups. More training on these specific items might be needed to increase EMS provider performance. Even so, the performance on these single items and the CPSS (that includes these items) was still good in our study, possibly because of the dichotomous scoring.

Prehospital studies on inter-rater agreement between EMS providers and stroke physicians (equivalent to the neurology physicians in our study) vary with respect to size (from 31 to 171 patients) and results.[4,18–21] Different measures of the agreement have been reported with a K_w of 0.58, 0.69 and 0.818 when using the full-scale NIHSS,[18] NIHSS-8[21] and the Rapid Arterial Occlusion Evaluation Scale (RACE),[19] respectively, and with an ICC of 0.89 when using the CPSS.[4] An association seems to exist between stroke-scale complexity and the inter-rater agreement.[18,19,21] The use of the simple dichotomous CPSS produced a high ICC (0.89),[4] a result very close to our result (rIRA at 0.84). In contrast, the use of the full-scale NIHSS produced the lowest K_w among the presented studies.[18] Also, the high inter-rater agreement found in our study might explain why these simple scores more easily get implemented and used in the emergency medical service compared with more advanced scales.[6,9] With only small differences in prehospital

performance,[9,10] the stroke scales with the highest usability should be preferred when the difference in use accounts for up to 20%.[9]

Previous studies on inter-rater agreement represent different study setups with individual strengths and limitations. Studies in which the EMS providers used a stroke scale in the prehospital environment represent a great strength by not only reporting on the interpretation of a scale but also its usage.[18,20] A downside to this setup is the timing of the sequential assessment by the stroke physician, taking place upon hospital admission and thereby being separated in time why symptoms may potentially have changed.[18,20] A small study where video telemedicine was used reported a very good inter-rater agreement (K_w = 0.818) between stroke physicians and EMS providers using RACE. Unfortunately, the study only included 31 out of 52 eligible patients.

A methodological strength of the present study is the use of video-recorded standardised patient assessments as all raters were presented with the same patient assessment, which brought a number of benefits: the presentation of symptoms was the same, especially as partial paresis was not assessed repeatedly; no time delays were present during which symptoms might change; the answers to the question about current age and present month could not be remembered from one assessment to another; raters could not be influenced by other raters (a circumstance not clearly described in other studies);[4,19] and all raters viewed the patients from the same angle, in the same light, etc.

One limitation to this study is that the raters did not instruct the patients to perform the item constituting the CPSS and PASS. For that reason, the agreement between raters applies only to the interpretation of the patient examination and not to the scales full use. Moreover, the video recorded assessments were performed with the patients placed in hospital beds, a familiar setup for the stroke physicians but not the EMS providers. Performing in an unfamiliar study setup, the EMS providers could have assessed patients restrictively to avoid mistakes maybe explaining the lower score tendency, maybe introducing a respondent bias.

We did not perform sub analyses according to time from symptom onset to video assessment. No patients were included within the treatment window of i.v. thrombolysis, and some patients were included weeks after the initial admission which might have included weak symptoms (partial paresis).

Finally, the number of neurology consultants and residents were relatively low. It was a convenience sample, but we believe the included neurology physicians to be representative.

In conclusion, this study provided evidence for a high inter-rater agreement between healthcare professionals with different educational backgrounds and seniority when using the CPSS and PASS. The agreement was measured with a variety of methods at both an individual and a group-based level. These findings underpin the hypothesis that CPSS and PASS may be used in routine clinical practice where they may facilitate effective communication between the EMS and the hospital.

Abbreviations

EMS, Emergency medical service; CPSS, Cincinnati Prehospital Stroke Scale; PASS, Prehospital Acute Stroke Severity Scale; ICC, Intra-class correlation; IRA, Inter-rater agreement; rIRA, relative IRA; aIRA, absolute IRA; RR, Rating reliability; AUC, Area under the receiver-operating characteristic curve; LVO, Large-vessel occlusion; PreSS, Prehospital Stroke Score; NIHSS, National Institute of Health Stroke Scale; LoA, Limits of agreement; PPV, Positive predictive value; NPV, Negative predictive value; Kw, Weighted Cohen's kappa.

Acknowledgments

The authors take this opportunity to express their gratitude to the ambulance services in the Central Denmark Region for their participation in the study.

Funding

The study was supported by a grant from the Danish non-profit foundation TrygFonden (grant number 117615) and by a grant from the Laerdal Foundation. The foundations did not influence the study, the drafting of the manuscript or the interpretation of the results.

Disclosure

Dr Rolf Blauenfeldt reports speaker fees from Novo Nordisk and Bayer, outside the submitted work. Prof. Dr Claus Z. Simonsen reports grants from Novo Nordisk Foundation and Health Research Foundation of Central Denmark Region, during the conduct of the study. The authors report no other conflicts of interest in this work.

References

1. Powers WJ, Rabinstein AA, Ackerson T, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: a Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. *Stroke*. 2019;50(12):e344–418.
2. Kobayashi A, Czlonkowska A, Ford GA. European Academy of Neurology and European Stroke Organization consensus statement and practical guidance for pre-hospital management of stroke. *Eur J Neurol*. 2018;25(3):425–433. doi:10.1111/ene.13539
3. Kothari R, Hall K, Brott T, Broderick J. Early stroke recognition: developing an out-of-hospital NIH Stroke Scale. *Acad Emerg Med*. 1997;4 (10):986–990. doi:10.1111/j.1553-2712.1997.tb03665.x
4. Kothari RU, Pancioli A, Liu T, Brott T, Broderick J. Cincinnati Prehospital Stroke Scale: reproducibility and validity. *Ann Emerg Med*. 1999;33 (4):373–378. doi:10.1016/S0196-0644(99)70299-4
5. Hastrup S, Damgaard D, Johnsen SP, Andersen G. Prehospital Acute Stroke Severity Scale to predict large artery occlusion: design and Comparison With Other Scales. *Stroke*. 2016;47(7):1772–1776. doi:10.1161/STROKEAHA.115.012482
6. Gude MF, Blauenfeldt RA, Behrndtz AB, et al. The Prehospital Stroke Score and telephone conference: a prospective validation. *Acta Neurol Scand*. 2022;145(5):541–550. doi:10.1111/ane.13580
7. Zhelev Z, Walker G, Henschke N, Fridhandler J, Yip S. Prehospital stroke scales as screening tools for early identification of stroke and transient ischemic attack. *Cochrane Database Syst Rev*. 2019;4(4):Cd011427. doi:10.1002/14651858.CD011427.pub2
8. Brott T, Adams HP Jr, Olinger CP, et al. Measurements of acute cerebral infarction: a clinical examination scale. *Stroke*. 1989;20(7):864–870. doi:10.1161/01.STR.20.7.864
9. Nguyen TTM, van den Wijngaard IR, Bosch J, et al. Comparison of Prehospital Scales for predicting large anterior vessel occlusion in the ambulance setting. *JAMA Neurol*. 2021;78(2):157–64.
10. Duvekot MHC, Venema E, Rozeman AD, et al. Comparison of eight prehospital stroke scales to detect intracranial large-vessel occlusion in suspected stroke (PRESTO): a prospective observational study. *Lancet Neurol*. 2021;20(3):213–221. doi:10.1016/S1474-4422(20)30439-7
11. Vidale S, Agostoni E. Prehospital stroke scales and large vessel occlusion: a systematic review. *Acta Neurol Scand*. 2018;138(1):24–31. doi:10.1111/ane.12908
12. Purrucker JC, Härtig F, Richter H, et al. Design and validation of a clinical scale for prehospital stroke recognition, severity grading and prediction of large vessel occlusion: the shortened NIH Stroke Scale for emergency medical services. *BMJ open*. 2017;7(9):e016893. doi:10.1136/bmjopen-2017-016893
13. Smith EE, Kent DM, Bulsara KR, et al. Accuracy of prediction instruments for diagnosing large vessel occlusion in individuals with suspected stroke: a systematic review for the 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke. *Stroke*. 2018;49(3):e111–e122. doi:10.1161/STR.0000000000000160
14. Lyden P. Using the National Institutes of Health Stroke Scale: a Cautionary Tale. *Stroke*. 2017;48(2):513–519. doi:10.1161/STROKEAHA.116.015434
15. Perez de la Ossa N, Carrera D, Gorchs M, et al. Design and validation of a prehospital stroke scale to predict large arterial occlusion: the rapid arterial occlusion evaluation scale. *Stroke*. 2014;45(1):87–91. doi:10.1161/STROKEAHA.113.003071
16. Okuno Y, Yamagami H, Kataoka H, et al. Field Assessment of Critical Stroke by Emergency Services for Acute Delivery to a Comprehensive Stroke Center: FACE(2)AD. *Transl Stroke Res*. 2020;11(4):664–670. doi:10.1007/s12975-019-00751-6
17. Václavík D, Bar M, Klečka L, Holeš D, Čábal M, Mikulík R. Prehospital stroke scale (FAST PLUS Test) predicts patients with intracranial large vessel occlusion. *Brain Behav*. 2018;8(9):e01087. doi:10.1002/brb3.1087
18. Larsen K, Jæger HS, Hov MR, et al. Streamlining acute stroke care by Introducing National Institutes of Health Stroke Scale in the emergency medical services: a Prospective Cohort Study. *Stroke*. 2022;53(6):2050–2057.
19. Hackett CT, Rahangdale R, Protetch J, et al. Rapid Arterial Occlusion Evaluation Scale Agreement between Emergency Medical Services Technicians and Neurologists. *J Stroke Cerebrovasc Dis*. 2020;29(6):104745. doi:10.1016/j.jstrokecerebrovasdis.2020.104745
20. Mulkerin WD, Spokoyny I, Francisco JT, et al. Prehospital identification of large vessel occlusions using modified national institutes of health stroke scale: a pilot study. *Front Neurol*. 2021;12:643356. doi:10.3389/fneur.2021.643356
21. Demeestere J, Garcia-Esperon C, Lin L, et al. Validation of the National Institutes of Health Stroke Scale-8 to detect large vessel occlusion in ischemic stroke. *J Stroke Cerebrovasc Dis*. 2017;26(7):1419–1426. doi:10.1016/j.jstrokecerebrovasdis.2017.03.020
22. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach*. 2012;34 (11):960–992. doi:10.3109/0142159X.2012.703791
23. Huebner A, Lucht M. Generalizability theory in R. *Pract Assess Res Evaluation*. 2019;24(1):5.
24. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: a practical guide to study design, implementation, and interpretation. *J Sch Psychol*. 2014;52(1):13–35. doi:10.1016/j.jsp.2013.11.008
25. Brennan RL. *Generalizability Theory*. Vol. 1. New York, NY: Springer; 2001.
26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428. doi:10.1037/0033-2909.86.2.420
27. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–160. doi:10.1177/096228029900800204
28. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014;35(29):1925–1931. doi:10.1093/eurheartj/ehu207

29. 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. doi:10.1159/000094857
30. Hov MR, Røislien J, Lindner T, et al. Stroke severity quantification by critical care physicians in a mobile stroke unit. *Eur J Emerg Med*. 2019;26(3):194–198. doi:10.1097/MEJ.0000000000000529
31. 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(7):2507–2511. doi:10.1161/STROKEAHA.108.532069
32. Meyer BC, Hemmen TM, Jackson CM, Lyden PD. Modified National Institutes of Health Stroke Scale for use in stroke clinical trials: prospective reliability and validity. *Stroke*. 2002;33(5):1261–1266. doi:10.1161/01.STR.0000015625.87603.A7