OPEN ACCESS

Original research

National Institutes of Health Stroke Scale scores obtained using a mobile application compared to the conventional paper form: a randomised controlled validation study

Helge Fagerheim Bugge ,[1,2,3] Mona Marie Guterud,[1,2,4] Jo Røislien,[1,5] Karianne Larsen,[1,6] Hege Ihle-Hansen,[3] Mathias Toft,[2,3] Maren Ranhoff Hov,[1,3,7] Else Charlotte Sandset[1,3]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjinnov-2022-000968).

For numbered affiliations see end of article.

Correspondence to
Mr Helge Fagerheim Bugge, Norwegian Air Ambulance Foundation, Oslo, Norway; helge.bugge@norskluftambulanse.no

Received 29 March 2022
Accepted 25 May 2022

To cite: Bugge HF, Guterud MM, Røislien J, *et al*. *BMJ Innov* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjinnov-2022-000968

ABSTRACT

Background Prehospital delay contributes to treatment delay in acute stroke. Numerous prehospital stroke scales exist for stroke identification, but they lack the diagnostic accuracy of the in-hospital National Institutes of Health Stroke Scale (NIHSS). We have developed a mobile application to aid paramedics assessing prehospital NIHSS. This study explores agreement between NIHSS scores obtained using the Paramedic Norwegian Acute Stroke Prehospital Project (ParaNASPP) application compared with conventional assessment.

Methods 25 physicians working with stroke were randomised to an application group or control. 20 unique videos portraying acute stroke symptoms were scored by both groups. 95% limits of agreement (LoA) were calculated using Bland-Altman's method, comparing the groups to predefined scores, and each other. LoAs within ±3 NIHSS points were considered acceptable. Cohen's kappa was calculated on dichotomised NIHSS scores.

Results The ParaNASPP application group had 95% LoA of −2.33 to 2.71. The control group had LoA of −2.60 to 2.55. Direct comparison between the groups gave LoA of −3.12 to 3.55. When compared with the dichotomised predefined scores kappa was 0.93 in the application group and 0.89 in the control group. Kappa was 0.84 for direct comparison between the groups.

Discussion There was very good agreement between the application and both the predefined score and the control group. Scores from the ParaNASPP application differ slightly more than our predefined goal when compared with the control group, but is well within when compared with the predefined NIHSS scores. We consider this acceptable and the ParaNASPP application validated for further clinical studies.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Prehospital stroke scales are not up to the standard of the in-hospital National Institutes of Health Stroke Scale (NIHSS), which is considered to complex for prehospital use.

WHAT THIS STUDY ADDS

⇒ Using a mobile application developed for paramedics to aid NIHSS assessment has a very good agreement with the conventional paper form.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

⇒ This is the first step toward implementing prehospital NIHSS, though more research and clinical testing is needed.

BACKGROUND

Medical evidence and new technologies change the way patients who had an acute stroke are diagnosed and treated,[1] and short onset-to-treatment time remains an important predictor of outcome.[2 3]

Better prehospital and in-hospital stroke assessment, including early prehospital identification of stroke, triage to the right level of care and reduced door-to-needle time may improve stroke prognosis.[4] Recognition of acute stroke symptoms in prehospital settings is essential in order to transport patients to the right level of care, and a consensus statement issued by the European Stroke Organisation and the European Academy of Neurology recommend specific training for paramedics.[5]

The National Institutes of Health Stroke Scale ((NIHSS) online supplemental material I) is the most validated tool for assessment of patients who had an acute stroke among doctors and nurses in hospitals.[6] It is measured as a continuous scale from 0 to 40, divided into 11 categories, and a higher score indicates more severe stroke.[7] An ideal prehospital stroke scale is easy to use, quick, reliable and ensures good communication with in-hospital stroke physicians. Paramedics in Europe, and Norway, predominantly uses the Face Arm Speech Test (FAST) to assess patients with suspected acute stroke.[8] FAST and most other prehospital scales are derived from NIHSS, but lack the diagnostic accuracy of NIHSS.[9–11] NIHSS has been shown to be better at predicting large vessel occlusion (LVO) than prehospital stroke scales.[12] Due to its complexity, NIHSS has been considered less suited to the prehospital setting,[9 13] but new studies are showing it might be a valuable prehospital tool.[14]

In the Paramedic Norwegian Acute Stroke Prehospital Project (ParaNASPP), we have developed a mobile application to aid paramedics using NIHSS on patients with suspected acute stroke ((ParaNASPP application) online supplemental material II), including a specific paramedic training programme. This concept may enable prehospital NIHSS as an efficient, accurate and objective tool for a common language in the chain of stroke survival. With mobile devices and applications now widely available in clinical settings, trained paramedics can use the application to perform accurate NIHSS scoring in suspected stroke patients and directly transfer patient data to the on-call stroke physician.[15]

In this study, we aim to explore the agreement in NIHSS scoring obtained using the ParaNASPP application compared with NIHSS scores obtained using the standard paper version. A good agreement would validate the ParaNASPP application for further clinical use in prehospital and in-hospital settings.

METHODS

Participants

Physicians working with stroke were voluntarily enrolled from two Norwegian hospitals, Akershus University Hospital (AHUS) and Molde Hospital. Each participant received a compensation for time spent on study-specific tasks. In total, 25 physicians (consultants and residents in geriatrics and neurology) signed up for participation. We opted to use physicians working with stroke to assess the inter-rater agreement between NIHSS from the ParaNASPP application and the standard paper NIHSS.

Participants were randomised into an application group and a control group using a block randomisation on GraphPad.com.[16] One block was AHUS (n=21) and one was Molde Hospital (n=4). The randomisation created two study groups, one that used the ParaNASPP application to score NIHSS (n=13) and a control group using the conventional paper version of NIHSS (n=12).

Cases

We created 20 cases with different acute stroke scenarios, which were acted by a physician working with stroke and filmed. Each of the 20 cases had a predefined symptom presentation and total NIHSS score defined by the research team. The NIHSS scores varied from 0 to 31, with a median score of 7. The variation and total score of the cases was made to reflect a real-world stroke population.[17] The examiner in each video is an NIHSS certified[18] paramedic with no prior knowledge of the cases. The cases were numbered from 1 to 20. All participants were given access to the videos through a web page (https://www.akuttportalen.no/).

Application group

None of the participants had used the ParaNASPP application prior to this study. Instructions on how to download the application to an Apple iOS mobile device were provided. Information registered in the application (mobile number of the user, case number, time used in minutes and total NIHSS) was sent directly to a secure database provided by Tjenester for sensitive data (TSD) at the University of Oslo.

Mobile application

The ParaNASPP application is based on pictograms and text guiding the scorer through the NIHSS examination. Each of the items in the standard paper version of NIHSS is made as a separate screen in the application, and all scoring options have their own pictogram showing the corresponding symptom. This makes the application easy to use and the scoring process more intuitive.

Once an item has been scored the application automatically moves on to the next item. When all items of the NIHSS is completed a summary screen showing the total score and all separate items appears. In addition, for future use in the clinical ParaNASPP study[15] with data transfer to the on-call stroke physician, it prompts registration of time of onset, sex, age, blood pressure, blood glucose and anti-thrombotic medication in addition to NIHSS. In this study, a modified version of the application limited to NIHSS registration was used. Figure 1 is an example of a pictogram.

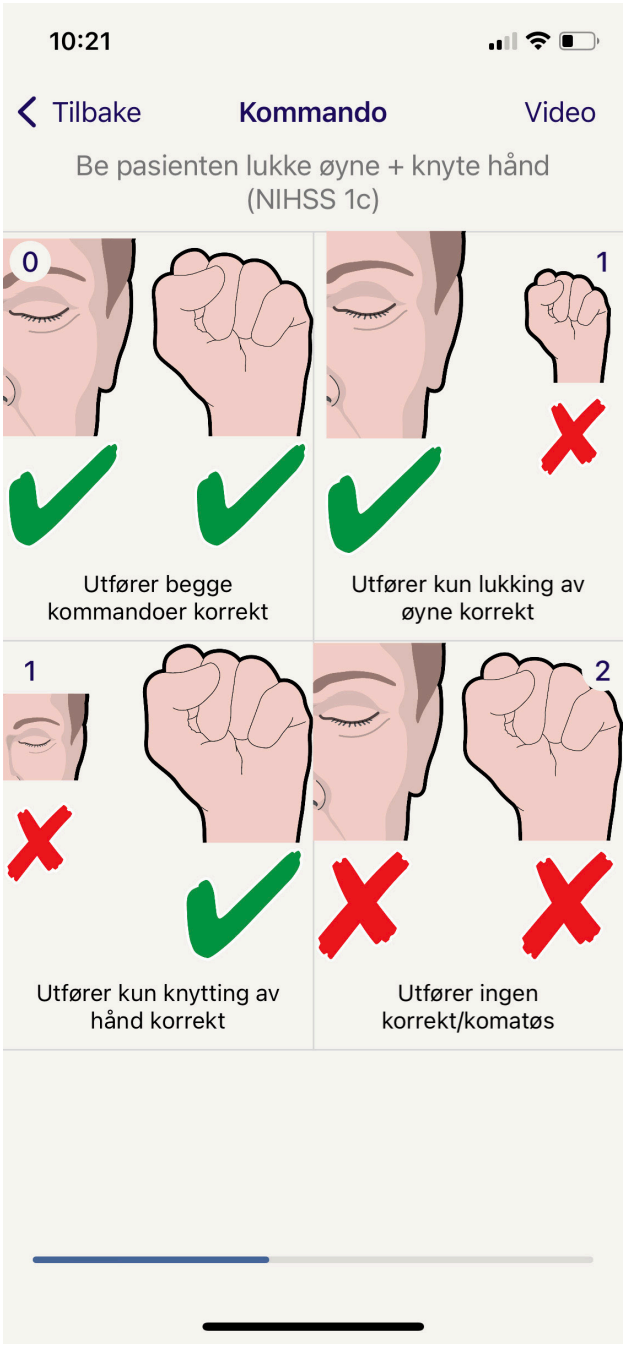

Figure 1 Level of consciousness command 1c as shown in application. NIHSS, National Institutes of Health Stroke Scale.

Control group

The control group used the validated Norwegian NIHSS paper version provided by the study group[19] (online supplemental material I). Time intervals were recorded by the participants, form video start to completed NIHSS score. The study form including time recordings, every aspect of NIHSS and total NIHSS score was returned to the research team and manually transferred to TSD.

Patient and public involvement

No patients were involved in this study. We used a voluntary physician working with stroke to act out all the cases to ensure that each participant had the same cases and symptoms to score.

STATISTICS

The ParaNASPP application and standard paper scorings were compared with the predefined score of each video using Bland-Altman's method for method comparison.[20] The 95% limits of agreement (LoA) were calculated for each Bland-Altman plot. A width of the LoAs of ±3 points was a priori deemed a clinically acceptable variation in NIHSS measurements.[21] Separate comparisons between each of the application and control subgroups and the predefined score were calculated, as well as a direct assessment of agreement between NIHSS scores when using the application and paper using a mixed models generalisation of the Bland-Altman approach.[22]

A dichotomisation of the continuous NIHSS scores into two categories: a low score group (NIHSS ≤5) and a high score group (NIHSS ≥6) was performed. This cut-off was selected as it is considered clinically relevant as NIHSS ≥6 is often used as an eligibility criterion for endovascular therapy and has a relatively high sensitivity for detecting LVO.[23–25] Cohen's kappa inter-rater agreement was performed to calculate agreement between the scripted score of each case and the score in the application and paper groups, respectively. A score variability leading to a change in category was considered to potentially result in altered triage and treatment options. Since a mixed method Cohen's kappa for comparing multiple raters does not exist, a crude Cohen's kappa estimate for comparing NIHSS scores between application and control groups, combining all value pairs in the same cross table, was performed. Kappa ≤0.2 represents poor agreement, 0.21–0.4 fair agreement, 0.41–0.6 moderate agreement, 0.61–0.80 good agreement and 0.81–1.0 very good agreement.[26]

Statistical analysis was performed with R V.4.0.3 and STATA statistical software (V.16.1).

RESULTS

Of the 25 enrolled physicians working with stroke, 23 completed the study tasks, of which nine were neurologists, 13 were resident neurologists and one was a resident in geriatrics, while two opted out without returning their scores. Median clinical experience of participating physicians was 5.5 years, with a range of 0.7–20 years. Overall, 10 of 23 participants were NIHSS certified.[18] There were no major differences in baseline characteristics between the participants randomised to application or control groups (table 1).

The application group scored NIHSS with a mean difference of 0.19 and LoA from −2.33 to 2.71 as compared with the predefined scores. For the paper group, the mean difference was −0.027 and LoA from −2.60 to 2.55 as compared with the predefined scores. Figure 2 shows the corresponding Bland-Altman plots.

Table 1 Baseline characteristics of application and control groups

	Application	Control
Participants	12	11
Specialist	5	4
Median years experience	5 years and 3 months (5.28 years) Min: 0.66 Max: 20	5 years and 9 months (5.76 years) Min: 1 Max: 18
Hospital	AHUS: 10 Molde: 2	AHUS: 9 Molde: 2
NIHSS certification	5	5

AHUS, Akershus University Hospital; NIHSS, National Institutes of Health Stroke Scale.

Given the clinical significance of NIHSS values being ≤5 or ≥6, we calculated LoAs also for these subsets separately. For the application subgroups, the LoA was −0.70 to 1.34, with mean difference 0.32 for NIHSS scores ≤5 and −2.91 to 3.14 with a mean difference of 0.12 for NIHSS scores ≥6. In the control subgroups, the ≤5 group had 95% LoA from −1.34 to 1.49 with a mean difference of 0.08. In the NIHSS ≥6 group on paper, the 95% LoA was −3.10 to 2.94, with a mean difference from the predefined score of −0.08.

In a direct assessment of agreement between the application and paper groups, the mean difference was 0.21 with LoA from −3.12 to 3.55. When compared with predefined scores, the inter-rater agreement of the dichotomised NIHSS categories (≤5 and ≥6) was kappa 0.93 in the application group and kappa 0.89 in the control group. For predefined scores ≤5, 5.9% of the application group scored ≥6, whereas this only occurred in 1.3% among the control group. Oppositely, when predefined scores were ≥6, underestimation of the scores happened in 1.2% of the application scores and 6.9% of the paper scores.

A direct Cohen's kappa score between the two groups showed very good agreement with a kappa score of 0.842.

DISCUSSION

Efficiency and accuracy in the hyper-acute phase of stroke requires a common language. Prehospital NIHSS may ensure that prehospital personnel recognise symptoms of acute stroke; however, better tools are needed. The physicians working with stroke using the ParaNASPP application has very good inter-rater agreement of NIHSS scores compared with physicians working with stroke in the control group using the conventional paper version.

When comparing the application group and the control group to the predefined NIHSS scores, both groups had similar scores to the predefined score,

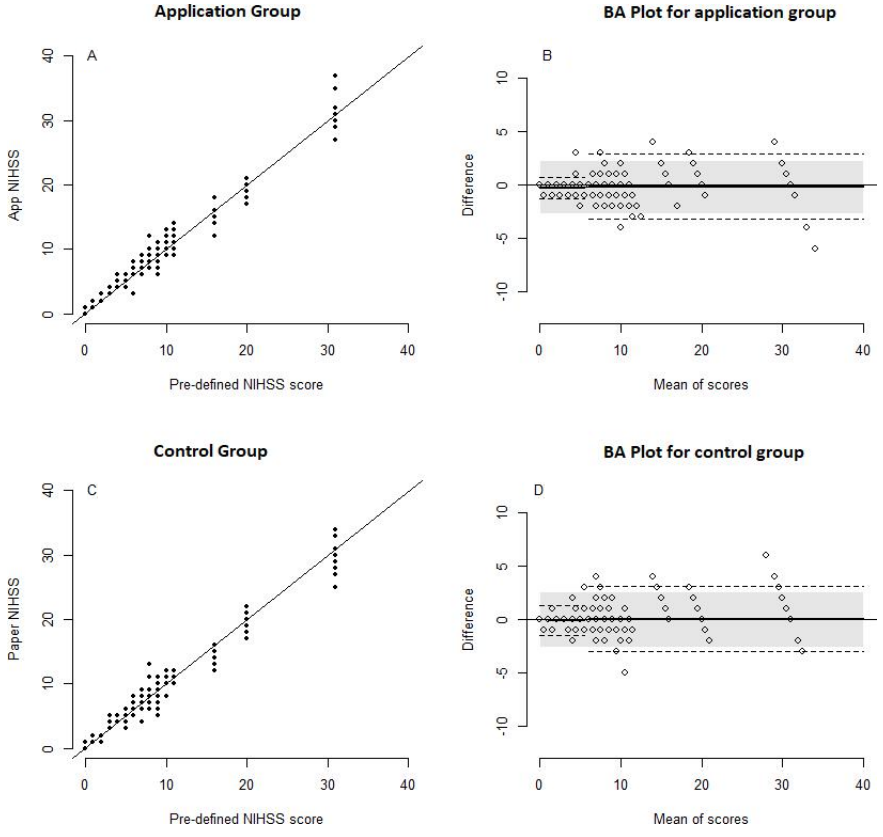

Figure 2 (A) Shows National Institutes of Health Stroke Scale (NIHSS) scores form the application group versus the predefined scores for each case. (B) Bland-Altman (BA) plot for application group versus predefined scores. (C) NIHSS scores from control group versus predefined scores. (D) Bland-Altman plot for control groups versus predefined NIHSS scores.

within the ±3 point margin that we had deemed a good result. The mean difference was slightly higher in the application group than in the paper groups. In the direct comparison of the application to the paper group, the mean was slightly higher in the application group. This is suggestive of slightly higher NIHSS scores being given in the application group, though the variance is most correctly portrayed using the LoAs.

When exploring the dichotomised NIHSS categories, the LoAs in both the application group and the paper group were narrower for the low score category (NIHSS ≤5). For the groups in total and for the NIHSS score ≤5 subgroups, all 95% LoAs were within the predetermined ±3 points. There was very good agreement of NIHSS scores between the application group and the paper groups with a Cohen's kappa of 0.84. A clinical test is valid if it accurately portrays or describes the symptoms of a patient. In our study, there is a difference of more than our predefined ±3 points between the application and the paper groups; however, the accuracy between the application and the predefined score is very good. Hence, despite being outside our strict margins (when the application is directly compared with the paper scores), the application still accurately portrays the actual symptoms enacted in the videos. In our study setting, and unlike the real world, we had a predefined, correct score for each case. One could argue that the ±3 points limit is best applied to this correct score, and not to the scores from the control group. The low score subgroups had narrower LoAs compared with the high score groups, which is of more importance when the stroke symptoms are subtle. A change of ±3 points might change the triage of the patient in the low score group, while a three point difference will be less likely to influence treatment decisions in the high score group.

For the dichotomisation of NIHSS into groups ≤5 and ≥6 as a cut-off point was chosen as this is in accordance with other studies and has been proposed as a limit for identifying LVO.[24 27] In the application group, there was a tendency of over-scoring, and more cases with a predefined score of ≤5 were scored as ≥6, while the paper group tended towards lower scores and had more cases defined as ≥6 scored as ≤5. Overestimation of scores while using the application could potentially lead to an over-triage of patients who had a suspected stroke. An over-triage of 30% is currently considered acceptable, cost beneficial and necessary to ensure patients suffering stroke do not miss treatment options.[28 29] Over-triage from an application to be used in a prehospital setting is in accordance with this and likely benefits patients suffering from stroke.

Overall, our results show very good inter-rater agreement between scores obtained by the application group compared with both the predefined scores and the paper group scores. NIHSS is a broader and more thorough examination than the other prehospital stroke scales, with better diagnostic accuracy and the advantage that it is also used in-hospital. Using the same examination, prehospital and in-hospital, ensure compatible communication through the stroke treatment chain. These results are reassuring for the use in the prehospital setting. The ParaNASPP application also transfers the results from the prehospital examination to the in-hospital stroke team, ensuring that safe handover of patient information. The ParaNASPP application including prehospital NIHSS will be further tested in the large prehospital randomised clinical trial (ParaNASPP).

Limitations

Not all the participants in this validation study were certified in NIHSS, and this may have contributed to a larger spread within both the application and paper groups, but is likely to represent a generalisable sample of Norwegian physicians working with stroke. This study only uses videos of a physician acting cases with different stroke symptoms, and not real patients. This is a limitation as the application is not tested assessing patients in a real clinical setting in this study, but also a strength as the video format provides standardised cases to score and ensures that both groups assess equivalent stroke symptoms.

The lack of a proper statistical method for calculating a Cohen's kappa directly between the application and the paper group that adjust for replicate measurement and multiple raters is another limitation. Still, the comparison between the groups and the predefined scores gives a very good indicator that the application can be used to accurately assess stroke symptoms.

CONCLUSION

The ParaNASPP application groups had very good agreement of NIHSS scores compared with the predefined NIHSS scores, with the application performing slightly better than the paper version. We consider the application validated and ready to be tested in a clinical environment. The ParaNASPP study will provide further evidence if the application improves prehospital identification of stroke.

Author affiliations
[1]Norwegian Air Ambulance Foundation, Oslo, Norway
[2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[3]Department of Neurology, Oslo University Hospital, Oslo, Norway
[4]Division of Prehospital Services, Oslo University Hospital, Oslo, Norway
[5]Faculty of Health Sciences, University of Stavanger, Stavanger, Norway
[6]Institute of Basal Medical Science, University of Oslo, Oslo, Norway
[7]Oslo Metropolitan Univeristy, Oslo, Norway

Acknowledgements We would like to thank all the participants in this study, Bernd Müller, Head of Neurology Department at Molde Hospital, Fredrik Kaupang, app-developer at Norwegian Air Ambulance Technology and Bjørn Fossen, System Manager at Norwegian Air Ambulance Technology for development and technical support throughout the study.

Contributors ECS, MRH and HFB conceived the study. HFB collected the data and administered the study and wrote the first draft of this manuscript. JR and HFB did statistical analyses. All

authors critically reviewed and approved the final version of the manuscript. HFB is acting as guarantor.

Funding This validation study is funded by the Norwegian Air Ambulance Foundation. Grant number is not applicable.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by The Regional Ethics Board south-east, Norway approved this study (number: 2018/2310) as part of the larger Paramedic Norwegian Acute Stroke Prehospital Project clinical trial. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Helge Fagerheim Bugge http://orcid.org/0000-0002-5614-7867

REFERENCES

1 Rabinstein AA. Update on treatment of acute ischemic stroke. *Continuum* 2020;26:268–86.
2 Meretoja A, Keshtkaran M, Tatlisumak T, *et al*. Endovascular therapy for ischemic stroke: save a minute-save a week. *Neurology* 2017;88:2123–7.
3 Meretoja A, Keshtkaran M, Saver JL, *et al*. Stroke thrombolysis: save a minute, save a day. *Stroke* 2014;45:1053–8.
4 Berglund A, Svensson L, Sjöstrand C, *et al*. Higher prehospital priority level of stroke improves thrombolysis frequency and time to stroke unit: the hyper acute stroke alarm (HASTA) study. *Stroke* 2012;43:2666–70.
5 Kobayashi A, Czlonkowska A, Ford GA, *et al*. European Academy of Neurology and European stroke organization consensus statement and practical guidance for pre-hospital management of stroke. *Eur J Neurol* 2018;25:425–33.
6 Lyden P. Using the National Institutes of health stroke scale: a cautionary tale. *Stroke* 2017;48:513–9.
7 Brott T, Marler JR, Olinger CP, *et al*. Measurements of acute cerebral infarction: lesion size by computed tomography. *Stroke* 1989;20:871–5.
8 Fassbender K, Balucani C, Walter S, *et al*. Streamlining of prehospital stroke management: the golden hour. *Lancet Neurol* 2013;12:585–96.
9 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:e111–22.
10 Brandler ES, Sharma M, Sinert RH, *et al*. Prehospital stroke scales in urban environments: a systematic review. *Neurology* 2014;82:2241–9.
11 Nguyen TTM, van den Wijngaard IR, Bosch J. Comparison of prehospital scales for predicting large anterior vessel occlusion in the ambulance setting. *JAMA Neurol* 2020.
12 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;20:213–21.
13 Kesinger MR, Sequeira DJ, Buffalini S, *et al*. Comparing National Institutes of health stroke scale among a stroke team and helicopter emergency medical service providers. *Stroke* 2015;46:575–8.
14 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:Strokeaha121036084:2050–7.
15 Bugge HF, Guterud M, Bache KCG, *et al*. Paramedic Norwegian acute stroke prehospital project (ParaNASPP) study protocol: a stepped wedge randomised trial of stroke screening using the National Institutes of health stroke scale in the ambulance. *Trials* 2022;23:113.
16 GraphPad. Randomly assign subjects to treatment groups 2021. Available: https://www.graphpad.com/quickcalcs/randomize1/
17 Fjærtoft H, Skogseth-Stephani R, Indredavik B. Årsrapport for 2020 Med plan for forbedringstiltak. Trondheim: Nasjonalt sekretariat for Norsk hjerneslagregister. *Seksjon for medisinske kvalitetsregistre* 2021.
18 NIH Stroke Scale International [Internet], 1999. Available: http://www.nihstrokescale.org/
19 Kristensen DV, Johnsen NT, Amthor K-F, *et al*. Culturally adapted validated translation of the NIHSS. *Sykepleien Forsk*:e–82736.
20 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
21 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;43:3271–7.
22 Carstensen B, Simpson J, Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. *Int J Biostat* 2008;4:Article 16.
23 Hansen CK, Christensen A, Ovesen C, *et al*. Stroke severity and incidence of acute large vessel occlusions in patients with hyper-acute cerebral ischemia: results from a prospective cohort study based on CT-angiography (CTA). *Int J Stroke* 2015;10:336–42.
24 Heldner MR, Zubler C, Mattle HP, *et al*. National Institutes of health stroke scale score and vessel occlusion in 2152 patients with acute ischemic stroke. *Stroke* 2013;44:1153–7.
25 Scheitz JF, Abdul-Rahim AH, MacIsaac RL, *et al*. Clinical selection strategies to identify ischemic stroke patients with large anterior vessel occlusion: results from SITS-ISTR (safe implementation of thrombolysis in stroke international stroke thrombolysis registry). *Stroke* 2017;48:290–7.

26 Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992;304:1491–4.
27 Vanacker P, Heldner MR, Amiguet M, *et al*. Prediction of large vessel occlusions in acute stroke: National Institute of health stroke scale is hard to beat. *Crit Care Med* 2016;44:e336–43.
28 Acker JE, Pancioli AM, Crocco TJ, *et al*. Implementation strategies for emergency medical services within stroke systems of care: a policy statement from the American heart Association/American stroke association expert panel on emergency medical services systems and the stroke Council. *Stroke* 2007;38:3097–115.
29 Liberman AL, Choi H-J, French DD, *et al*. Is the cost-effectiveness of stroke thrombolysis affected by proportion of stroke mimics? *Stroke* 2019;50:463–8.